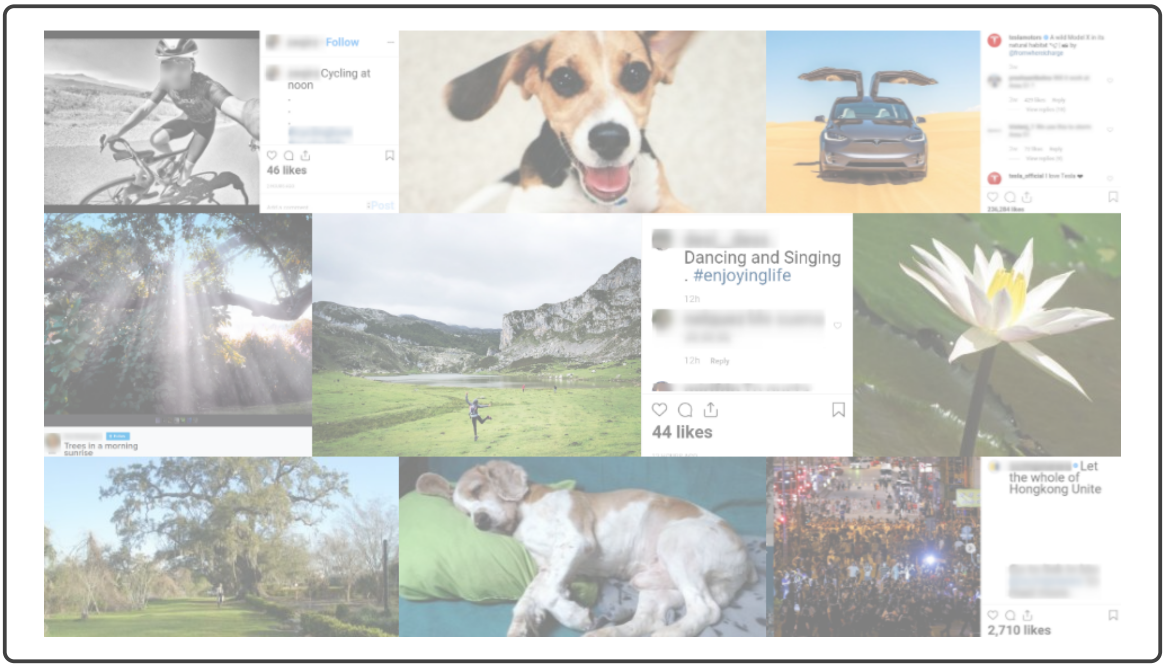
DISSERTATION

AFFECTIVE IMAGE CAPTIONING
EXTRACTION AND SEMANTIC ARRANGEMENT OF IMAGE INFORMATION WITH DEEP NEURAL NETWORKS

Generating Human-Like image descriptions with Deep Learning methods


Follow
Cycling at noon
46 likes
Dancing and Singing . #enjoyinglife
44 likes
Trees in a morning sunrise
Let the whole of Hongkong Unite
2,710 likes

Thesis approved by the
Department of Computer Science
of the TU Kaiserslautern
for the award of the Doctoral Degree

DOCTOR OF NATURAL SCIENCES
(DR. RER. NAT.)

to

TUSHAR KARAYIL

Date of the viva: 10.06.2020
Dean: Prof. Dr. Jens Schmitt

Reviewers:
Prof. Dr. Prof. h.c. Andreas Dengel
Prof. Dr. Shanley E.M. Allen

TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

D 386

Tushar Karayil: **Affective Image Captioning***: Extraction and Semantic Arrangement of Image Information with Deep Neural Networks*

SUPERVISORS:
Prof. Dr. Prof. h.c. Andreas Dengel

LOCATION:
Kaiserslautern, Germany

ABSTRACT

In recent years, the Internet has become a major source of visual information exchange. Popular social platforms have reported an average of 80 million photo uploads a day. These images, are often accompanied with a user provided text one-liner, called an *image caption*. Deep Learning techniques have made significant advances towards automatic generation of *factual image captions*. However, captions generated by humans are much more than mere factual image descriptions. This work takes a step towards enhancing a machine's ability to generate image captions with human-like properties. We name this field as *Affective Image Captioning*, to differentiate it from the other areas of research focused on generating factual descriptions.

To deepen our understanding of human generated captions, we first perform a large-scale Crowd-Sourcing study on a subset of Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M). Three thousand random image-caption pairs were evaluated by native English speakers w.r.t different dimensions like focus, intent, emotion, meaning, and visibility. Our findings indicate three important underlying properties of human captions: **subjectivity**, **sentiment**, and **variability**. Based on these results, we develop Deep Learning models to address each of these dimensions.

To address the **subjectivity** dimension, we propose the Focus-Aspect-Value (FAV) model (along with a new task of *aspect-detection*) to structure the process of capturing subjectivity. We also introduce a novel dataset, *aspects-DB*, following this way of modeling. To implement the model, we propose a novel architecture called *Tensor Fusion*. Our experiments show that Tensor Fusion outperforms the state-of-the-art cross residual networks (XResNet) in aspect-detection.

Towards the **sentiment** dimension, we propose two models: *Concept & Syntax Transition Network (CAST)* and *Show & Tell with Emotions (STEM)*. The CAST model uses a graphical structure to generate sentiment. The STEM model uses a neural network to inject adjectives into a neutral caption. Achieving a high score of 93% with human evaluation, these models were selected as the top-3 at the *ACMMM Grand Challenge 2016*.

To address the last dimension, **variability**, we take a generative approach called Generative Adversarial Networks (GAN) along with multimodal fusion. Our modified GAN, with two discriminators, is trained using Reinforcement Learning. We also show, that it is possible to control the properties of the generated caption-variations with an external signal. Using sentiment as the external signal, we show that we can easily outperform state-of-the-art sentiment caption models.

ACKNOWLEDGMENTS

My sincere gratitude to my supervisor, Prof. Andreas Dengel for his continuous support, encouragement and technical guidance directing this work to its conclusion. I would also like to thank Prof. Shanley Allen for her reviews and valuable feedback.

SDS group at DFKI, where I work, has been a wonderful and diverse collection of talents. My colleagues at MADM sub-group have been most supporting during these four years. Thank you Philipp for the long technical discussions, which often times, went to deep philosophical topics (at the end of which, we would stare through our office-windows in silence). Jörn and Federico, for helping me organize, review and complete this thesis. To Joachim, who would often drop by in my room with a completely random topic, resulting in an interesting discussion. I have also enjoyed the company of my colleagues Sebastian, Benjamin, Patrick, Andrey, Stanislav, Shailza and Fatemeh. Damian Borth, who guided me during the early years of the work has provided me with interesting technical directions. To Christian, who kept all the machines up-to-date and to Brigitte, who would often find me quick meeting appointments in an otherwise busy schedule of Andreas.

The support of my family and friends has been immense. I would like to thank them all. My wife, Madhuri, has been most encouraging during these years. To my son Siddharth, who is 1.9 years old when this work is published. To my father-in-law, Bhima Raju and mother-in-law, KrishnaVeni for having travelled to take care of Siddharth (while I was busy writing the thesis). To my sister, Sisira and brother-in-law, Ravi, for all the help. To dearest Raman-Uncle, whose persistent encouragement, to see me reach the highest in academics, has proved fruitful. To my uncle Ramachandran and aunt, Sushuma, whose life-ideals are an inspiration to me. To Nachu, for always being a supportive brother. To Anoop, Deepak, Sheefeque and Ganesh for ..., well, being friends! (difficult to thank friends).

Most importantly, to my parents, Ravindran and Sairandhri. My Father and Mother have taught me how to remain stoic and move forward while facing adversities in any path. Their life ethics and unconditional love, have inspired me in all facets of life.

Finally, to Sri M, for being a teacher and a guiding light in my life.

— Thank you, Tushar

GRANTS

This work has been supported by the TU Kaiserslautern CS department's PhD Program, the BMBF projects MOM (Grant 01IW15002), DeFuseNN (Grant 01IW17002) and the NVIDIA AI Lab program (NVAIL).

PUBLICATIONS RESULTING FROM THIS THESIS

Parts of the research and material (including figures, tables and algorithms) in this thesis have already been published in (or accepted in):

JOURNAL

Philipp Blandfort*, Tushar Karayil*, Jörn Hees, and Andreas Dengel. "The Focus-Aspect-Value Model for Predicting Subjective Visual Attributes." In: *International Journal of Multimedia Information Retrieval* (January 2020), pp. 1–14. DOI: 10.1007/s13735-019-00188-5. URL: https://link.springer.com/article/10.1007/s13735-019-00188-5.

CONFERENCE

Tushar Karayil*, Philipp Blandfort*, Jörn Hees, and Andreas Dengel. "The Focus-Aspect-Value Model for Explainable Prediction of Subjective Visual Interpretation." In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. ACM. 2019, pp. 16–24.

Tushar Karayil, Asif Irfan, Federico Raue, Jörn Hees, and Andreas Dengel. "Conditional GANs for Image Captioning with Sentiments." In: *Proceedings of the 28th International Conference Artificial Neural Networks, vol 11730*. Springer. 2019. DOI: 10.1007/978-3-030-30490-4_25. URL: https://link.springer.com/chapter/10.1007/978-3-030-30490-4_25.

Tushar Karayil*, Philipp Blandfort*, Damian Borth, and Andreas Dengel. "Generating affective captions using concept and syntax transition networks." In: *Proceedings of the 24th ACM international conference on Multimedia*. ACM. 2016, pp. 1111–1115. DOI: 110.1145/2964284.2984070. URL: https://dl.acm.org/doi/10.1145/2964284.2984070.

* Equal Contribution from Authors

Philipp Blandfort*, Tushar Karayil*, Damian Borth, and Andreas Dengel. “Introducing concept and syntax transition networks for image captioning.” In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM. 2016, pp. 385–388. DOI: 10.1145/2911996.2930060. URL: https://dl.acm.org/doi/10.1145/2911996.2930060.

Philipp Blandfort, Tushar Karayil, Federico Raue, Jörn Hees, and Andreas Dengel. “Fusion Strategies for Learning User Embeddings with Neural Networks.” In: *Proceedings of the 2019 International Joint Conference for Neural Networks* (2019).

WORKSHOP & ARXIV

Philipp Blandfort*, Tushar Karayil*, Damian Borth, and Andreas Dengel. “Image Captioning in the Wild: How People Caption Images on Flickr.” In: (2017), pp. 21–29.

Tushar Karayil*, Philipp Blandfort*, Jörn Hees, and Andreas Dengel. “The Focus-Aspect-Polarity Model for Predicting Subjective Noun Attributes in Images.” In: *arXiv preprint* (2018).

* Equal Contribution from Authors

CONTENTS

V SUMMARY AND OUTLOOK

VI APPENDIX

VII REFERENCES & CV

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CNN	Convolutional Neural Network
RNN	Recurrent Neural Networks
GAN	Generative Adversarial Network
CAST	Concept and Syntax Transition Network
STEM	Show and Tell with Emotions
MSCOCO	Microsoft Common Objects in Context
YFCC100M	Yahoo Flickr Creative Commons 100 Million
FAV	Focus Aspect Value
CGAN	Conditional Generative Adversarial Network
LSTM	Long Short Term Memory
FC	Fully Connected
RL	Reinforcement Learning
NLP	Natural Language Processing
DSB	DeepSentiBank
LCS	Longest Common Subsequence
BLEU	Bilingual Evaluation Understudy
METEOR	Metric for Evaluation for Translation with Explicit Ordering

ROUGE	Recall Oriented Understudy for Gisting Evaluation
CIDER	Consensus-based Image Description Evaluation

Part I

INTRODUCTION

1

INTRODUCTION

Internet as multimodal communication medium

The exponential growth of the Internet has brought-in multiple changes in our lives. One of these prominent changes has been in our (social) communication. Not only have our communication methods become faster, they have also become multimodal. With modern web-pages becoming more animated and interactive, the visual and text modalities have dominated these mediums. The introduction of Web 2.0 has also brought an era of user generated content. This has been furthered by a proliferation of social media platforms like Facebook, Instagram, Flickr etc. With nearly 75% of "Internet-Surfers" using social media, Instagram alone has reported an average of 80 million photo uploads a day[1] [1]. A similar report from Facebook states that its users have uploaded on an average 350 million photos a day since its inception[2] (till 2018).

visual and text medium

Of the vision and text modalities, the visual medium dominates the communication with an estimated projection of 80% by the year 2020 [2]. Although the visual medium dominates the communication, textual modality also plays an equally important role. This is because the uploaded visual contents are often accompanied by a short piece of text, usually a one-liner, called a *caption*. These text snippets mostly describe the visual content or provide subjective interpretations of the author. An analysis of one million random Flickr images, from the Yahoo Flickr Creative Commons 100 Million (YFCC100M), shows that 70% of the uploaded images have a user generated text accompanying them (Figure 1.1). The caption lengths vary, with the majority falling in the range of $[2-5]$ words (as depicted in Figure 1.2). This estimate excludes the automatically generated machine/camera image captions.

An *image caption*, therefore, is a piece of text which usually describes the content of the image or is in some manner connected to the context of the same. The motivating question then becomes:

Why are image captions important?

More importantly, in the context of this thesis:

Why should machines generate image captions?.

1.1 MOTIVATION

There are multiple reasons as to why users write an image caption. It varies from communication, engagement, entertainment to product-marketing. We explore these in the following sections.

[1] https://instagram-press.com/blog/2015/09/22/celebrating-a-community-of-400-million/

[2] https://www.brandwatch.com/blog/facebook-statistics/

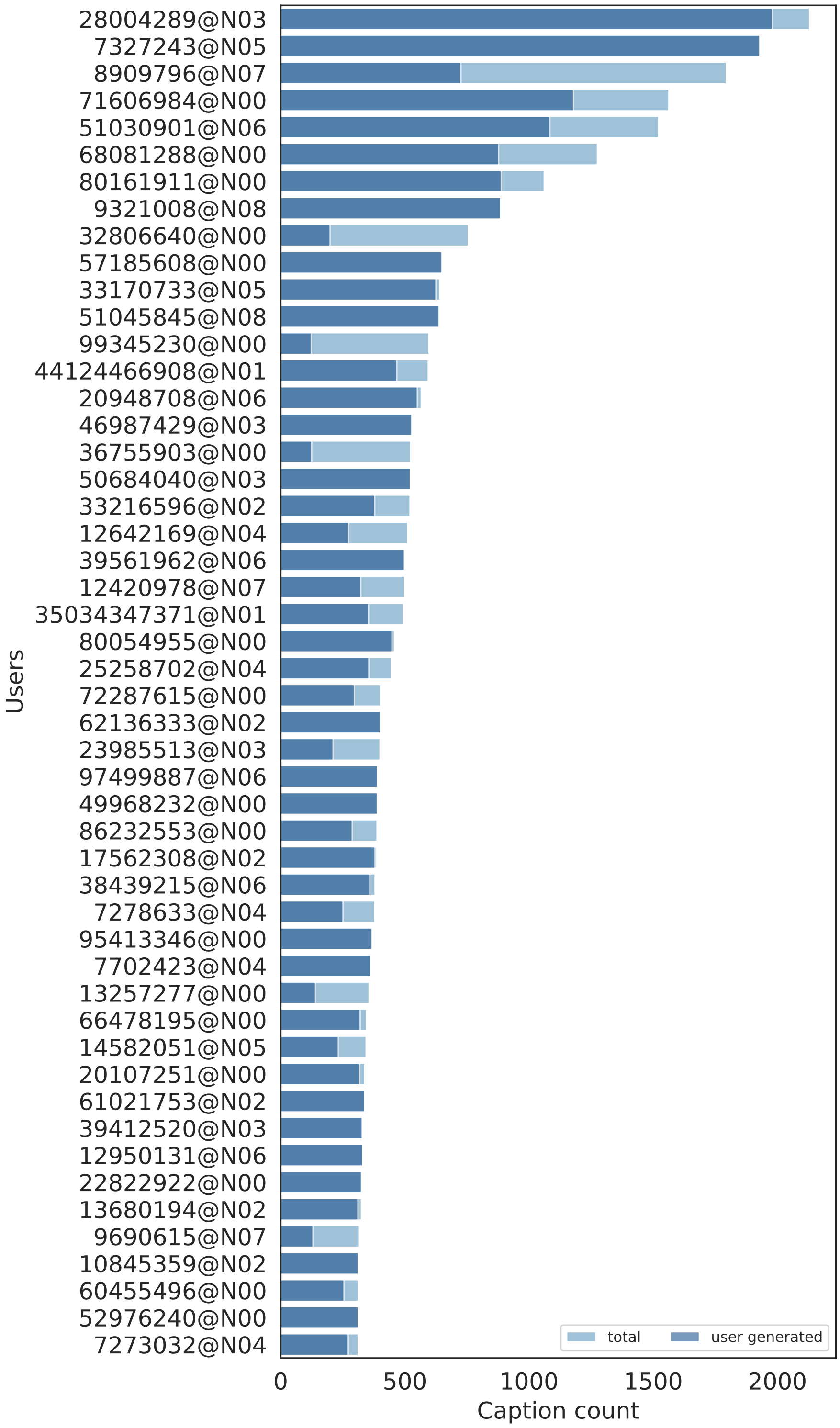


Figure 1.1: The distribution of user generated caption for 50 randomly sampled users on Flickr. On an average, the user generated captions dominate the text as compared to machine generated

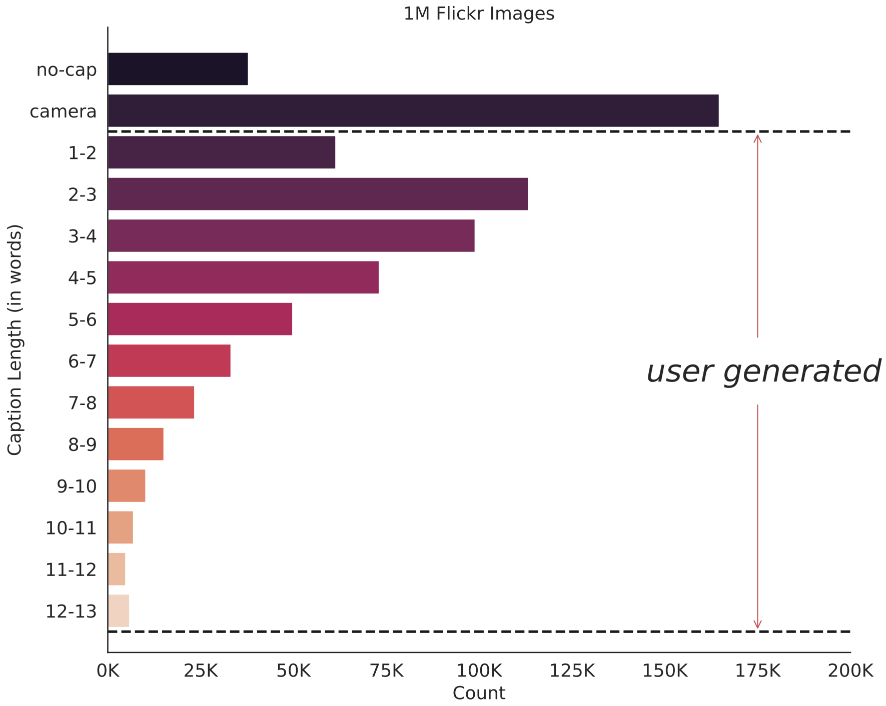


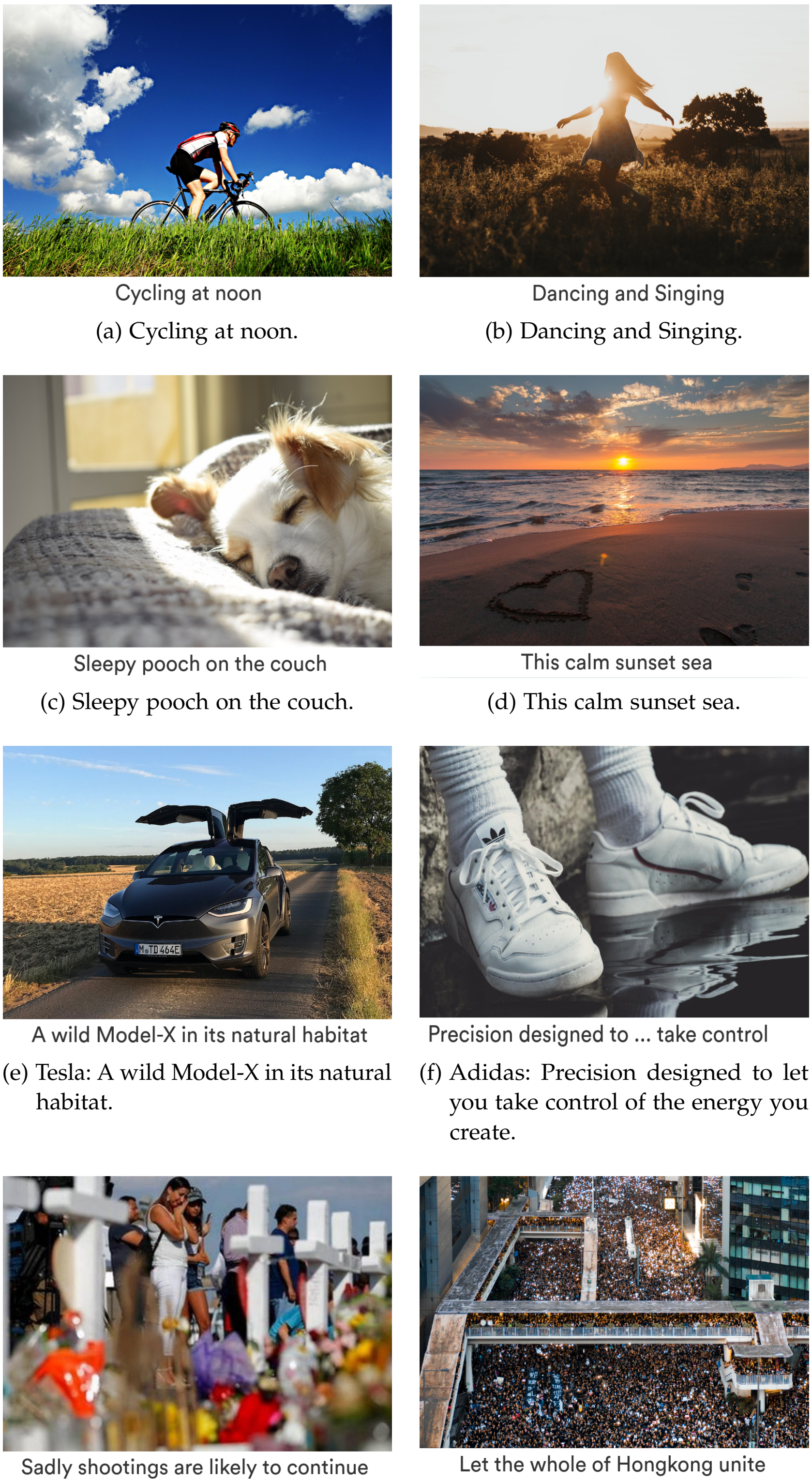
Figure 1.2: The distribution of caption lengths across 1 million randomly sampled Flickr images. If we exclude the camera generated (camera) and untitled (no-cap) images, we can see that an estimated 73% of images have a caption associated with them.

1.1.1 *Self Presentation in Social Media*

personality in social media

In 2019, an estimated 3.48 billion people are using social media worldwide, a reported increase of 9% every year[3]. The user base of social media is no longer limited to teenagers, but increasingly, $35 - 44$ years old have also populated the ranks of joiners and spectators. In the context of such large user base, social media postings, can be considered as an extension of Goffman's popular theory of *Presentation of Self* [3]. It states that, "During any social interaction, people often have a desire to control the impressions that others form about them". Extending this to the online world, users often create posts and use language that is consistent with ones desired identity [4]. For example, writing funny captions or uploading images of oneself in trendy outfits matching this desired identity. More often, this is done through voluntary self-disclosure. Thus choosing the style of language, hence captions, translate as the means to control the "impression management" across the social media. Figure 1.3, a and b, show examples of the same.

[3] https://blog.hootsuite.com/simon-kemp-social-media/, last accessed 15.07.2019


Cycling at noon
Dancing and Singing
Sleepy pooch on the couch
This calm sunset sea
A wild Model-X in its natural habitat
Precision designed to ... take control
Sadly shootings are likely to continue
Let the whole of Hongkong unite

(a) Cycling at noon.

(b) Dancing and Singing.

(c) Sleepy pooch on the couch.

(d) This calm sunset sea.

(e) Tesla: A wild Model-X in its natural habitat.

(f) Adidas: Precision designed to let you take control of the energy you create.

(g) Sadly shootings are likely to continue.

(h) Let the whole of Hongkong unite.

Figure 1.3: Examples of social media posts. These postings convey different messages like identity projections (a and b), regular captions (c and d), marketing (e and f), public opinions (g and h). These symbolic social media posts have been recreated by the author which reflect the real examples (to avoid copyright issues). Images used fall under the Creative Commons Zero License (used under CC0).

1.1.2 *Advertisement/Marketing*

With increased consumer presence across the Internet, business entities have increasingly started using the social media as a platform to market their products. Firms have employed dedicated internal/external social media teams to assist user engagement. The core method of *brand involvement* (creating positive feelings of attachment to a brand) revolves around increasing engagement of brand's social media posts. Brand images along with captions are used here to create a sense of intimacy with the consumers. Moreover, image captions are also used to highlight the product features in subtle ways. Figure 1.3, e and f, portray such marketing strategies employed by firms. These teams often employ software/bots to assist the engagement process [5, 6]. As part of this, automatic chat-bots also initiate conversations with users or write engaging comments to user uploaded images.

Advertisement methods used by brands

1.1.3 *Visually Impaired*

Describing images through language acts as an important aid for the visually challenged population. Visually impaired individuals often rely on software that read captions aloud in order to understand the visual contents. Such captions not only reveal objective information, but also enrich it by including subjective and sentiment dimensions. Such dimensions can help in magnifying details of entities present inside the image and highlight their individual attributes.

environment understanding

1.1.4 *Effective Communication*

Of all the applications, effective communication can be postulated as the most influential use of captions. Here, images along with captions play an important role in conveying sentiments/emotions which are otherwise difficult using a single modality. Users often use such multimodal methods in order to convey their strong opinions/resentments. These messages, often times, also highlight certain inherent problems that exist in their immediate surroundings/location. For example, the 2011 Egyptian revolution starting with Tunisia garnered widespread coverage through such communication methods [7]. Figure 1.3, g and h, show few examples of such communication.

public opinion through captions

1.2 AUTOMATIC CAPTION GENERATION

In the above sections we have seen the areas where image captions play an important role. Therefore, an enhanced human-like caption/description generation by machines has far reaching consequences. Machines can act as assistants to humans, providing them with different options with caption generation. Visually impaired individuals can

automatic captioning for enhanced applications use automatic machine generated descriptions as an aid in scene/image understanding [8, 9]. Automatic marketing bots can have effective communication with consumers making advertisement more efficient. A variety of other applications have already been developed around automatic image captioning. *Automatic Personal Diary*, which summarize ones vacations/trip photos is one such example [10]. Moreover, other fields like image retrieval, image indexing can also use this technology in order to enhance their existing algorithm. More importantly, the advances in this direction shall take machines one step closer towards efficient human interaction.

1.3 AFFECTIVE IMAGE CAPTIONING

move away from objective descriptions

Significant advances have been made in automatically generating image captions [11–16]. However, this line of research has primarily been focused on generating *factual/objective* description of images as captions. Although factual captioning has significant applications, this area of research ignores the human element of caption generation (as the core focus is on image description).

When humans write image captions, they are much more that mere factual description of images. Captions generated by humans often include different dimensions, contexts and more [17]. This field of automatic captioning has been less explored and therefore, shall be the main focus of this work. We shall call this field *Affective Image Captioning* to differentiate it from the above mentioned captioning research.

1.4 RESEARCH QUESTIONS AND GOALS

The main research question of this thesis is:
Q:

Can machines generate human-like image captions?

The above question can be decomposed into five sub-questions with corresponding goals:

- **Q.1** :

  What properties do human generated captions contain?

  Goal: Conduct a worldwide human involved crowd-sourcing study with user generated image captions. Collect human feedback on these captions (through a Questionnaire) and arrive at relevant properties through statistical analysis.

- **Q.2** :

  > *Can machines capture the subjectivity prevalent in the visual medium?*

  Goal: Check the feasibility of Deep Learning models to capture subjectivity in images. Arrive at a framework such that it can be further used for image captioning.

- **Q.3** :

  > *Can machines generate captions with a sentiment dimension?*

  Goal: Research the methodology of enhancing the sentiment of an image caption. Investigate the feasibility of using Deep Learning models to implement these methods. Evaluate these models quantitatively and qualitatively.

- **Q.4** :

  > *Can machines generate controlled variations in captions?*

  Goal: Investigate the feasibility of using Generative Deep Learning models for generating variations in captions. Additionally, design the models such that the properties of these variations can be controlled by the user. Provide quantitative and qualitative evaluation of these models.

- **Q.5** :

  > *Do we have representative non-biased datasets for the tasks?*

  Goal: Most datasets crawled through social-media have a high bias towards the positive sentiment. Investigate the availability of balanced datasets for the above mentioned sub-questions. If found unavailable, develop new datasets for the task and make them available to the community.

In the following section, we list-out our contributions w.r.t these questions.

1.5 CONTRIBUTIONS

The already published contributions of this work can be summarized into three broad areas: a) Human Captioning Study, b) Affective Captioning Models and c) Datasets

1.5.1 *Human Captioning Study*

crowd-sourcing

A worldwide study to evaluate human generated captions was performed over a span of 8 months. This study involved participants from countries where English is a native language. The study collected 15,000 responses and played a crucial role in evaluating properties of human generated captions. The study also revealed other interesting non-intuitive aspects of captioning behavior and set the platform for researching into captioning models to emulate these properties. The findings of this study are published in *ACMMM 2017* [17].

1.5.2 *Affective Captioning Models*

Deep learning models were developed and implemented to include human-like properties of Subjectivity, Sentiment, and Variations.

1.5.2.1 *Subjectivity*

aspect-detection

The Focus-Aspect-Value model (FAV) was designed to capture the subjectivity prevalent in image interpretations w.r.t focus. In this regard, a new task of *aspect-detection* was introduced into the community. Additionally, a novel method of information fusion called *Tensor Fusion* was also introduced. The Tensor Fusion method and the FAV model are published in ICMR 2019 [18] and the extended version has been accepted in journal IJMIR 2020 [19].

1.5.2.2 *Sentiment*

CAST/STEM

Two models were developed and evaluated as a part of the sentiment dimension for captioning:

- Concept and Syntax Transition Networks (CAST): A graphical model for sentiment captions.
- Show and Tell with Emotions (STEM): An end-to-end Deep Learning model for sentiment injection.

The above mentioned models scored high on the human evaluation track of the *ACMMM Captioning Grand Challenge* and *ICMR* [20, 21].

1.5.2.3 *Variations*

generate controlled variations

Generative models for introducing variation and sentiment into captions were developed using Generative Adversarial Networks (GAN). A novel way of context (sentiment) fusion using GAN and a *two-phase* training approach were also proposed to stabilize GAN training. It was also shown that using Policy Gradients algorithm from Reinforcement Learning, it is possible to overcome the diminishing gradients for GANs. The generative captioning model and the fusion methods were published in ICANN 2019 [22]

1.5.3 *Datasets*

Two important datasets were published as a part of this work. These datasets are open to the public.

1.5.3.1 *aspectsDB*

publicly available

As a part of the FAV model a new dataset was also introduced which aided training using this way of modeling. The dataset contains 155,539 images (with 19 aspects) and is balanced w.r.t the distribution of sentiment, which was found lacking in the previous datasets [18].

1.5.3.2 *captionsDB*

The evaluations and responses from the Human Captioning Study was compiled into a dataset to aid research in other areas like caption interpretation, user behavior analysis etc.
All datasets are publicly available online and can be downloaded at `https://madm.dfki.de/downloads`.

1.6 THESIS STRUCTURE

This thesis is organized into five major sections: Introduction, State-of-the-art, Human captioning study, Affective captioning models and Summary. The rest of the organization can be briefly explained as following:

1.6.1 *State-of-the-art*

background

Chapter 2 explains the background work and relevant algorithms required in the field of automatic image captioning. The chapter also summarizes the state-of-the-art research related to this field. General image captioning architectures and quantitative caption evaluation metrics are also discussed. A brief introduction to Reinforcement Learning (RL) is provided as a part of this chapter. RL algorithms have been used for generative approaches in the latter half of this work.

1.6.2 *Human Captioning Study*

human caption property

Chapter 3 describes the large scale human involved study conducted to evaluate properties of human generated captions. The details of the setup, experiment and statistical analysis are discussed here. The chapter concludes by providing important properties to pursue for affective image captioning.

1.6.3 *Affective Captioning Models*

subjectivity

Chapter 4 introduces the new task of aspect detection and the FAV model for subjective visual interpretation. A novel fusion approach called Tensor Fusion is explained as a method for implementing the FAV model. The experiment results prove that this method of information fusion outperforms the state-of-the-art Cross Residual Networks (XResNet) in aspect prediction.

sentiment

Chapter 5 describes how sentiment can be injected into image captions. Two relevant models are proposed that can generate image captions with high sentiment value. The evaluation of these models through human subjects are discussed including the ACMMM Captioning Grand Challenge.

variability

Chapter 6 explains the generative approach to image captioning. It also describes how Reinforcement Learning can be used to stabilize GAN training and generate variations in captions. It goes on to describe how a property (e.g., sentiment) can be fused inside a GAN to control the corresponding property in the generated captions.

application

Chapter 7 describes the architecture and organization of the real-time captioning demo *Captittude* using Tensorflow.

1.6.4 *Summary & Outlook*

Chapter 8 and Chapter 9 summarize the work and chart the future possible directions for the course of this research respectively.

1.6.5 *Appendix*

caption evaluation and crowd-sourcing lessons

Appendix A, describes the mathematical details behind the evaluation of captions. We derive the equations accompanying popular quantitative evaluation metrics. Appendix B lists the general best practices and cautions to be taken into account while setting up Crowd-Sourcing experiments.

Part II

STATE OF THE ART

2 BACKGROUND AND RELATED WORK

This chapter describes the state-of-the-art algorithms and models used in the field of image captioning. There have been many image captioning approaches proposed in the last decade. These methods range from basic detection/recognition based techniques to the present encoder-decoder models. Recently, generative models have also been proposed for image captioning. Other studies that have also tried to explore the non-factual aspects in the generated caption, e.g., sentiment. Therefore, the intention of this chapter is to introduce these architectures to the reader and make him aware of the state-of-the-art in this field of research.

chapter structure

The chapter is organized as follows: Section 2.1 describes the popular deep learning networks used in building image captioning models. Section 2.2 and Section 2.3 describe the basic approaches used in image captioning. Section 2.4 explains the neural network based encoder-decoder models. Section 2.5 discusses attention based approaches, where attention mechanisms are used to improve caption accuracy. Section 2.6 deals with the generative approaches. Here, the concepts from Reinforcement Learning are also discussed as many generative models use Reinforcement Learning algorithms for training. The common datasets and evaluation methods used in captioning are presented in Section 2.7 and Section 2.8 respectively. Section 2.9 gives an introduction to methods used for capturing subjectivity in images and Section 2.10 concludes the chapter with a summary.

2.1 DEEP LEARNING NETWORKS

The revolution in deep learning has been powered by two basic network architectures: Convolutional Neural Network (CNN) and Long Short Term Memory, Long Short Term Memory (LSTM). While CNN is mostly used for image processing, LSTM is used for sequence analysis and Natural Language Processing (NLP). In the sections below, we provide a short introduction to these architectures and their variants.

2.1.1 *Convolutional Neural Networks*

image processing

The architecture of a CNN is inspired by the visual cortex of the human brain, where individual neurons respond to stimuli only in the restricted region of the visual field. A CNN can capture the spatial and temporal dependencies of an image. It does this by successive *convolution* operations over the image. Briefly put, the network consists of two main components: feature learning and classification. Figure 2.1 shows a typical architecture of a CNN.

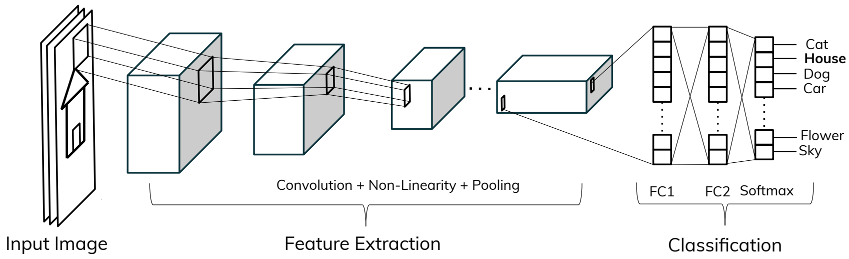


Figure 2.1: A CNN consists of two main components. The feature learning and the classification component. In the feature learning part, multiple convolutional layers are stacked on each other to learn features starting from basic ones (like edges) to more complex ones. The classification layer consists of stacked FC layers that assign the input to classes.

2.1.1.1 *Feature Extraction*

Features in a CNN are learned through successive convolutional layers. The level of abstraction learned increases as we go deeper with each convolutional layer. The first level of convolutional layers learn low level features like edges, corners etc. These features are then provided as input to succeeding convolutional layers which learn to combine these features to learn higher level of abstraction. In order to check values from exploding and to introduce non-linearity, the outputs from each layer are usually *clamped* using different mathematical functions. These are called as *activation* layers. Popular choices for these functions are sigmoid, ReLU etc. To reduce the spatial dimensions of intermediate representation of images, an operation called *pooling* is applied. The general choice for pooling is max-pooling, wherein, the largest value is selected as a representative of a part of an image. Therefore a convolutional layer refers to the combination of the three: convolution + non-linearity + pooling.

conv block

2.1.1.2 *Classification*

The classification block consists of flattening the convolution matrices and addition of FC layers to the network. The flattened matrices are converted into a vector and feed forward neural network. The intuition is that the features learned by the convolutional layers can be used by multiple FC layers to make a decision. The final layer in a classification block is the *softmax* layer, which converts the activations into a probability distribution over the label space.

decision block

2.1.1.3 *CNN Variants*

There are four popular variants of a CNN which have been used widely in image captioning research. These are: AlexNet, GoogLeNet, ResNet and VGGNet. The AlexNet architecture was developed by authors of [23] and started the deep learning revolution when it won the ILSVRC in 2012. Two years later, ILSVRC-2014 was won by another CNN variant introduced called GoogLeNet [24] introduced by Google. Another popular architecture called as VGGNet [25] was also introduced the same year. ILSVRC-2015 saw the introduction of very deep CNN architectures called Residual Network or ResNet [26] with up to 256 layers.

ILSVRC winners

2.1.2 *Long Short Term Memory Network*

Recurrent Neural Networks (RNN) are used to process sequential data because they can take previous information into account while processing the present task. This is done by providing the current output backwards as an additional input to the next step. Although this architecture works well for shorter sequences, the performance of RNN depreciates over longer sequences and deeper networks. The main challenge as discovered by [27] is called the *vanishing/exploding gradient problem*. Simply put, when the gradients are propagated backwards through a deep network they either explode or vanish due to repetitive multiplications. This makes RNN incapable of learning long term dependencies.

sequence processing

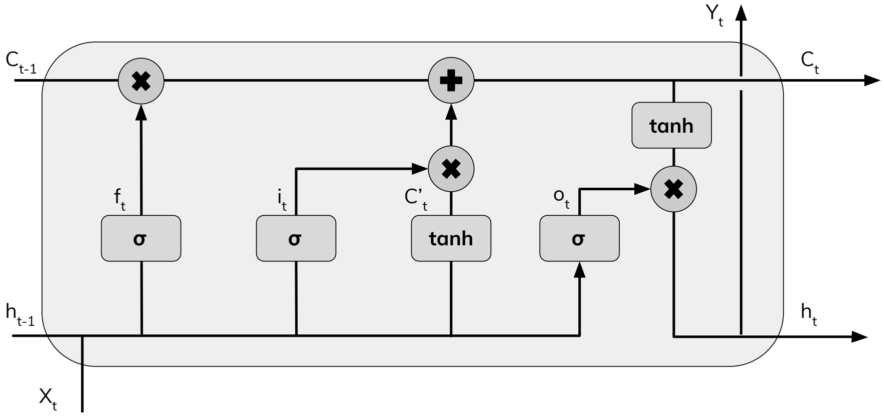


Figure 2.2: The LSTM cell architecture contains a cell state along with three gates: input, output and forget. Gates are sigmoid functions that control that flow of data through them. This allows the cell to remember long sequences and keep the memory over time-steps to take decisions.

In order to overcome this problem, the authors of [28] introduced the LSTM architecture. The simplest form of an LSTM contains a cell with three gates: input, output and forget. Gates inside an LSTM cell are sigmoid activation functions (meaning they output values in the range [0-1]). Therefore, they decide when data should be stored and given away from the cell. This represents a form of *memory* for the LSTM, which then can keep contexts over long sequences. The gates can be represented by:

gate mechanism

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i), \tag{2.1}$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f), \tag{2.2}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o), \tag{2.3}$$

where i_t, f_t and o_t represent the input, output and forget gate respectively w_x, b_x, represents the weights and biases of the corresponding gates. x_t is the input at timestamp t. h_{t-1} represents the input from the previous timestamp.

The cell state equations can be represented by:

$$c'_t = tanh(w_c[h_{t-1}, x_t] + b_c), \tag{2.4}$$

$$c_t = f_t * c_{t-1} + i_t * c'_t, \tag{2.5}$$

$$h_t = o_t * tanh(c_t), \tag{2.6}$$

where c'_t, c_t are the intermediate cell state and cell state respectively.

2.1.2.1 *LSTM Variants*

gate variations

Various architectures of LSTM have become popular over time. Authors of [29] introduced *peephole connections* inside the cell, allowing the gates to look at the cell state. Others combined different gates together. However, the most popular variant is the *Gated Recurrent Unit* introduced by the authors of [30]. In this architecture, the forget and input gates are combined into a single *update* gate. Additionally, the cell state and hidden state are also merged into one. The authors claim that this simplification provides an enhanced performance over regular LSTM cells in the task of machine translation.

Having discussed the basic neural network architectures used in image captioning, we move on to describing the evolution of image captioning approaches in the next sections.

2.2 BASIC CAPTIONING APPROACHES (TRIPLET BASED)

triplets

Some of the early approaches that were used with respect to image captioning were to first extract important objects from the image and then construct sentences based on these objects. In the work, [31], the authors first map each of the images to a meaning space which consists of $(object, action, scene)$ triplet and then use these triplets to generate the sentences describing images. This work is similar to [32] where again objects, visual attributes, and spatial relationships are first extracted as a phase selection step. These extracted entities are used to form a sentence in the phase fusion step. Another recognition based approach is used in [33] which first detects important objects, its attributes using classifiers and then generate the sentences using a Conditional Random Field. Other interesting studies using similar detection techniques are in the works [34–37].

2.3 basic captioning approaches (projection based)

The second set of methods bring the images and sentences into a single multi-dimensional space by converting each of them into vectors. Thereafter a set of distance measures are used to find the closest matching description of a given image ([38, 39]). The work of [39] uses neural networks to map images and sentences into the same vector space. Although the above mentioned methods have shown promising results they cannot be used for generating novel descriptions. Hence the performance of these methods drop when there are new compositions of objects in a given image (even though individual objects have been observed during training).

multimodal projection

2.4 encoder decoder captioning approaches

The state-of-the-art image captioning models today use the encoder-decoder setup [11, 15, 40–42]. The inspiration for the encoder-decoder setup comes from the advances made in the field of machine translation. The work in [43], called as the *seq2seq* model, paved way for this. In seq2seq model two LSTM networks are used for the purpose of machine translation. Briefly put, the first LSTM reads the source sentence and encodes it into a n-dimensional vector. The second LSTM uses this n-dimensional vector to decode it into the target sentence. Image captioning can also be considered as a translation problem. The difference here is that the translation is across multiple modalities, namely from *visual to text* modality. Authors of [11] were the first to use this architecture to generate objective image captions. This architecture is called as *Show-n-Tell*. In Show-n-Tell, a CNN is used as the encoder and a LSTM acts as the decoder. The CNN encodes the image into a n-dimensional vector (similar to seq2seq) model and the LSTM decodes this vector into a caption. Figure 2.3 shows a brief illustration of this architecture. Many other works have explored this technique with different degree of variation. In [44], the authors use a bidirectional RNN to not only generate the captions from the images but to also get the visual features (given a sentence describing the images). In [45], the authors explore the use of Diverse Beam search instead of normal search suggesting improved captioning quality. Other interesting studies using similar techniques are [46–49].

end-to-end neural network based

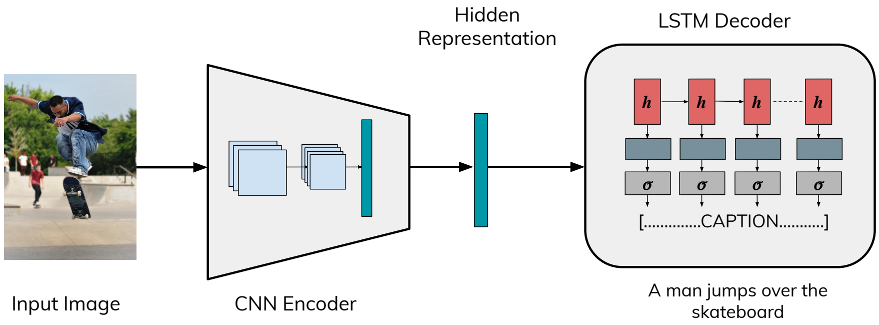

Hidden Representation
LSTM Decoder
h
h
h
h
σ
σ
σ
σ
[.............CAPTION..........]
Input Image
CNN Encoder
A man jumps over the skateboard

Figure 2.3: In the encoder-decoder model, the CNN takes the image as input and converts it into an n-dimensional vector based on the last FC layer. This n-dimensional vector as initial hidden vector for the LSTM decoder (h), which decodes it into a sentence/caption. Image used is from the publicly available Microsoft Dataset [84]

2.5 ATTENTION BASED CAPTIONING APPROACHES

focus based variant

Attention mechanism used in RNN is an idea loosely based on how humans observe a scene. When humans look at a scene, we focus on certain parts of the scene in "high-resolution" while other parts remain in peripheral "low-resolution". The brain uses this mechanism to attend to specific details and adjust the focal point across regions. Attention based methods for captioning have been proposed in [15]. Here an LSTM cell with attention is trained to look at certain part of the image while generating a particular word. The advantage here is that the generated captions are more precise as irrelevant parts of the image are discarded. Attention mechanism has also been combined with additional visual information in [50, 51]. Here, the authors use attention together with information from top-down (encoder-decoder approach) and bottom-up (detection based) approaches inside the RNN.

2.6 GENERATIVE CAPTIONING APPROACHES

Generative approaches have been have been quite successful in domains like image generation. These approaches work by learning the distribution of the training data and generating similar samples. Of the generative methods, the most successful have been the *Generative Adversarial Networks*. In the following section we discuss this architecture in detail.

2.6.1 *Generative Adversarial Networks*

Generative Adversarial Network (GAN) is a type of network which consists of two components: a *Generator* and a *Discriminator*. The task of the generator is to learn the distribution of the data and produce samples which are as close to the distribution as possible. The discriminator tries to identify the real data (from the training set) from the fake ones (which are generated by the generator). The training method of a GAN is called as *adversarial training*. During adversarial training, the two entities (generator and discriminator) compete against each other. The generator tries to fool the discriminator (by producing samples very close to the real distribution) whereas the discriminator tries not to get fooled. As these networks get trained, both of them improve until a point where the data generated by the generator is indistinguishable from the real data.

adversarial mechanism

In the regular setup of a GAN, the generator receives a noise variable $z \sim P_z(z)$, which the generator maps to dataspace as $G(z; \theta_g)$, where G is a differentiable function represented by the neural network and θ_g are its parameters. The discriminator represents another differentiable function $D(x; \theta_d)$ that outputs a single scalar value. The output of $D(x)$ represents the discriminators "belief" that x came from the training distribution or was generated by G. Therefore the discriminator is trained to maximize the probability of determining both true and fake samples correctly. The discriminators objective, O_D, can be represented by the following:

min-max game

$$O_D = \mathbb{E}_{x \sim P_{data}(x)} [log\, D(x)] + \mathbb{E}_{z \sim P_z(z)} [log\,(1 - D(G(z)))], \quad (2.7)$$

where, $P_{data}(x)$ and $P_z(z)$ are the distributions of data and the noise variable z respectively.

The generators objective is to fool the discriminator and generate samples as close to the training data. Therefore its objective can be defined as the following:

$$O_G = \mathbb{E}_{z \sim P_z(z)} [log\,(D(G(z)))], \quad (2.8)$$

If we combine Equation 2.7 and Equation 2.8, we realize that together they play a min-max game with the following value function, $V(G, D)$:

$$V(G, D) = \min_G \max_D \mathbb{E}_{x \sim P_{data}(x)} [log\, D(x)] + \mathbb{E}_{z \sim P_z(z)} [log\,(1 - D(G(z)))]. \quad (2.9)$$

generator fools discriminator

The convergence point of the training is when the generator learns to generate samples which the discriminator cannot assign a label correctly (real or fake). Figure 2.4 shows a brief illustration of this architecture.

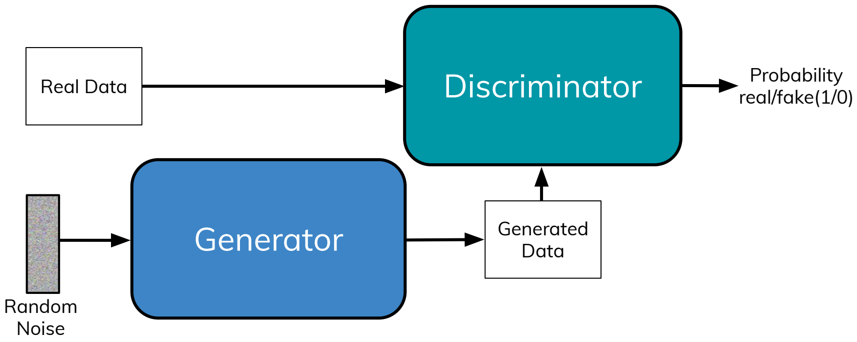


Figure 2.4: In the original setting of a GAN[52], the generator takes an n-dimensional random vector as input and generates data which is close to the original training distribution. The discriminator tries to differentiate between the generated data and the real one.

2.6.2 *Conditional Generative Adversarial Networks*

external conditionals

In a Conditional Generative Adversarial Network (CGAN), the architecture of GAN is extended by conditioning the generator and discriminator on some extra information. This information, y, can be any kind of external input, e.g., data from other modality, class label etc. This conditioning is performed by feeding y into discriminator and generator as additional input layer. This means that if we use a neural network for G and D, each of these networks would have an additional input of y. The modified value function can be represented as follows:

$$V(G,D) = \min_G \max_D \mathbb{E}_{x\sim P_{data}(x)}[log\, D(x|y)] + \mathbb{E}_{z\sim P_z(z)}[log(1 - D(G(z|y)))] \tag{2.10}$$

Having discussed the generative model with GAN and CGAN, we shall briefly describe the methods of Reinforcement Learning which are used to train generative models. The following section covers these methods of Policy Gradients and Monte-Carlo rollouts.

2.6.3 *Reinforcement Learning*

agent and environment

In a Reinforcement Learning (RL) setting, we have an *agent* and an unknown *environment*. The agent's current *state* is denoted by $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of states. To interact with the environment, the agent performs an *action*, $a \in \mathcal{A}$ (where $\mathcal{A}$ is the set of actions) in turn receiving a *reward*, $r \in \mathcal{R}$. This also causes the agent to change to a new state, $s' \in \mathcal{S}$. The goal of the agent is to maximize the cumulative reward by choosing the right set of actions for interacting with the environment (Figure 2.5 shows a brief overview of this setting). The agent's *policy* $\pi(s)$ provides the decision mechanism to choose the optimal action in a state in order to maximize the total reward. A policy usually contains a set of parameters θ and is generally denoted by $\pi_\theta(s)$.

The interaction between agent and environment is captured in discreet time steps $t = 1, 2, 3, \ldots, T$. If the state, action, and reward at a time-step t are denoted by S_t, A_t, and R_t, respectively, then we can describe the interaction sequence or *trajectory*, terminating in time T as : $S_1, A_1, R_2, S_2, A_2, R_3, \ldots, S_T$ (the reward for time-step i only arrives in time-step $i+1$). A transition step from state s to s', with an action a, providing a reward r is represented by a tuple, (s, s', a, r).

Each state s is associated with a *value function*, $V(s)$, which measures the expected amount of future rewards received by being in this state and following the corresponding policy. The goal of reinforcement learning is to learn both the policy and the value function.

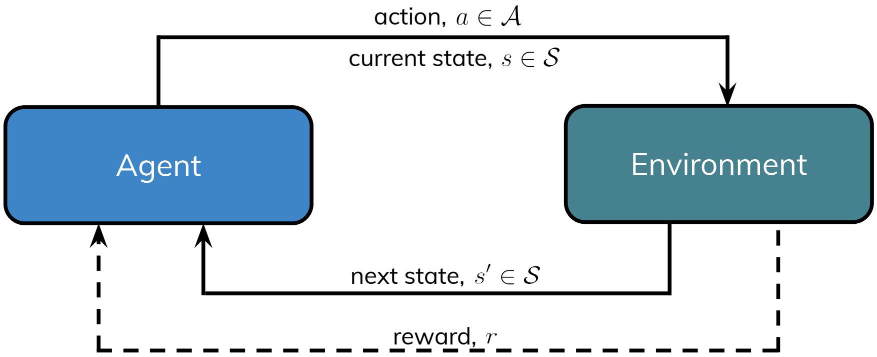

action, $a \in \mathcal{A}$
current state, $s \in \mathcal{S}$
Agent
Environment
next state, $s' \in \mathcal{S}$
reward, r

Figure 2.5: A reinforcement learning settings consists of an *agent* and an unknown *environment*. The agent takes an action a (from a set of actions $\mathcal{A}$) from current state s (from a set of states $\mathcal{S}$). As a result, the agent moves into another state s', collecting a reward r. The goal of the agent is to choose a set of actions such that it maximizes the cumulative reward.

2.6.3.1 *Policy Gradients*

Policy Gradients methods are one of the ways to optimize the policy parameters θ directly (in contrast to other methods which learn the state/action value function and then select actions accordingly). We begin by defining a reward function, $\mathcal{J}(\theta)$, as the expected return from a state. The goal of the algorithm, then, is to maximize the reward function.

optimizing for unknown environment

For the discrete space, $\mathcal{J}(\theta)$ can be defined as:

$$\begin{aligned} \mathcal{J}(\theta) &= V_{\pi\theta}\,(\mathcal{S}_1), \\ &= \mathbb{E}_{\pi\theta}\,[V1], \end{aligned} \tag{2.11}$$

where $\mathcal{S}_1$ is the initial state, V is the value function following the policy $\pi\theta$.

For the continuous space, $\mathcal{J}(\theta)$ can be defined as:

$$\begin{aligned} \mathcal{J}(\theta) &= \sum_{s \in \mathcal{S}} d_{\pi\theta}(s) V_{\pi\theta}(s), \\ &= \sum_{s \in \mathcal{S}} \left(d_{\pi\theta} \sum_{a \in \mathcal{A}} \pi(a|s, \theta)\, Q_{\pi}(s, a) \right), \end{aligned} \tag{2.12}$$

where $d_{\pi\theta}$ is the stationary distribution of a Markov chain w.r.t $\pi\theta$.

Optimizing algorithms like *gradient descent* can be used to find the best θ that produces the maximum value for $\mathcal{J}$ (highest return).

2.6.3.2 *Monte-Carlo Rollouts*

solution through simulation

The central idea behind Monte-Carlo methods is to learn by simulating trajectories/ episodes. This does not involve modeling of the environment and use the observed values as an approximation for the computed value. Therefore, to compute the value function of a state, $V(s) = \mathbb{E}[G_t|S_t = s]$, Monte-Carlo method would simulate different episodes starting form state s: $S_1, A1, R2, \ldots, S_T$ many times and compute G_t and $V(s)$ as following:

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}, \tag{2.13}$$

$$V(s) = \frac{\sum_{t=1}^{T} 1[S_t = s]\, G_t}{\sum_{t=1}^{T} 1[S_t = s]}, \tag{2.14}$$

where $1[S_t = s]$ is the Kronecker delta, R_t is the reward for time step t, γ is a discounting factor.

A similar approach is followed to learn the optimal policy with Monte-Carlo through iteration.

2.6.4 *GAN for Image Captioning*

language space non-differentiable

Traditionally, GAN have been used to generate images by learning a distribution from the image space of training set. In the case of an image, learning this distribution through back-propagation is a continuous mapping (which is differentiable) and therefore, can be done in an effective manner. However, generating languages form a discrete space which is non-differentiable. This makes language generation through GAN a challenging task. To alleviate these challenges, researchers often use methods from RL. Policy Gradients (discussed in the above section) which is popular RL optimization method is often employed in such scenarios [53–55]. While applying policy gradients for language generation in GAN, language is considered as a sequential decision making process. The generator is treated as a policy network and each token generation as an action. The discriminator becomes the evaluator which provides a reward for each action.

Authors of [53] were the first to apply policy gradients on an unconditioned GAN. The work in [55] extends it to CGAN, with image as the external condition. Similar approaches with attention mechanism have also been used in [56] where the generator uses hierarchical attention. RL methods have also been used for language generation without a GAN in [57–62]. However, these architectures and specific methods are out of the scope of this work.

2.7 CAPTIONING DATASETS

High quality datasets which can act as benchmarks (for researchers to evaluate their methods) are important in all fields of research. For image captioning, Microsoft Common Objects in Context (MSCOCO) is the dataset used to evaluate new captioning methods. Although, there have been smaller datasets like Flickr 8K and Flickr 30k [63], these datasets have faded away from use and are no longer considered for performance benchmarks. Off late, other datasets like YFCC100M have also gained popularity among researchers. We discuss these datasets in the following sections.

2.7.1 *MSCOCO*

objective captions

MSCOCO image captioning dataset was released by Microsoft in 2015. The images were collected by searching for pairs of eighty object and scene categories in Flickr. The objective was to collect images with multiple objects along with natural context. The dataset was generated using human subjects on Amazon Mechanical Turk (a crowd-sourcing platform). The subjects were given specific instructions to describe the image shown:

1. Describe all the important parts of the scene.
2. Do not start the sentences with “There is”.
3. Do not describe things that might happen in future or past.
4. Do not give people proper names.
5. The sentence should contain eight words.

In the final dataset, each image contains three different captions related to it. There are 1,026,459 training images along with 40,504 validation and 40,775 test images. The dataset can be downloaded at `http://cocodataset.org/`

2.7.2 *YFCC100M*

Flickr captions

YFCC100M was created by Yahoo as a part of their Webscope program. It is the largest collection of multimedia with 99.2 million images and 0.8 million videos. This multimedia data was collected from Flickr which were uploaded between 2004 and 2014, published under the Creative Commons license. The unique aspect of YFCC100M is its user-generated content. This implies that the labels and descriptions are unmodified and created under unrestricted scenarios (in contrast to MSCOCO discussed above). Therefore captions and tags contain natural descriptions and slang which are used by Flickr users. User generated content data also comes with additional challenges. For example, the data tends to have high noise content, in terms of non-English wordings, missing images etc. Therefore, using YFCC100M always mandates an additional step of noise removal. The dataset can be downloaded at https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67

2.8 CAPTION EVALUATION METHODS

Image captioning experiments are evaluated both quantitatively and qualitatively. While quantitative evaluation uses the evaluating standard metrics against the test set, the qualitative evaluation involves using humans to look at the generated captions and provide feedback.

2.8.1 *Quantitative Evaluation*

n-gram matching

The quantitative evaluation of image captions is done by comparing the generated captions against the ground truth captions. In general, these metrics aim at evaluating how close the generated captions are to the ground truth. This is done by matching n-grams of the two captions (generated and ground truth). There are four important metrics which are used for this: Bilingual Evaluation Understudy (BLEU) [64], Recall Oriented Understudy for Gisting Evaluation (ROUGE) [65], Metric for Evaluation for Translation with Explicit Ordering (METEOR) [66], and Consensus-based Image Description Evaluation (CIDER) [67]. The BLEU metric calculates the n-gram precision between two sentences and has shown good performance over corpus-level matching where a good number of n-grams usually match. However, for sentence-level matching it performs poorly. The ROUGE is a set of metrics for evaluating text summarization algorithms. Briefly put, it performs an n-gram recall over ground truth text. METEOR extends the BLEU idea with an addition of n-gram matching including synonyms into account. The CIDER metric performs a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each of the n-grams.

2.8.2 *Human Evaluation*

In order to provide a qualitative evaluation of image captions, humans are involved into the loop. This is commonly accomplished through crowd-sourcing platforms. These platforms provide annotators who can provide feedback on the generated captions in return for a fee. The evaluation questions along with image captions are uploaded to these platforms and an odd number of annotators (generally three) are asked to evaluate the questionnaire. The right answer is decided by the majority vote. Such platforms also provide methods for quality control like annotator ratings and validation questions. However, researchers also design their validation methods on the answers provided in order to ensure high quality results. The common crowd-sourcing platforms that are used in the research community are *Amazon Mechanical Turk*[1] and *CrowdFlower*[2].

qualitative feedback

2.9 SUBJECTIVITY IN IMAGES

There have been many promising approaches which researchers have employed at detecting subjective parts of visual interpretation. While some works focused on attributes to enhance the quality of nouns [68–70], others focused on understanding the aesthetics [71, 72]. Authors in [70, 73] proposed the large scale visual sentiment ontology to detect adjective-noun pairs inside an image. Given an image, they propose to find a suitable adjective-noun pair to best describe an image from a set of adjective-noun pairs. Although adjective-noun pairs capture the sentiment to an extent, they do not reveal the degree to which this sentiment applies. Moreover, relying on a single adjective-noun pair to describe the whole image would mean only the most prominent noun is focused upon.

holistic approaches

The work in [74] proposes a cross-modal mapping from a visual semantic space onto a linguistic space in order to automatically annotate images with adjectives. The mapping is performed by a projection function that maps the vector representation of an image tagged with an object/attribute onto the linguistic representation of the object/attribute word. This mapping function can then be applied to any given image to obtain its linguistic projection. The main advantage claimed is that of zero-shot learning, i.e., unseen attributes (not present in training) can be predicted. However, in this approach the whole image is mapped onto an adjective without focusing on any particular noun or aspect.

[1] www.mturk.com

[2] www.crowdflower.com

2.9.1 *Detecting Adjective-Noun Combinations*

Chapter 4 of this work builds on a line of work originating from the Visual Sentiment Ontology [75], which aims at detecting adjective-noun combinations from images. So far, the best performing method within this direction are cross-residual networks (XResNet) [76]. For any given image, XResNet outputs scores for adjective-noun combinations as well as scores for all individual adjectives and nouns separately. This means that it separates the more subjective parts of interpretation (represented by the adjectives) from the more objective (represented by the nouns) ones.

There are two major datasets that have been used for training the above-mentioned architectures: The Visual Sentiment Ontology (VSO) [75] and the Multilingual Visual Sentiment Ontology (MVSO) [77]. These datasets have been created from the popular photo-sharing platform Flickr. However, the data in these cases suffers from a clear bias towards the positive attributes/adjectives [78]. In the work that introduces XResNet [76], some efforts have been taken for achieving a better overall balance, however, for any given noun the number of associated adjectives is typically very small and the distribution heavily skewed. More importantly, the "feasible" adjectives for a given noun are in most cases not mutually exclusive. At times, adjectives that come with the same noun are even similar in meaning (e.g., "smiling person" and "happy person"), and yet, predicting any adjective that is not identical to the ground-truth adjective is typically considered wrong. For example, "smiling" and "sad" would count as equally bad if the ground truth was "happy". This makes it harder to interpret performances on these datasets in terms of ability to capture subjective aspects.

adjective-noun pairs

2.9.2 *Attribute Datasets*

There are several popular attribute datasets available for computer vision research. The Visual Genome [79] contains over 10,000 images with fine-grained annotations, including region descriptions, object instances and visual attributes in the order of millions. However, the attributes in this dataset mostly relate to objective information. Hence, most common attributes are colors like white, blue red, black .Despite the large number of total annotations in Visual Genome, it was found that the number of subjective attribute instances is too low for the purpose of Chapter 4. aPascal and aYahoo [80] are two attribute datasets containing natural object-based images with attribute annotations. Here again, the included attributes correspond to objective features, such as parts of a face like eyes, nose and so on, which deems it inappropriate for analyzing subjective interpretation. Another attribute dataset is the SUN Attribute Dataset [81], which contains scene at-

objective attributes

tributes of the four categories “functions / affordances” (e.g., “diving”, “climbing”), “materials”, “surface properties” and “spatial envelope”. The former three categories are restricted to objective information, and while there are several subjective attributes (such as “scary” or “stressful”) in the “spatial envelope” category, all of these annotations are describing the scene in a holistic manner.

2.10 CONCLUSIONS

The methods used to generate automatic image captions have evolved from basic template approaches to complex deep learning methods. The chapter began with a discussion of basic approaches that exist in image captioning. The basic approaches were soon replaced with the end-to-end encoder-decoder models with neural networks. We also discussed generative captioning models with GAN. In this regard, RL approaches of policy gradients was also introduced as they are used for training generative models for captioning. The major datasets for captioning, namely, MSCOCO and YFCC100M we also discussed. Quantitative caption evaluation metrics: BLEU, ROUGE, CIDER and METEOR were briefly introduced. Qualitative caption evaluation techniques using human-in-the-loop with crowd-sourcing platforms were also discussed.

Finally, the methods for capturing subjectivity were introduced along with the existing datasets in this field. Having provided the reader with an overview of the field of image captioning and subjectivity, we now move on to the main contents of this thesis in the upcoming chapters.

Part III

HUMAN CAPTIONING STUDY

3 HUMAN CAPTIONING ANALYSIS

This chapter presents the large-scale study performed in order to understand relevant properties of human generated captions. To accomplish this, 3,000 image caption pairs were randomly sampled from the social media platform Flickr[1]. These image/caption pairs were evaluated by native English speakers across the world. The evaluations spanned across different dimensions of these image/caption pairs like intent, visibility, focus, user-preference and caption-clarity. Each pair was evaluated by 5 individuals and the decision was taken with a majority vote. The procedure of this study and the statistical analysis are discussed in detail in the following sections.

publication & chapter structure

The results of the study presented in this chapter have appeared in ACMMM 2017 [17]. The rest of the chapter is organized as follows. Section 3.1 introduces the motivating problem of the chapter. Section 3.2 talks about the dataset used in this experiment. Section 3.3 describes the experiment settings, the task and annotator background. Section 3.4 analyzes the results both qualitatively and quantitatively. Section 3.5 concludes the chapter by summarizing the task.

3.1 PROBLEM DEFINITION

find human-like properties

Humans use a wide variety of ways to describe images using text. These textual descriptions or captions often show large variations in terms of styles, intent, length, focus etc. This implies that there is no right or wrong way to caption an image since it depends on the perspective of the user. Given such variations, it would be interesting to find if there are few key properties that define human generated captions. If such prominent properties exist, it would be beneficial to study and understand them for automatic image captioning. The benefits of such a study are two-fold:

- First, in order for machines to emulate human captions, it is important to first understand the properties of human generated captions. If such underlying properties are statistically studied and established, we can develop algorithms for machines to emulate these properties. Moreover, a better understanding of human captioning behavior will also enable machines to enhance their interaction with humans in multi-modal environments such as social media platforms.

[1] https://www.flickr.com/

- Second, with the surge of Deep Learning, creating representative datasets have become an important part of research. Therefore, statistics related to human captioning behavior are potentially useful in generating such datasets for automatic captioning research.

Therefore, our motivating question is:

How does "captioning in the wild" actually look like?.

More precisely, how can we formally capture the variety prevalent in human generated captions?

3.2 DATASET

To perform this study 3,000 random image captions prevalent across the social media platform, Flickr, were selected. The Flickr platform, given its global presence and hosting a large user base, is a popular choice among researchers in the automatic captioning community [82].

natural captions

In order to collect user generated image captions, the YFCC100M[2] dataset was used. The YFCC100M dataset [83] contains Flickr captions taken between the period 2004 – 2014. It contains 99.2 million images and 0.8 million videos. Given this large number, YFCC100M can be considered as a representative example of Flickr. Additionally, all this data is user-generated. This feature makes it unique and differentiates it from other datasets like MSCOCO [84] or Flickr30k [63]. MSCOCO and Flickr30k were created in a controlled environment with the help of paid annotators, making these datasets less suitable for human caption analysis.

noise

However, the inherent diversity of YFCC100M also comes with its own challenges. For example, as it is user-generated, the content includes high amount of noise in the form of camera generated titles, non-existent titles, spams etc. Since Flickr also has a global presence, it means that non-English titles are also present to a great extent [85]. This implies that the dataset needs to be cleaned in order to weed out unnecessary noise. Figure 3.1 shows a few samples of images present in the dataset.

3.3 EXPERIMENT

Studies that collect human responses need to be carefully planned and executed. This is because, annotators are easily susceptible to unintended biases and they may come from different cultural/social backgrounds. Moreover, it is also important to verify that evaluators

[2] https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67, last accessed 15.07.2019

have understood the task correctly, without unintentionally priming them (in the process of explaining too much). Furthermore, on Crowd-Sourcing platforms, it is also important to ensure that annotators do no cheat or collude in order to maximize their payments. To alleviate these issues, we have followed the best practices as mentioned in [86] and [87] (see Appendix B). The following sections describe the data cleansing and analysis in detail.

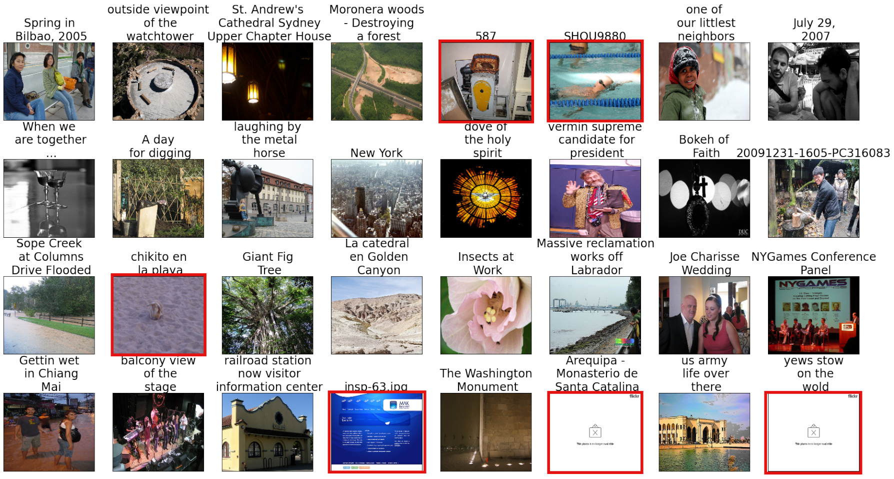

Spring in Bilbao, 2005
outside viewpoint of the watchtower
St. Andrew's Cathedral Sydney Upper Chapter House
Moronera woods - Destroying a forest
587
SHOU9880
one of our littlest neighbors
July 29, 2007
When we are together ...
A day for digging
laughing by the metal horse
New York
dove of the holy spirit
vermin supreme candidate for president
Bokeh of Faith
20091231-1605-PC316083
Sope Creek at Columns Drive Flooded
chikito en la playa
Giant Fig Tree
La catedral en Golden Canyon
Insects at Work
Massive reclamation works off Labrador
Joe Charisse Wedding
NYGames Conference Panel
Gettin wet in Chiang Mai
balcony view of the stage
railroad station now visitor information center
insp-63.jpg
The Washington Monument
Arequipa - Monasterio de Santa Catalina
us army life over there
yews stow on the wold

Figure 3.1: Illustration of randomly sampled image caption pairs from the YFCC100M dataset to show the variation in captions. The captions found here vary from something as simple as a number (e.g., "587") to complex subjective interpretation (e.g.,"when we were together"). The empty white images on the bottom row constitute one class of noise images present in the dataset wherein the user has removed the corresponding image. The images marked with a red boundary represent some examples of noise in the dataset for this task

3.3.1 *Data and Filtering*

Given the amount of noise that comes with user generated data, it is imperative to weed out the noise. In case of Flickr, this is generally done by using predefined filters which take out unwanted captions. Since the YFCC100M also contains a variety of languages, an additional language filter was also required. Figure 3.2 shows few examples of the noisy data. The well known filter published by the MM Commons Yahoo-Flickr Grand Challenge[3] is the general choice used for noise removal in captions. Briefly put, this filter performs the following:

noise filter

- All punctuation characters are removed.
- All letters are converted to lower case.

[3] https://multimediacommons.wordpress.com/tag-caption-prediction-challenge/, last accessed 20.06.2019

- If the caption has less than 5 words it is removed.
- If any of the words in the caption do not appear in an English dictionary, the caption is dropped.

However this filter was designed for a specific purpose in the context of the grand challenge. Therefore, it still lacks additional rules to address the noise specific to YFCC100M. To address this challenge, an additional enhanced filter was designed which was inspired from the grand-challenge filter. This enhanced filter removed the noise using the following conditions:

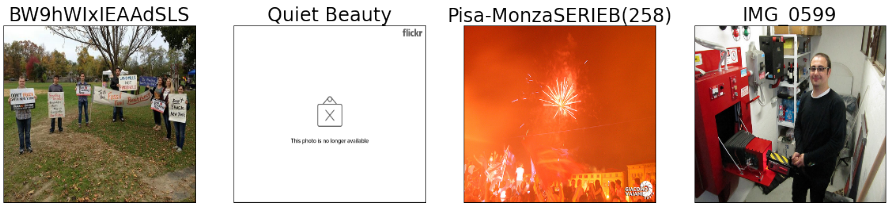


Figure 3.2: A few sample images which represent the noise for this task. From random-letter captions in the left, to image not found errors, to camera generated ones in the right.

elimination criteria

- *empty*: The caption is an empty string.
- *generic*: The caption is generic (e.g., IMG_321). This is checked using a regular expression.
- *short*: There are less than n (threshold) words in the caption. (For this study a threshold of 2 was used.)
- *non-en*: The caption is not English. To compute this, we remove most punctuation characters and for the resulting string check whether the fraction of words that are contained in the enchant dictionary [88] is greater than or equal to a given threshold. (Set to 0.5 for this study.)
- *numeric*: The main part of the caption is made of numeric expressions, such as single numbers or dates.
- *valid*: All captions that do not fall into any of the previous categories are considered to be useful and pass the filter.

It should be noted that, the original caption is kept unchanged even though during the intermediate steps it is modified.

3.3.2 *Crowd-Sourcing Task*

In order to engage a world-wide audience in our annotation task, we chose the Crowd-Sourcing platform CrowdFlower[4]. Such platforms

[4] https://www.crowdflower.com/

allow human evaluators to be involved in the task and provide an evaluation for a remuneration in return. The task involved annotation of 3,000 image/caption pairs from YFCC100M. Each of these image/-caption pairs were evaluated by 5 different human evaluators. The annotations were collected for the following questions for each pair:

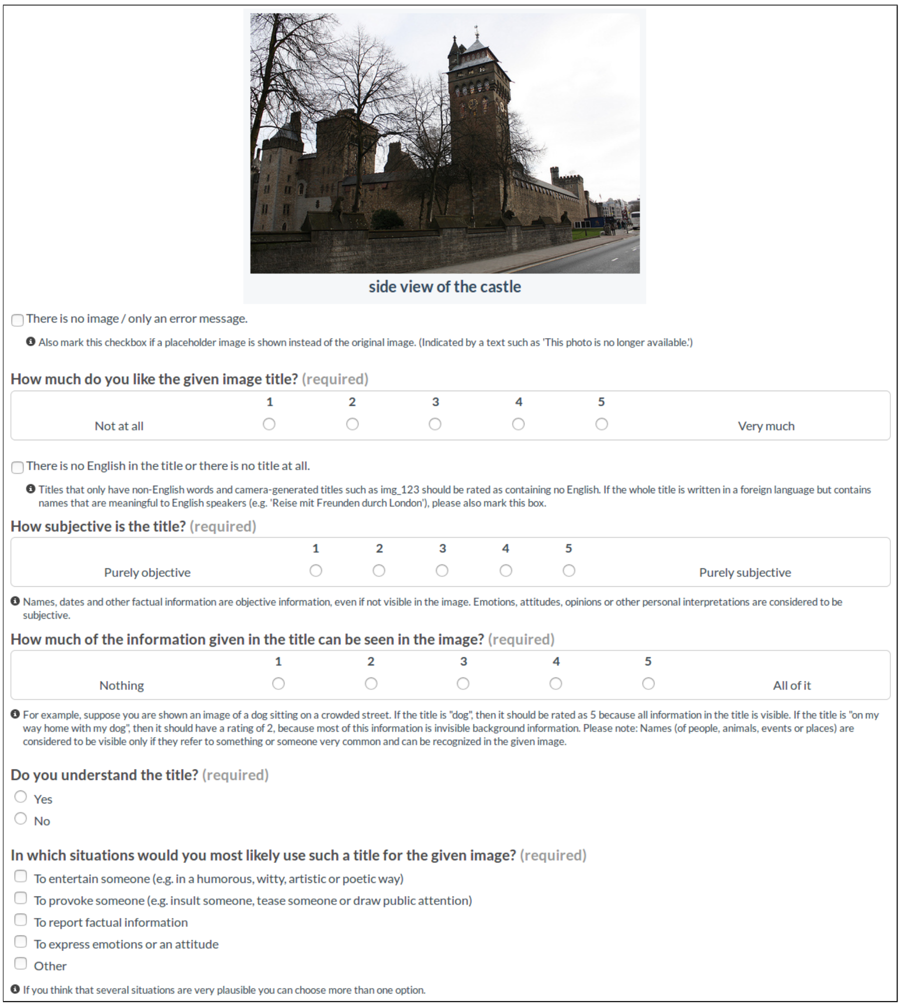
caption evaluation

- *Image error*: Some images become unavailable and answering the remaining questions is not very useful for these cases.
- *Preference*: Annotators are asked how much they personally like the given image title on a scale from 1 (*not at all*) to 5 (*very much*).
- *No English*: Non English captions which have escaped through the filter.
- *Subjectivity*[5]: The question "How subjective is the title?" is answered by rating the image title on a scale from 1 (for *purely objective*) to 5 (for *purely subjective*). High subjectivity indicates that there are many other options for captioning the given image.
- *Visibility*[6]: The question "How much of the information given in the title can be seen in the image?" is answered by rating the image title on a scale from 1 (for *nothing*) to 5 (for *all of it*). Information from the title (e.g., people, objects, famous landmarks etc.) is considered to be *visible* if it can be directly identified in the image. Among other things, visibility is important for computational feasibility.
- *Understanding*: For being able to filter out captions that were not understood by our annotators, we added a field for specifying a lack of understanding.
- *Intent*: The last task is to specify in which situations one would most likely use such a title for the given image. This is done with respect to the intent categories "to *entertain* someone" (e.g., in a humorous, witty, artistic or poetic way), "to *provoke* someone" (e.g., insult someone, tease someone or draw public attention), "to report *factual* information", "to *express* emotions or an attitude" and "*other*". This information is particularly useful when filtering for a specific image captioning purpose.

The image/caption pairs were randomly shuffled and presented to human evaluators. A screenshot of the interface for the complete task is shown in Figure 3.3.

[5] By *subjective* we mean based on or influenced by personal feelings, tastes, or opinions, as opposed to *objective* which relates to facts such as names, dates or other factual information.

[6] Famous landmarks (e.g., London Bridge, Statue of Liberty, Eiffel Tower etc.) can be considered as visible since they can be identified by most of the people, whereas for example pet names are non-visible information since only a few people are able to identify the pets.



side view of the castle

□ There is no image / only an error message.

Also mark this checkbox if a placeholder image is shown instead of the original image. (Indicated by a text such as 'This photo is no longer available.')

How much do you like the given image title? (required)

Not at all 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very much

□ There is no English in the title or there is no title at all.

Titles that only have non-English words and camera-generated titles such as img_123 should be rated as containing no English. If the whole title is written in a foreign language but contains names that are meaningful to English speakers (e.g. 'Reise mit Freunden durch London'), please also mark this box.

How subjective is the title? (required)

Purely objective 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Purely subjective

Names, dates and other factual information are objective information, even if not visible in the image. Emotions, attitudes, opinions or other personal interpretations are considered to be subjective.

How much of the information given in the title can be seen in the image? (required)

Nothing 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ All of it

For example, suppose you are shown an image of a dog sitting on a crowded street. If the title is "dog", then it should be rated as 5 because all information in the title is visible. If the title is "on my way home with my dog", then it should have a rating of 2, because most of this information is invisible background information. Please note: Names (of people, animals, events or places) are considered to be visible only if they refer to something or someone very common and can be recognized in the given image.

Do you understand the title? (required)

○ Yes
○ No

In which situations would you most likely use such a title for the given image? (required)

□ To entertain someone (e.g. in a humorous, witty, artistic or poetic way)
□ To provoke someone (e.g. insult someone, tease someone or draw public attention)
□ To report factual information
□ To express emotions or an attitude
□ Other

If you think that several situations are very plausible you can choose more than one option.

Figure 3.3: The Crowd-Sourcing task as seen by an annotator. Annotators arrive at this screen after having passed a relevant test set of similar questions. This ensures that the annotators have understood the task well.

3.3.3 *Annotation Quality Assurance*

two stage testing

The experiment ran for a span of 6 months and had a total of 298 annotators participating (through the full questionnaire) at various stages. Each annotator was allowed to annotate a maximum of 100 captions. The task involved reading and understanding different dimensions present in captions. Given the difficulty level of the task, the annotators were given detailed instructions at the beginning. This was also supplemented with an articulate explanation of each field and an exemplary image/caption pair was also provided (for each field) in order to aid their understanding.

Having gone through the instructions and examples, annotators were asked to take a set of test questions designed to evaluate their understanding of the task. Only participants who scored satisfactorily were allowed to proceed to the actual task.

To ensure that a sufficient annotation quality was maintained throughout the annotation task, test items were intermixed with the actual items and annotators who went above a threshold of error rate on the test items were dropped. The annotators were also required to be from English speaking countries. Figure 3.4 shows the distribution of annotators on the world map. Each item was evaluated by 5 different annotators.

3.3.4 *Annotator Background*

CrowdFlower provides basic information about the annotators like their country, region and city. To augment this information, we asked the annotators to answer additional questions about their age, nationality, educational level, profession/job, and level of English. The vast majority of annotators were from the USA within the age range of 20 – 40 (Figure 3.4, Figure 3.5). Close to 33% were college graduates or held a bachelor's degree. With regards to the profession, students, teachers, and homemakers constituted the majority. Figure 3.5 shows the distributions w.r.t the highest education and age of the annotators.

demographics

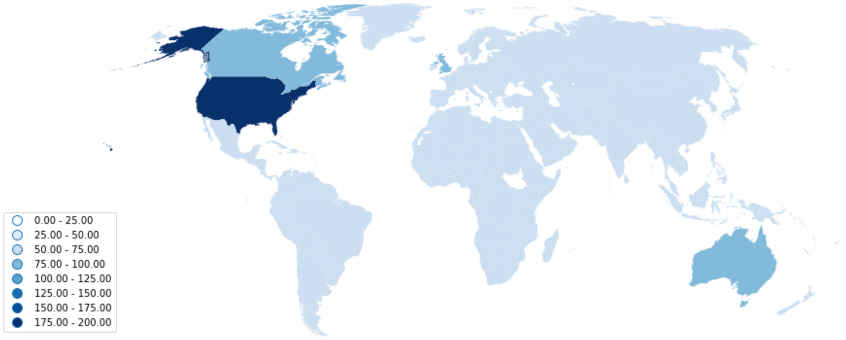
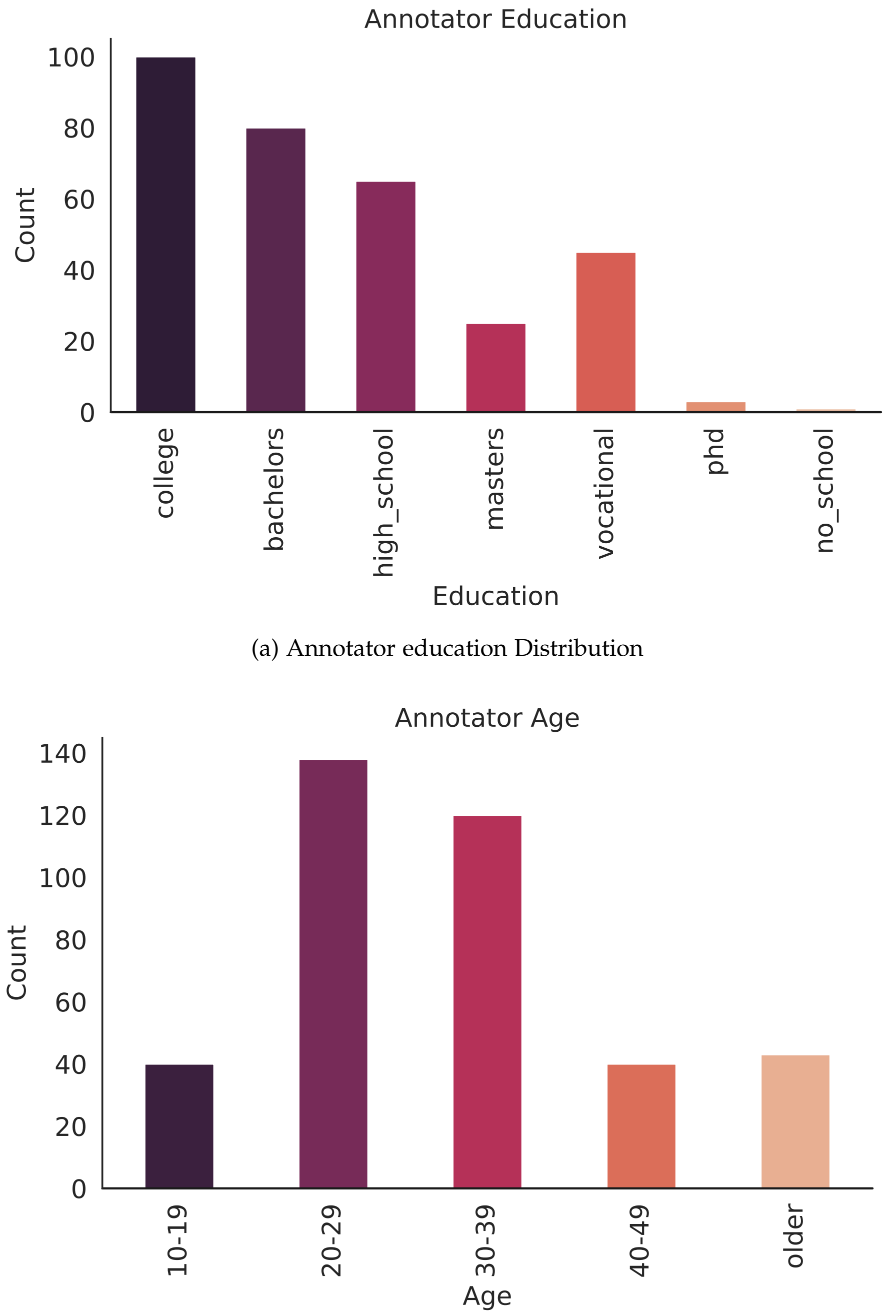


Figure 3.4: Annotator distribution normalized on the world map. Annotators were selected from countries where English is the native/majority language. These countries were the USA, England, Ireland, Scotland, New Zealand, Australia. The majority of the annotators came from the USA.

3.4 ANALYSIS

The analysis of user evaluations reveal certain interesting/non-intuitive perspectives of human captioning. First of all, captioning itself can be understood in several ways and depending on this understanding, the caption will be shaped accordingly. For example, it was found that many captions are titles in a rather "classical sense", i.e. they consist of only a few words which could be understood as a distinguishing name for the image. Names of places or people are often used to label



(a) Annotator education Distribution

(b) Annotator age distribution

Figure 3.5: Education and Age distributions of annotators. Majority of the annotators fall within the age range 20-40 and have a college/bachelors degree.

the image in this way, but more artistic names are not very uncommon as well.

Similarly, image titles can be used to merely (objectively) summarize what is in the image. This is essentially a translation process from vision to language and captions of this type have been the focus for the vast majority of papers in the field of automatic image captioning [11–13, 16]. Additionally, image captions can be considered to be a part of a multi-modal message consisting of text and image. Here, captioning is part of a communication process. So, in this setting, captioning an image would mean to find a suitable phrase or short sentence that together with the image forms a single message and conveys some desired meaning. This meaning could be a subjective interpretation of the image but captions that provide context information (such as names of places or what happened just before the image was taken) or guide the viewer towards a certain interpretation of the image (e.g., by mentioning certain aspects of the image one would not typically see at first glance) also fit nicely into this way of modeling. It should be noted that these interpretations are strongly related to different purposes or intents the author might have for creating the caption.

captions as communication

Apart from these, the properties observed in human captions can be classified into three broad categories: a) Subjectivity Space, b) Sentiment Space and c) Variations Space.

major spaces

3.4.1 *Subjectivity Space*

The subjectivity space is the result of analysis of answers provided for the visibility and focus related questions. These are observed as variations in interpretations of the image. There are two sources of subjectivity that have been seen in captions. The first type of subjectivity arises because of contextual interpretation of images, e.g., complex artistic interpretations of the image. This can be seen in captions where poem snippets are used to represent the image. These sort of captions require a great deal of contextual/cultural information (of a different sort) and is out of the scope of this work. The second type of subjectivity is the result of individuals describing different parts of an image with different attributes (to varying degrees). For example, in a picture of a garden, few individuals would focus on the trees (and its properties) and few others on the flowers (and its properties). This work focuses on the subjectivity of the latter kind. Roughly, it can be modeled as a two stage process where individual differences are observed at each of the two stages:

focus based

1. Focus: The caption relates to parts of the image or the image as a whole.

2. Evaluation: These attributes of the focus point are then interpreted, which gives the final information written in the caption.

3.4.1.1 *Sub-Spaces*

Various other subspaces are present within the broader umbrella of *Subjectivity*. These arise as a result of varying levels of visibility within the captions. We shall briefly discuss these sub-spaces below. Figure 3.6 shows the distribution of captions for different combinations of visibility and subjectivity (based on median values of all responses for each given image/caption pair).

To simplify interpretation, we separated this space into 9 boxes (as indicated by the lines in Figure 3.6). Detailed statistics about each of these boxes can be found in Table 3.2. We can characterize these categories as follows:

most common category

High visibility, low subjectivity. This makes up the largest category. There are many captions that could be seen as titles in a classical sense. Very often, additional context information (such as adjectives as attributes, naming rare objects/animals or place information) is included in the caption as well and some captions guide the viewer towards a certain interpretation of the image. Later we will analyze other aspects of this category.

High visibility, medium subjectivity. Typically some aspects of the image are interpreted in a subjective way in the caption. Often there is a clear relation to the image contents.

High visibility, high subjectivity. This combination is rather unusual. Usually the image content is described in a very subjective way and the captions can be of an artistic kind.

Medium visibility, low subjectivity. Captions here typically pick up on something visible in the image while giving additional background information

Medium visibility, medium subjectivity. Quite heterogeneous category. Some are rather subjective interpretations of the image, some merely give background information and some are rather artistic

Medium visibility, high subjectivity. Interesting category with lots of entertaining and expressive captions. Most of the captions can be seen as subjective interpretations of (parts of) the image. Here, the relation to the image can be quite complex and captions should generally be treated as parts of a multi-modal message.

Low visibility, low subjectivity. For the most part, these captions provide invisible background information (usually about the location where the image was taken, sometimes also about the time it was taken or about entities on the image - such as names of people).

Low visibility, medium subjectivity. Uncommon (smallest) and heterogeneous category. Some captions give invisible background information, some relate to the image in rather complex ways and some "classic" titles.

contextual information

Low visibility, high subjectivity. Intent here is for most cases either entertainment or expression. Majority of captions are subjective interpretations of the image content. Relation to the image is at times rather complex. There are also a few cases where there is no clear relation between the image and the title at all.

Since the high-visibility/low-subjectivity category is by far the largest and most existing work focuses on captions from this category, we had a closer look at it. In particular we wanted to see how captions in this category relate to captions from other existing datasets such as MSCOCO. We found is that even though a very large amount of captions are in this general category, most of the captions are subjective or include attribute information that is based on the interpretation of the author. However, for most image captioning datasets out there, all of the captions are visible and there is typically no subjectivity at all.

presence of subjective attributes

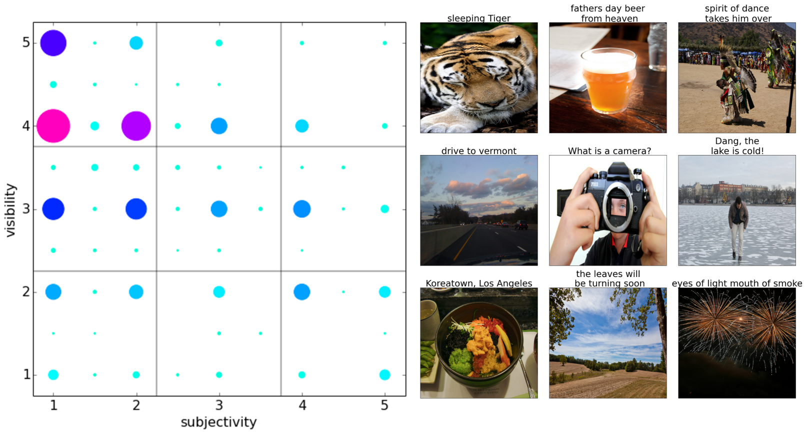


Figure 3.6: Illustration of the variation of visibility vs subjectivity. Each caption was assigned to a single visibility or subjectivity value based on the median values of the annotations for the respective field. The size of the circles represent the number of similar points on the graph. Although, the majority of the captions lie in a high visibility, low subjectivity area, a qualitative analysis reveals that they are not truly objective. The figure on the right hand side shows the corresponding example for each of the 9 categories on the left.

When having a closer look at these captions in our dataset, we find that there are mainly the following sub-types:

- Subjective attributes enhancing each of the noun in the image, including naming uncommon things or animals, and named entities (such as “the beautiful great wall of China")

- Classic titles (like “sleeping Tiger” or “Rails”)

- Highly descriptive titles (such as “Girls working in rice paddies”) do exist but are actually very rare.

It should not be very surprising that there are not many highly descriptive titles because this would in most circumstances be a violation of Grice’s *maxim of quantity* [89]. The maxim of quantity states that

factual	express/entertain	provoke	other
59.74%	49.48%	3.99%	1.13%

Table 3.1: Intent distributions from annotation results. The express/entertain caption takes the second next major share after factual captions.

one gives as much information as is needed and no more (while being as informative as he can). The right part of Figure 3.6 displays one exemplary image-caption pair for each corresponding category on the left side

3.4.2 *Sentiment Space*

presence of adjectives

The sentiment space arises as a result of the analysis of captions w.r.t the "intent" field. Annotators often voted captions as expressive (positive/negative) and entertaining . This led us to investigate the sentiment component of the sampled captions. In natural language, sentiments are often induced by the inclusion of adjectives. These adjectives amplify the users interpretation of the noun attributes inside the image. Therefore, as the first step we analyzed the frequency of adjective usage and the occurrence of *adjective-noun pairs*[7] [70] for the captions. This was done by tagging all captions with Part-Of-Speech tags in order to extract adjectives and nouns. Popular library NLTK[8] [90] was used for tokenization and POS tagging.

For the unfiltered case, 30.03% captions have adjectives and for the filtered captions, this was higher at 54.9%. Adjectives seem to be followed by nouns in 67.5% of the cases (a similar behavior can also be seen in the intent distribution as reported in Table 3.1) . Table 3.3 shows examples of most common adjectives, the corresponding adjective-noun pairs and their usage found in captions. Although factual captions dominate the intent category, it can be seen that expression/entertain is the next major category with close to 49% captions Table 3.1. We also performed an automated sentiment analysis with NLTK-VADER. The sentiment analysis in case of the captions was done mainly to address the following important question:

Do people prefer captions with sentiment over neutral ones?

preference towards sentiment

Figure 3.7b shows the distribution of sentiment of a caption against the preference as evaluated by annotators. In the sentiment range of [0.0 – 0.4] (neutral caption), it can be seen that the preferences are almost equally distributed. However, in the sentiment range [0.5 – 1] (absolute sentiment, positive or negative), annotators have clearly

[7] Cases where adjectives are followed by nouns
[8] http://www.nltk.org/

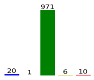
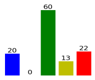
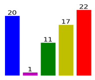
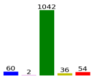
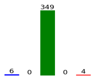
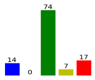
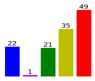
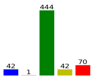
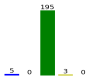
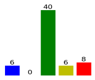
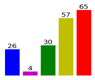
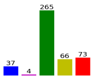
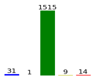
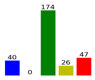
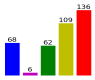
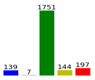
↕ **visibility**	↔ **subjectivity**	low subjectivity (median 1-2)	medium subjectivity (median 2.5-3.5)	high subjectivity (median 4-5)	any subjectivity
high visibility (median 4-5)	#captions	1008 (45.04%)	115 (5.14%)	71 (3.17%)	1194 (53.35%)
	intent distribution ambiguous factual entertainment provoke expression				
medium visibility (median 2.5-3.5)	#captions	359 (16.04%)	112 (5.00%)	128 (5.72%)	599 (26.76%)
	intent distribution ambiguous factual entertainment provoke expression				
low visibility (median 1-2)	#captions	203 (9.07%)	60 (2.68%)	182 (8.13%)	445 (19.88%)
	intent distribution ambiguous factual entertainment provoke expression				
any visibility	#captions	1570 (70.15%)	287 (12.82%)	381 (17.02%)	2238 (100.00%)
	intent distribution ambiguous factual entertainment provoke expression				

Table 3.2: Numbers of captions in the visibility-subjectivity partitions. We assigned the captions to the different partitions based on the median values of the annotations for the respective question. The "numbers of captions" rows give the total numbers (or percentages respectively) of captions that fall within this category in the aggregated data set. The intent distributions show how many frequent the different intents are within the given category (based on majority votes). The last row and last column contain aggregated statistics (combining all visibility levels or all subjectivity levels respectively).

Adjective	Adjective-Noun	Example
new	city, haircut, picture	′**new city** for the day′ ′stylish **new haircut** for cole′ ′**new picture** at the house′
big	house, head, tree	′the **big head** troubled boy′ ′the **big tree** near my house′ ′a progress of my **big house**′
old	dog, man, wood	′dog with **old style** mustard beard′ ′**old man** in the rain′ ′**old wood** fence′
great	clock, sand, party	′the **great clock** tower′ ′**great party** last week′ ′**great sand** dunes national park′
good	food, weather, house	′**good progress** being made′ ′cycling in **good weather**′ ′the **good food** truck is here′
small	plane, flower, church	′ferns growing on the **small church**′ ′**small flower** with spider′ ′boarding **small plane** at den′
grand	opening, canal, river	′**grand opening** day at seven trees′ ′**grand canal** from academia bridge′ ′man at the **grand river**′

Table 3.3: Table shows the seven most common adjectives, the associated ANP and examples from the study dataset. 30.03% captions have adjectives and for the filtered captions this was higher at 54.9%. Adjectives seem to be followed by nouns in 67.5% of the cases.

shown a preference of 3 or above. This can also be validated through the preference distribution of expressive/emotional captions containing higher sentiment score. It can be seen from Figure 3.7a that this distribution has a bias towards higher preference. Therefore, the results seem to suggest that people like captions with a sentiment dimension. A qualitative analysis also showed that this preference was higher for positive sentiment than negative. Additionally, most of the *provoking* captions apparently ended up receiving very low preference scores. For the other intent categories correlations are not as clear.

high sentiment preferred

From qualitatively looking at captions of very low and very high preference, we found that the captions people do not like at all tend to be more generic (e.g., "Pictures from dig cam 10_06 026"), include "noisy" parts (e.g., "Pg 076i Historical Sites") or lack information apart from names (e.g., "Jennie & Colette"). On the other side of the spectrum, people's favorite captions typically relate to the image in a rather clear manner and often make the viewer see the image in some emotional way.

3.4.3 *Variations Space*

This space arises as a result of people using different sentences to describe the same/similar image/s or similar concepts inside an image. These are a result of individual language styles chosen by authors. Authors often try to formulate the meaning they want to convey in their own individual ways. Our backgrounds (cultural/society), education/vocabulary range, moods play as important priming factors here. For example, a crowded railway station can be described as: a) "Monday morning commute" or b) "Crowded morning at station on a Monday". These descriptions refer to a similar situation of a "crowded station" being described in different ways. Synonyms are also often used to describe the same entity referred to in the image. Figure 3.8 shows examples of different captions being used to describe similar images.

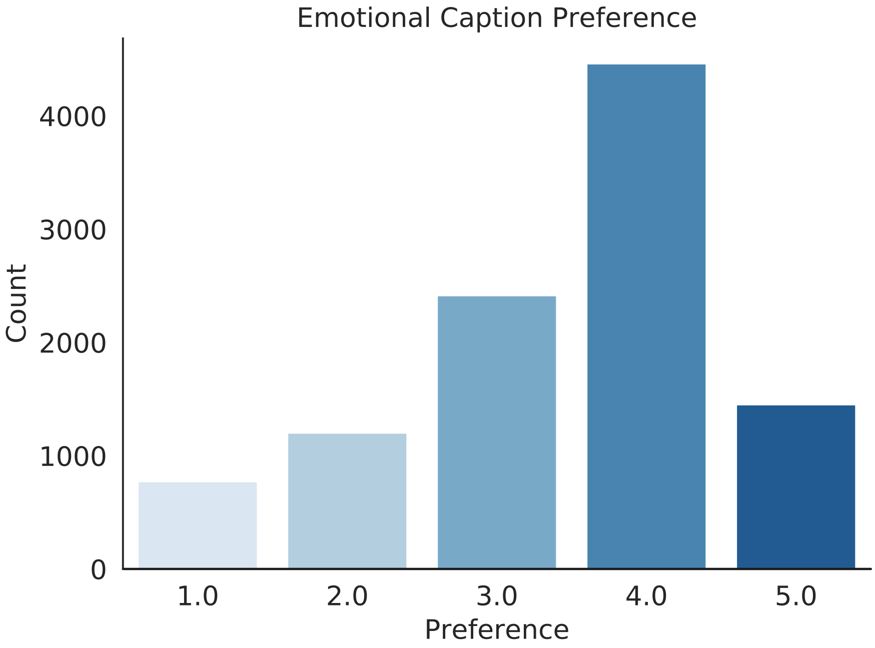
describe same content differently

Finally, variations also occur as a part of the length of the captions. Individuals resort to writing single word captions to longer sentences as a part of the captioning process. Figure 3.9 shows these variations in 1 million Flickr captions.

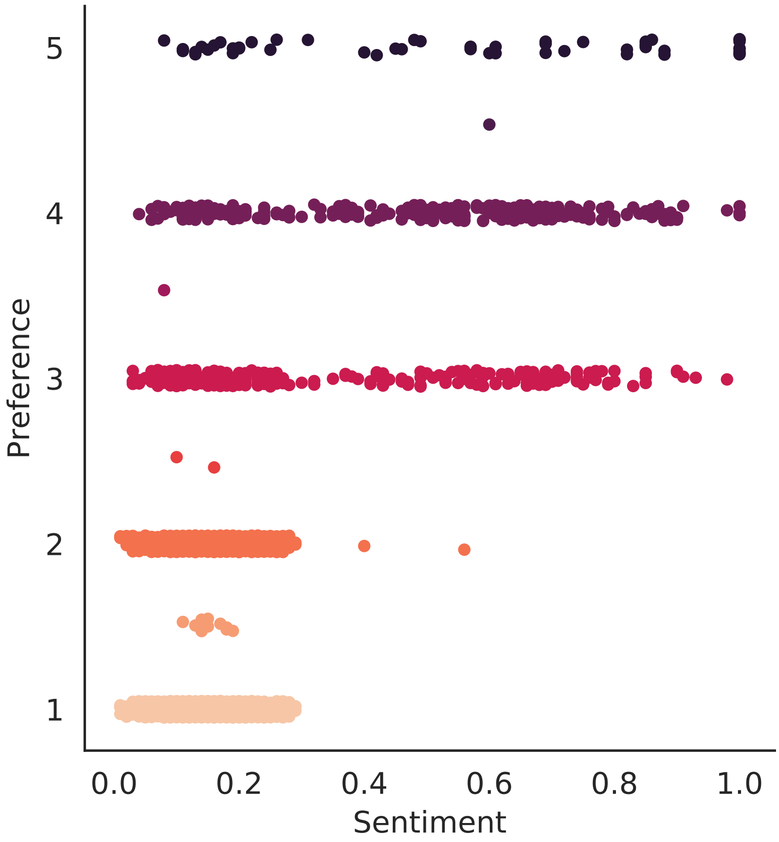
3.4.4 *Other Minor Spaces*

Analysis of the intent category "other" also reveals certain minor spaces. These are the spaces of *Humor* and *Sarcasm*. Although statistically less represented (less than 1.3%), they are equally important. However, the biggest challenge in analyzing these spaces arise from the context required in order to understand them. The humor part of a caption is often highlighted only when viewed through a certain

requires prior contextual knowledge

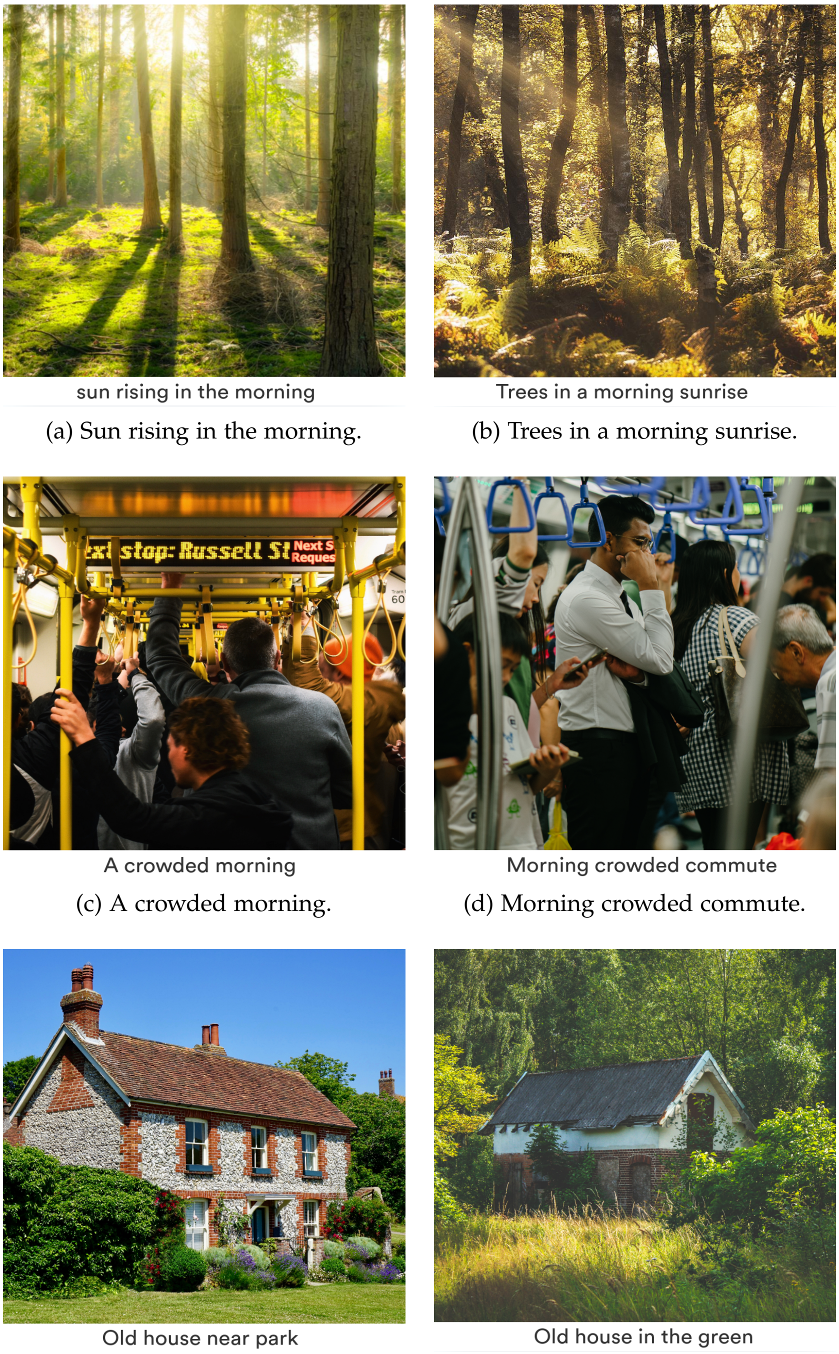


(a) Preference distribution for emotional captions



(b) Preference distribution for sentiment calculated with NLTK

Figure 3.7: Annotator Preference for emotional captions. Figure 3.7a shows the preference distribution with respect to captions which have an intent of Expression/Entertainment. As can be seen, annotators have shown high preference for such captions. Figure 3.7b shows the preference distribution w.r.t sentiment as calculated by NLTK [90]. It can be seen that for high sentiment values annotators have shown higher preference



(a) Sun rising in the morning.

(b) Trees in a morning sunrise.

(c) A crowded morning.

(d) Morning crowded commute.

(e) Old house near park.

(f) Old house in the green.

Figure 3.8: Example of variations in captions. Each row shows two similar images which are captioned differently by Flickr users. These variations try to describe similar things using different sentences. These symbolic social media posts have been recreated by the author which reflect the real examples (to avoid copyright issues). Images used fall under the Creative Commons Zero License (used under CC0).

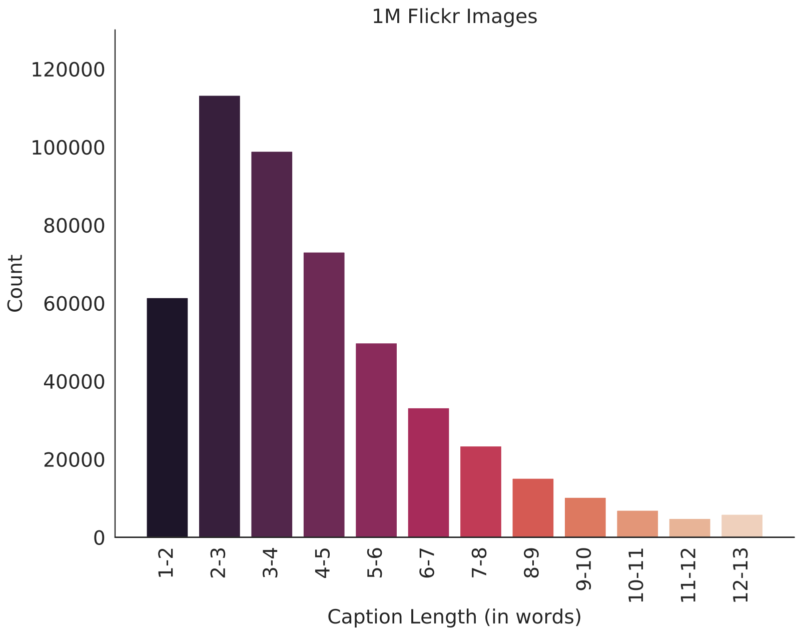


Figure 3.9: Length variations observed among captions as present in 1M Flickr images. These captions exclude the machine generated titles and are filtered for human generated captions. The lengths vary from one word captions to long captions of ten words.

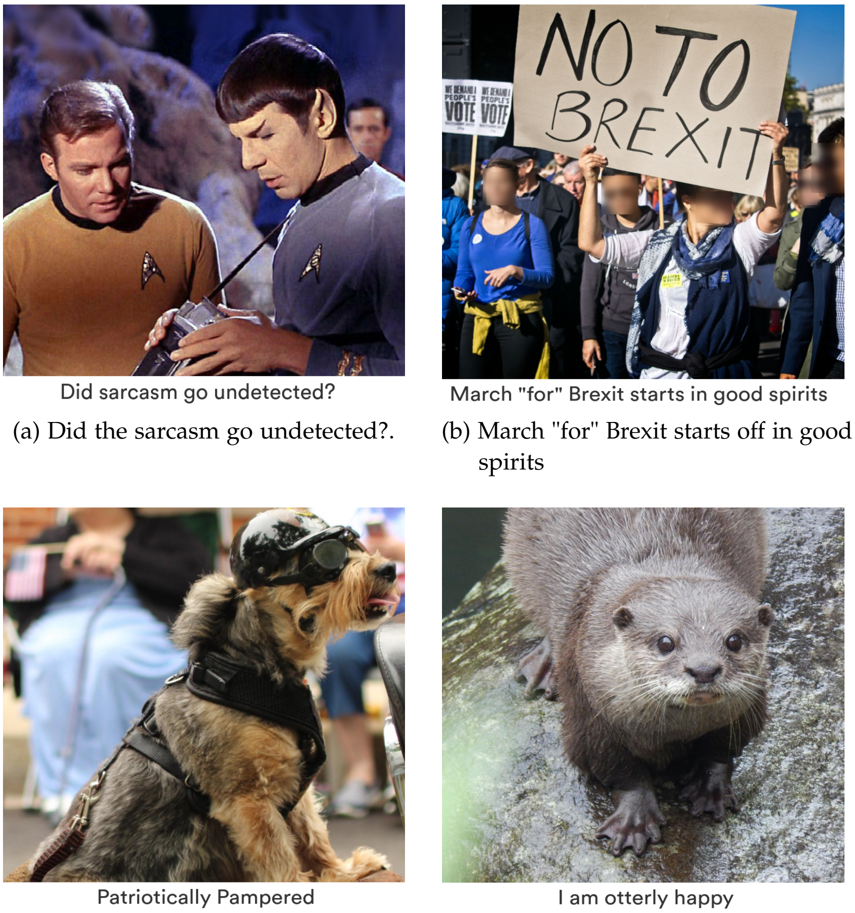
context. This condition also applies equally to understanding sarcasm. Such contexts require prior knowledge and are quite unique to cultures/regions. It assumes that the reader has the required knowledge in this regard (like current political events). For this reason, these spaces are difficult to study and emulate. Moreover, the entities visible in the image are rarely present in the caption as such. Therefore such spaces pose a different set of challenges for automated analysis and are out of the scope of this work. Figure 3.10 shows examples of such spaces.

3.5 CONCLUSIONS

We began by asking the question:

What are the properties of human-generated captions?

In order to analyze human generated captions, we conducted an annotation study through the crowd-sourcing platform CrowdFlower. We generated a dataset by sampling 3,000 random captions from YFCC100M dataset. To weed out the noise, we proposed a new filter as an upgrade to the existing popular caption filter. The experiment ran for a span of 6 months. In the process of qualitative and quantitative analysis of human caption annotations, we found that certain promi-


Did sarcasm go undetected?
NO TO BREXIT
WE DEMAND A PEOPLE'S VOTE
March "for" Brexit starts in good spirits
Patriotically Pampered
I am otterly happy

(a) Did the sarcasm go undetected?.

(b) March "for" Brexit starts off in good spirits

(c) Patriotically Pampered

(d) I am otterly happy.

Figure 3.10: Examples for minor spaces. As can be seen, understanding humor/sarcasm requires certain knowledge about the context. These context range from political (Brexit:Figure 3.10b) to contemporary culture (Star Trek:Figure 3.10a). Often the visibility of entities present in the captions is very low. These symbolic social media posts have been recreated by the author which reflect the real examples (to avoid copyright issues). Images used fall under the Creative Commons Zero License (used under CC0)

nent spaces/properties are observed. Of these spaces, there are three major ones:

1. Subjectivity
2. Sentiment
3. Variations

Although there were other minor spaces, we found that they could not be automatically analyzed because of the context-knowledge assumption. Therefore, going forward we assume that in order for machines to generate human-like captions, they must emulate at least the above mentioned three important properties. Having established a ground for human generated captions, our next research question becomes:

Can machines emulate the properties of Subjectivity, Sentiment and Variations?.

In the following chapters, we shall address each of these properties and develop suitable algorithms to emulate them. We also make the dataset publicly available for further research.

Part IV

AFFECTIVE CAPTIONING MODELS

4 EXPLAINING SUBJECTIVITY FOR IMAGE CAPTIONS

This chapter presents the methods and models proposed to capture visual subjectivity prevalent in images. Subjective visual interpretation is a challenging yet important topic in computer vision. Many approaches have reduced this problem to the prediction of adjective- or attribute-labels from images. However, most of these do not take attribute semantics into account, or only process the image in a holistic manner. There is also a clear lack of relevant datasets with fine-grained subjective labels. In order to overcome these challenges, the chapter proposes a new task of *Aspect Detection* and a novel Focus Aspect Value (FAV) model to implement aspect detection. Different fusion strategies are tried and a new form of fusion, *Tensor Fusion* is also proposed. A new dataset *aspects-DB* is also introduced as a part of aspect detection task. We run experiments on this dataset to compare several deep learning methods and find that incorporating context information based on Tensor Fusion outperforms the default way of information fusion (concatenation).

publication & chapter structure

The methods and models presented in this chapter have appeared in ICMR 2019 [18] under the *Brave New Ideas* category. An extension of this work has been accepted to appear in the journal IJMIR 2020. The rest of the chapter is organized as follows. Section 4.1 discusses the problem of subjectivity in detail. Section 4.2 introduces the FAV model and the task of aspect detection. Section 4.3 explains the compilation procedure of the new dataset. Section 4.4 explains the two tasks that form the core of this method. Section 4.5 gives a detailed description of the experiments and architectures used. Section 4.6 provides our insights and findings from the experiments. Section 4.7 concludes the chapter with a summary and future work.

4.1 PROBLEM DEFINITION

multiple reasons for subjectivity

Subjectivity is defined as the phenomenon where human perception is influenced by personal feelings, opinions, tastes etc. [91] The variance in human perception that arises because of this phenomenon has an important role in the visual domain. For example, the meaning that we infer from an image can depend on: our internal templates about the stimuli [92], expectations and learned biases about the visual object [93], context / prior visual input [94], random neural fluctuations in cortex [95] and other factors like personality of the interpreting individual. This innate diversity in interpretation has made evaluation and computational modeling of subjectivity a difficult task. The challenge in modeling subjectivity arises from two main sources.

1. Subjective interpretation by definition is arbitrary in a certain sense, since there is no a priori objective taste, feeling, or opinion, and at times such context information might not be accessible at all. In particular, this poses challenges to evaluation, and in many cases it is reasonable to expect that there will be a larger margin to a perfect score.

2. Subjectivity tends to be more fine-grained than objectivity. For example, in images, objectivity that is detected typically is about which entities are visible, while subjective information is rather about characterizing how these entities or the picture as a whole differ from some expectation [93, 94].

Additionally, there is also a clear lack of datasets with more fine-grained or structured subjective aspects. This is because existing attribute based datasets [79, 96] either have a skew towards the positive attributes or are based on objective attributes like color. We try to overcome these challenges in the following sections. Therefore, our motivating question becomes:

Can we capture the subjectivity prevalent in the visual medium?

4.2 FOCUS-ASPECT-VALUE MODEL

previous works on subjectivity

There have been many approaches proposed to capture the problem of subjectivity. These approaches either take a holistic view of subjectivity (e.g., in visual sentiment analysis, [97]) or mix subjective components with non-subjective components (as in adjective-noun pairs) [73]. The work of Borth et al. [75] shows that adjective noun combinations are often visible and reasonably simple to automatically detect in images, presumably because of how they contain both subjective (in the adjective) and objective (in the noun) information. A shortcoming is that, in evaluation, these components are mixed and might be hard to separate later on (when the original interest was to focus on subjective parts). Additionally, existing works which use this approach do not include any sophisticated structuring of the subjective components.

Therefore, we take a step back and start with the semantics of adjectives as described by Baroni and Zamparelli in [98], where adjectives are interpreted as modifiers of nouns. We see that visually detecting adjective noun combinations can be understood as a model that combines attention and evaluation. Here the noun describes where the viewer is focusing when interpreting the image and the adjective contains the subjective evaluation of this part of the image. For the adjective, we want to take a step further and acknowledge the fact that adjectives (or “attributes”) for the same noun are often semantically related. In other words, they can be organized along various dimensions of evaluation. For example, these dimensions can be size, age, cuteness or temperature etc. So instead of considering any non-ground-truth attribute as wrong and thereby largely ignoring semantic relations between attributes, we organize attributes into opposing lists. Arranging in this manner paves way for a more appropriate evaluation because:

1. Opposing attributes (mutually exclusive) cannot occur together for the same noun. For example, if we consider the opposing attributes in [“cute”, “adorable”] vs [“scary”, “ugly”], a classification of “cute” and “scary” cannot apply to the same puppy.

2. Semantically related adjectives are grouped together in such an arrangement. For example, a classification of “cute” or “adorable” of a puppy are semantically similar.

Such a pair of opposing attribute lists reflects a certain dimension of evaluation, which we call *“Aspect”* of the noun. Note that an aspect in our case is very similar to the concept of semantic adjective class (used for structuring adjectives in GermaNet [99]), such as *appearance* (“pretty”, “ugly”, ...), *size* (“small”, “big”, “large”, ...) or *age* (“young”, “old”, ...). The difference is that we group attributes of an aspect into mutually exclusive sets. Given a classification of aspect, an evaluation can point to which of the opposing lists the aspect belongs to. For example, for the aspect *size* of a puppy, the evaluation would point to the adjective list of (“small”, “tiny“, ..). We call this evaluation as *Value* of an aspect.

FAV model

In summary, we separate three potential sources of subjectivity in the FAV model:

1. *Focus*: Given a single image, there are typically different components one can pay attention to. For this work, we will assume that this place of focus can be captured by a noun. It should be noted that nouns can not only relate to an entity in the image (such as “dog” or “dude”), but also refer to the whole scene (as in “place”) or the picture itself (“shot”).

2. *Aspect*: Once the focus has been determined, there are several potential dimensions for evaluation. For example, people in the image can be evaluated with respect to their physical size, age, level of activity and so on. Selecting an aspect for evaluation is essentially about choosing a set of semantically related attributes.

3. *Value*:We chose all aspects to be represented by two mutually exclusive sets of adjectives, such that evaluating each aspect amounts to a binary decision problem. For example, physical size would have adjectives like “small”, “tiny”, “short” on one side and “tall”, “big”, “huge” on the other. Picking a certain

value then means to select a set of attributes that are appropriate to be used as an attribute for the given noun. Any adjective from this given set can then be used as an approximate attribute for the given noun.

Figure 4.1 provides a brief overview of the FAV model and illustrates its three components.

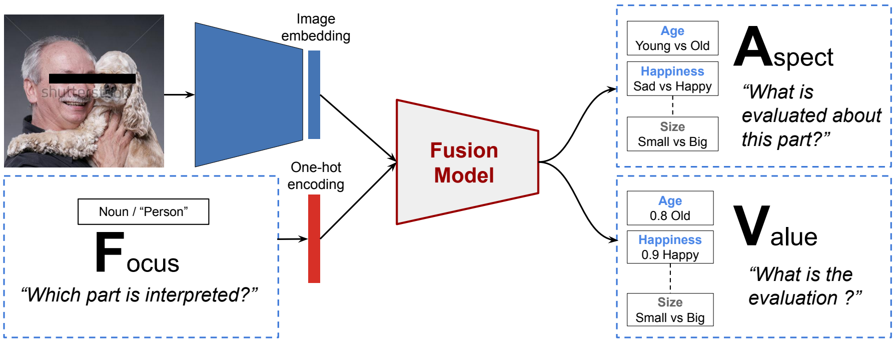


Figure 4.1: Illustration of the task. The model takes an image and a noun (*focus*) present in the image as input. It outputs the corresponding *aspects* and *value* (of each *aspect*). For the given image in the illustration, since the *focus* is on the noun "person", the model identifies the *aspects* "age" and "happiness" as the most appropriate. The *value* provided for each *aspect* determines which set of attributes suits the noun in the context of the image. For the *aspect* age in the given example, the *value* output indicates that suitable attributes for the person in the image would be "old", "elderly", "mature" or "senior", in contrast to "young". The input image is taken from Google's Conceptual Caption Dataset [100].

analogy to NLP methods

Modeling subjectivity in this manner has specific advantages. First, three different sources of subjectivity are disentangled. This brings about the possibility to evaluate these components separately, and helps to make results easier to understand. This way of modeling also offers an analogy to the NLP methods of opinion mining. Opinion mining is a task where topic, aspect and sentiment are extracted from text in order to explain prevalent opinions [97]. Second, semantic relations between attributes are respected. In particular, by detecting aspect values instead of individual attributes, we treat attributes of the same value as being synonymous for the given aspect. We thereby avoid to consider any attribute as wrong if it means the same but is merely phrased differently, as it is for example done when using adjective noun combinations or single attributes as independent class labels. Third, this modeling leads to a more sensible way of 0-shot learning for attribute detection, i.e., predicting subjective attributes for nouns for which they were not available during training time. We will explore this direction in the following sections.

4.3 DATASET

In order to overcome the shortcomings of the existing datasets mentioned above and to have a fair evaluation for experiments, we decided to create a new dataset called *aspects-DB* for subjective visual interpretation, following the FAV model. We will now describe the steps we took for building the dataset. Our dataset is build from Google's Conceptual Caption Dataset [100], which contains over 3 million images together with natural-language captions.

ANP extraction from Google-Captions

First, we ran a POS tagger (using NLTK [101]) over all these captions. From these POS-tagged captions, we compiled a list of all adjectives and nouns which appear in adjective-noun combinations (we selected the adjective-noun combinations which have appeared at least 200 times in the dataset). Our underlying assumptions were that whenever we find such an adjective-noun combination inside a caption (e.g., "cute puppy"), it is very likely that the adjective describes a property of the noun (or "attribute") and that the noun is visible in the image. Next, we manually organized all resulting adjectives (that are potentially visible) into *aspects*. This gave us an initial list of aspects with associated attributes (represented by adjectives) grouped into mutually exclusive sets.

additional filter

We now collected all images from the Conceptual Caption Dataset, which have an adjective-noun combination in their caption where the adjective is included in any one of our aspects. This gave us an initial dataset of over 400,000 images together with associated adjective-noun combination, aspect, and aspect value for each image. We then manually went through the list of remaining nouns, and excluded some words ("retro", "beautiful", "news", "cloudy") which were falsely labeled as noun by the POS tagger, or not clearly visible in images. Finally, we iteratively removed data until all the following criteria were satisfied:

- For each noun-aspect combination, there are at least 10 images for each value.
- For each aspect, there are at least 500 images for each value (across all nouns).
- For each noun, there are at least 50 images for each value (across all aspects).
- For each aspect, there are at most 20,000 images in total. (For aspects with more images, we did a simple down-sampling to reduce the number.)

Figure 4.2 illustrates the above mentioned dataset creation steps briefly.

final dataset

The final *aspects-DB* dataset contains 155,539 images in total and features 143 nouns for 19 aspects. A list with the 5 most common aspects

No.	Name	Values		#Images	#Nouns
1	color	COLORFUL (“blue”, “turquoise”, “green”, “colorful”, “red”, “purple”, “colored”, “colored”, “golden”, “yellow”, “silver”, “orange”, . . .)	COLORLESS (“white”, “black”, “gray”, “grey”, “bland”)	19914 (6728 vs 13186)	46
2	age	YOUNG (“modern”, “new”, “young”, “trendy”, “youthful”, “teenage”, “teen”, “contemporary”, “current”, “recent”)	OLD (“old”, “historic”, “colonial”, “medieval”, “ancient”, “historical”, “traditional”, “elderly”, “senior”, “aged”, “vintage”, . . .)	19793 (11169 vs 8624)	70
3	size	SMALL (“small”, “tiny”, “little”, “miniature”)	BIG (“large”, “giant”, “big”, “huge”, “massive”, “major”, “grand”, “enormous”, “oversized”, “astronomical”)	19177 (12918 vs 6259)	78
4	sun	SUNNY (“sunny”, “bright”, “clear”)	CLOUDY (“cloudy”, “rainy”, “misty”)	14321 (10641 vs 3680)	21
5	rareness	UNUSUAL (“unique”, “ornamental”, “creative”, “different”, “oriental”, “unusual”, “exotic”, “popular”, “stylish”, “magical”, . . .)	ORDINARY (“local”, “daily”, “typical”, “generic”, “general”, “regular”, “familiar”, “casual”, “usual”, “similar”, “normal”, “natural”, . . .)	13299 (6804 vs 6495)	94

Table 4.1: Five most common aspects (out of 19) in the *aspects-DB* dataset. We only list attributes that are included in any adjective-noun combination in the dataset. Numbers in parentheses indicate how many images the dataset contains for the two possible values. The remaining aspects in *aspects-DB* are (from most common to least common): AERIAL VS PANORAMIC, FIRST VS LAST, SMILING VS SAD, FRONT VS BACK, INTERIOR VS EXTERIOR, BRIGHT VS DARK, RURAL VS URBAN, HOT VS COLD, TINY VS TALL, GOOD VS BAD, BUSY VS LAZY, PRIVATE VS PUBLIC, OPEN VS CLOSED, WESTERN VS EASTERN.

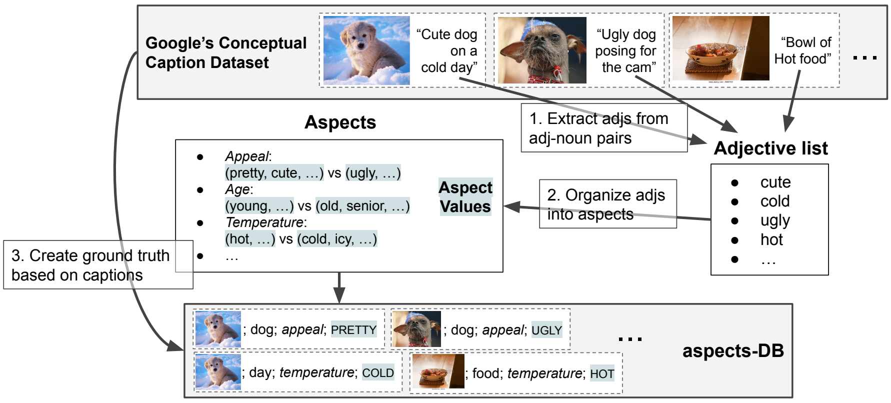

Google's Conceptual Caption Dataset
"Cute dog on a cold day"
"Ugly dog posing for the cam"
"Bowl of Hot food"
...
1. Extract adjs from adj-noun pairs
Adjective list
cute
cold
ugly
hot
...
Aspects
Appeal:
(pretty, cute, ...) vs (ugly, ...)
Age:
(young, ...) vs (old, senior, ...)
Temperature:
(hot, ...) vs (cold, icy, ...)
...
Aspect Values
2. Organize adjs into aspects
3. Create ground truth based on captions
; dog; appeal; PRETTY
; dog; appeal; UGLY
; day; temperature; COLD
; food; temperature; HOT
...
aspects-DB

Figure 4.2: Overview of the dataset creation process. The captions used for illustration were written by us. All images are part of Google's Conceptual Caption Dataset [100].

can be found in Table Table 4.1. Since the ground truth was obtained by adjective-noun pairs, we keep the adjective part in our dataset as extra information, so for each image, *aspects-DB* includes a noun, an aspect, the value of this aspect and the original adjective the noun was combined with in the caption. Table 4.2 shows a few examples of ground truth information for two particular aspect-noun combinations. The dataset is available to the public and can be downloaded at `http://madm.dfki.de/downloads`.

available for download

It needs to be emphasized that the ground truth in *aspects-DB* is meant to capture general tendencies in subjective interpretation (where we use tags as proxy). These tendencies must to some extent be corpus / domain specific and on item-level we cannot expect perfect performance. This means that the task is not to detect objectively correct labels as in many common image classification datasets, but to model general biases such as "for this image of a sleepy puppy and noun *dog*, people would typically interpret the image with respect to aspect *age*. Aspects *age*, *activity*, *evaluation* would likely be rated as having values *young*, *sleepy*, *good* respectively".

4.4 TASKS

Having defined the FAV model and created a dataset, we would like to evaluate our model on this dataset. This means, we need to articulate specific tasks on which the model can be evaluated. In the following sections we list out these tasks.

FAV tasks

4.4.1 *Aspect Prediction*

In the first task, an image and a noun are given and the task is to predict which one of the aspects in our dataset (see Table 4.1) a

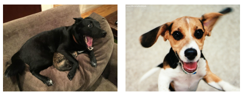
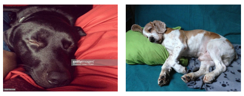
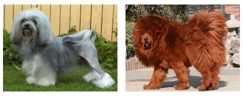
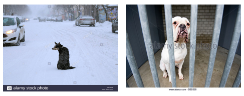
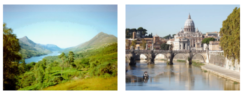
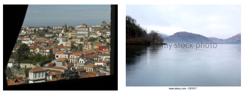
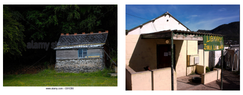
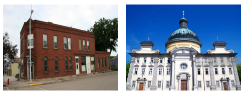
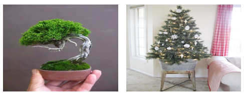
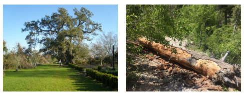
Aspect	Noun	Value	
activity	dog	ACTIVE 	LAZY 
wealth	dog	RICH 	POOR 
sun	view	SUNNY 	CLOUDY 
size	building	SMALL 	BIG 
size	tree	SMALL 	BIG 

Table 4.2: Examples of ground truth data in the proposed *aspects-DB* dataset. Each row represents an aspect for an example noun, and contains sample images which corresponds to the two possible values. All images are part of Google's Conceptual Caption Dataset [100].

subjective interpretation would most likely focus on. For example, given an image with a puppy together with the noun "dog", a likely aspect from our list would typically be *age*. This problem is modeled as multi-class classification task, where, for each given image and noun, only a single aspect is considered to be correct. We evaluate in terms of overall prediction accuracy.

4.4.2 *Aspect Value Prediction*

real number as value

Aspect Value Prediction is about deciding which value applies to a given noun for a given aspect in the context of the input image. Coming back to the previous puppy example of Section 4.4.1, the true value for aspect *age* would be YOUNG when given an image of a puppy with the noun context "dog". For training and evaluation we only consider one aspect at a time, hence this problem can be seen as binary classification task.

4.4.3 *Zero-Shot*

In zero-shot or *0-shot* value prediction, evaluation is done on noun-aspect combinations that were not available during training time. The ground truth data for other noun-aspect combinations with the same noun but different aspects or same aspect but different nouns is assumed to be available for training. For calculating overall accuracy for value prediction, we compute accuracy over all test set items.

unseen combinations

4.4.4 *Dataset Split*

We use two different dataset splits, the *standard* split and the *0-shot* split.

- For the *standard* split, all available data is for each value randomly split into 60% training, 20% development and 20% test data. This implies that aspect and value priors are identical for training, development and test set. We use the standard split for experiments on aspect prediction and aspect value prediction.
- For the *0-shot* split, we randomly split noun-aspect combinations, using 60% of the combinations for training, 20% for development and 20% for testing. We use this dataset split for experiments on aspect value prediction. It should be noted that 0-shot learning on aspect prediction cannot be done in the same way (unless the noun is left out completely for training): If we remove individual noun-aspect combinations and train a model on the remaining ones, the model generally learns that for any noun the excluded aspects are not feasible. This points to another problem in the adjective-noun way of modeling, where aspect and aspect value are both blended into adjective information.

4.5 METHODS

We shall now get into the details of the methods we compare in our experiments , where they are evaluated on the tasks described in the previous section. For all models, except the XResNet variants, visual features are extracted from the image by a ResNet-50 network [102], which was trained on ImageNet [103] and kept unchanged.

implementations of FAV model

4.5.1 *Logistic Regression*

As baselines, we deploy two models based on logistic regression which take visual features from the inception network as the only input:

- The *noun-agnostic* version does not consider noun information at any point. Aspect prediction is modeled as classification task with multiple classes. So for predicting the most likely aspect given an image and noun, a single logistic regression model is trained to output the corresponding class from the visual features, irrespective of the noun. Aspect value prediction is modeled as separate binary classification problems, i.e., for each aspect, one logistic regression model is trained to detect the value of the respective aspect from the image vector, again not taking the noun into account.

predict without noun

- In the *noun-specific* variant, separate models are trained for distinct nouns. For each individual noun, we then follow the same approach as described in the previous point. This means that for each noun we have one model predicting the most likely aspect, and for each noun-aspect combination we have one model for aspect value prediction. We explored this possibility as a simple way to take the noun context into account.

predict with noun

In both cases we have used the scikit-learn [104] implementation for training and inference.

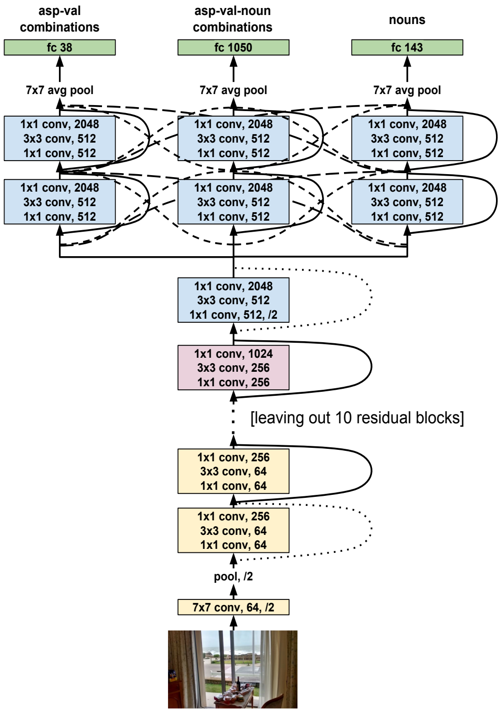


Figure 4.3: XResNet architecture (adapted from [76]). Solid shortcuts indicate identity, dotted connections indicate 1×1 projections, and dashed shortcuts indicate cross-residual weighted connections.

4.5.2 *Cross-Residual Network*

Cross-residual network, or XResNet, refers to an architecture which was introduced by Jou and Chang [76] for adjective-noun pair detection and is based on the well-known residual network (ResNet) architecture [102]. Figure 4.3 shows the modified structure of the XResNet architecture that has been used. The main difference of XResNet as compared to ResNet is that the network branches out at the end into three distinct heads, where these branches remain closely connected to each other via so-called cross-residual connections. The standard XResNet architecture has 50 layers and finally branches out to predict adjectives, nouns and adjective-noun pairs respectively.

XResNet for ANP

We adapt XResNet to our tasks, by replacing these original output branches by three branches specific to our tasks. Instead of adjectives, one branch outputs scores for all combinations of aspect and aspect value. Instead of adjective-noun pairs, scores for all combinations of aspect, aspect value and noun are predicted. The noun branch remains unchanged. This leaves us two possibilities for evaluation:

XResNet modification

- The final decision can be made based on the aspect-value branch (*asp-val*): For aspect prediction, we use the aspect of the aspect-value combination with the highest score as output. In case of value prediction for a given aspect, the value of the aspect-value combinations with the given aspect and highest score is considered. Note that in this version the noun context is ignored.
- The other version is based on the aspect-value-noun branch (*asp-val-noun*): All combinations with irrelevant nouns are removed. The rest is done completely analogous to the *asp-val* case.

4.5.3 *Concatenation + MLP*

The concatenation model is a straightforward application of information fusion, where a one-hot encoding of the noun is appended to the image embedding obtained from the inception network. This concatenated vector is then used as the input to a multi-layer perceptron (MLP) with one hidden layer. The MLP has two output branches, one for aspect prediction and one for detecting aspect value. More precisely, the hidden activation h is computed as

regular concatenation

$$h(x,n) = \tanh\left([W_1|W_2] \cdot \begin{bmatrix} x \\ n \end{bmatrix} + b\right) = \tanh\left(W_1 x + W_2 n + b\right),$$

where $\left[\frac{x}{n}\right]$ stands for the concatenation of x and n, b is a bias vector, and W_1, W_2 are weight matrices of suitable shapes. The dimension of h is a hyper-parameter and is referred to as number of hidden units. The model then estimates aspect likelihoods as

$$\text{softmax}(W_a \cdot h(x,n) + b_a),$$

and aspect values as

$$\tanh(W_p \cdot h(x, n) + b_p),$$

where b_a, b_p are bias vectors, and W_a, W_p weight matrices of suitable shapes such that the output dimension for both branches is equal to the number of aspects. It should be noted, however, that for value prediction during training and testing we only consider the output of the unit corresponding to the aspect which is processed at the time.

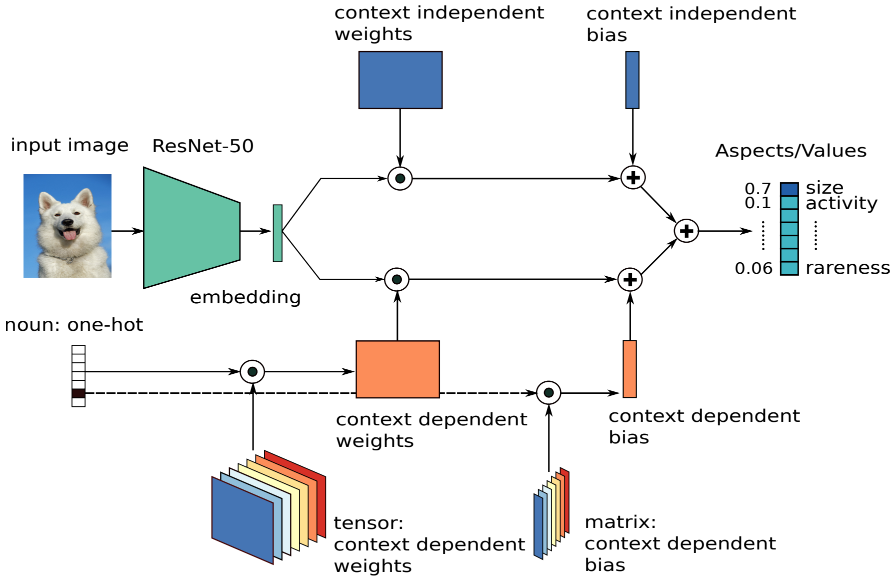


Figure 4.4: An overview of the Tensor Fusion model. Given as input a one-hot encoded noun and an image, the Tensor Fusion model embeds the image with a pre-trained ResNet network. This image embedding vector is then processed with the Tensor Fusion layer which consists of two linear layers, a context-independent and a context-dependent one, followed by element-wise additive fusion of their outputs. For the context-dependent path, the Tensor Fusion layer keeps a tensor with context-dependent weights and a matrix with context-dependent biases, which are multiplied by the noun vector to obtain weights and bias. (Since the noun is one-hot encoded, this multiplication amounts to a selection operation.) We use two separate Tensor Fusion models for our experiments, one for aspect prediction with aspect likelihoods and one for aspect value prediction with aspect values as output.

4.5.4 *Tensor Fusion*

Instead of merely concatenating image features and context we consider a slightly more sophisticated way of conditioning on the context, where higher-order interactions between input and context information are described by a weight tensor. Similar ways of using context information have been used in several publications in the field of natural language processing (for example [105–107]). In computer vision, a related approach for information merging can be found in the MUTAN model [108] for question answering, where question and image embeddings are merged under the use of Tucker decomposition.

context dependent fusion

The Tensor Fusion approach is illustrated in Figure 4.4. The core part is the *Tensor Fusion layer*, which can be understood as part of a neural network that combines the noun-agnostic and noun-specific logistic regression models: For each noun $i = 1, \ldots, 13$, there is a weight matrix W_i and a bias term b_i. In addition, the layer uses a weight matrix W_0 and a bias term b_0 that are independent of the noun context. Given as input the image embedding x and the i-th noun, the output of the Tensor Fusion layer is then computed as

$$(W_0 + W_i) \cdot x + b_0 + B_i \, .$$

We now represent nouns as one-hot vectors $n \in \mathbb{R}^{143}$ and put together all noun weight matrices W_i into a third-order Tensor $W \in \mathbb{R}^{143 \times 1000 \times 19}$ and all noun biases b_i into a bias matrix $B \in \mathbb{R}^{143 \times 19}$. The final layer function $T(x, n)$ can be formulated by using a multiplication operation between the noun context and the weight tensor to obtain the weight matrix for the given noun:

separate weight matrix per noun

$$\begin{aligned} \mathrm{T}(x, n) &= \left(W_0 + \sum_{i=1}^{143} W_i \cdot n_i \right) \cdot x + b_0 + \sum_{i=1}^{143} B_i \cdot n_i \\ &= (W_0 + W \circ n) \cdot x + b_0 + B \cdot n \, , \end{aligned}$$

where $W \circ n := \sum_{i=1}^{143} W_i \cdot n_i$.

We deploy separate Tensor Fusion models for the two tasks of aspect prediction and aspect value prediction.

4.5.5 *Linear Fusion*

Using the same notation for variables as above, we define the linear fusion layer as having the output

$$\mathrm{linear}(x, n) = W_0 \cdot x + b_0 + B \cdot n \, .$$

This enables another view of the Tensor Fusion layer, namely to interpret it as linear fusion plus a term for capturing higher-order interactions:

fusion without bias term

$$\begin{aligned} \mathrm{T}(x, n) &= W_0 \cdot x + b_0 + B \cdot n + (W \circ n) \cdot x \\ &= \mathrm{linear}(x, n) + (W \circ n) \cdot x \end{aligned}$$

Hence, we include linear fusion into our experiments in order to single out the role of the higher-order interaction term, which makes the majority of trainable parameters for Tensor Fusion.

Approach		Aspect accuracy	Value accuracy	
Model	Variant	standard	standard	0-shot
logistic regression	noun agnostic	65.67%	78.97%	60.45%
	noun specific	67.07%	84.79%	-
XResNet	asp-val	45.64%	80.36%	63.71%
	asp-val-noun	**70.30%**	85.34%	-
concatenation MLP	100 hidden units	50.89%	79.36%	68.43%
	5000 hidden units	53.23%	80.11%	61.24%
linear fusion	N/A	62.15%	80.34%	68.41%
Tensor fusion	N/A	69.46%	**86.34%**	**69.04%**

Table 4.3: Aspect prediction and aspect value prediction performances of all models. All methods except the XResNet models use a pre-trained ResNet to embed the image. Please refer to Section 4.5 for details on the individual models. Note that not all models are applicable to the 0-shot learning task.

4.6 RESULTS

For both tasks of aspect prediction and aspect value prediction, we ran experiments with all conditioning methods explained in Section 4.5. Table 4.3 lists out the results of these experiments. Hyper-parameters (learning rate, regularization weight, and number of hidden units for the concatenation method) were optimized based on performances on training and development data (see Section 4.4.4). The performances reported are on the test data.

4.6.1 *Aspect Prediction*

XResNet and Tensor Fusion perform high

For aspect prediction, XResNet is the best performing method (70.30% for *asp-val-noun*), closely followed by Tensor Fusion (69.46%). Both of these models outperform the logistic regression baselines (67.07% for noun-specific and 65.67% for noun-agnostic). The linear fusion model performs worse than the noun-agnostic logistic regression baseline (62.15%). This is unexpected because this model essentially computes the same as noun-agnostic logistic regression plus a linear part coming from the noun. We assume that this effect is due to differences in training and implementation, which was done with the sklearn implementation for logistic regression and using an own neural network implementation in Tensorflow. for the linear fusion model. The gap between performances of the linear fusion to the Tensor Fusion model shows clearly that incorporating higher-order interactions between image features and noun context is beneficial to the task at hand.

Interestingly, concatenation yields very poor performances for aspect prediction. An accuracy of 10% worse than the linear fusion model suggests that the concatenation models were not able to properly make use of the noun information.

4.6.2 *Aspect Value Prediction*

Tensor Fusion performs high

With both the standard dataset split and the 0-shot split, Tensor Fusion gave the best results for aspect value prediction (86.34% for standard, 69.04% for 0-shot). Using the standard split, XResNet was able to achieve comparable performance when using the asp-val-noun output branch (85.34%). For detecting values of unseen noun-aspect combinations (0-shot column), however, XResNet has to rely on the asp-val output, and clearly falls behind Tensor Fusion, linear fusion and one of the concatenation models.

Noun-specific logistic regression is almost on par with XResNet for the standard task (84.79% vs 85.34%), but cannot be applied to 0-shot learning, where the noun-agnostic logistic regression model lead to the lowest overall accuracy (60.45%).

Linear fusion, concatenation, the asp-val version of XResNet and noun-agnostic logistic regression all yield comparable performances (between 78.97% and 80.36%), around 6% lower than the top performing methods. Surprisingly, the corresponding 0-shot results of these models show much greater variation. In particular, concatenation with 100 hidden units gives the second-highest overall score for the 0-shot experiment (68.43%), but the second-lowest one for the standard task (79.36%).

4.7 CONCLUSIONS

We started with the question:

Can we capture the subjectivity prevalent in the visual medium?.

To answer this question and capture subjectivity prevalent in images, we introduced a new task called *Aspect Detection*. We also developed the FAV model as a means of implementing aspect detection. In order to alleviate the dataset related challenges, including the heavy bias towards positive tags / titles in social media, and to make it possible to separately evaluate different parts of subjective visual interpretation, we compiled a new dataset based on Google's recently released Conceptual Captions Dataset [100]. Four different architectures were proposed to run our experiments: logistic regression, XResNet, MLP, and Fusion-based approaches. A novel way of fusion called Tensor Fusion was also introduced as an architecture to implement the FAV model. In order to evaluate these architectures, the task of aspect-prediction and aspect-value prediction were proposed. It was also shown that with the new modeling, it is possible to perform *0-shot* learning to predict unseen noun-attribute combinations. We ran our experiments on the new dataset and reported results with these architectures. Given the prevalence of simple concatenation for combining information in deep learning approaches, we also find it interesting that Tensor Fusion performed better across experiments.

At this juncture, of the three important properties of subjectivity, sentiment, and variations for captioning, we have developed a method to address the subjectivity dimension. In the following chapters, (Chapter 5 and Chapter 6) we delve into the dimensions of sentiment and variations. We shall also show how the detected aspects/adjectives can be used to generate captions from images.

5 SENTIMENT IN IMAGE CAPTIONS

This chapter presents the methods proposed to generate image captions with a sentiment dimension. Humans descriptions of image content often include qualifying nouns or image contents, enhancing the sentiment quotient of the caption. However, most image captioning approaches focus on objective description of images. In order to overcome this challenge, the chapter proposes the Concept and Syntax Transition Network (CAST) networks and Show and Tell with Emotions (STEM) model for sentiment injection. We use the YFCC100M dataset to train our model and prove that we are able to competitively inject sentiment into captions. The models were also selected at the *ACMMM Grand Challenge 2016* as they scored high in the human evaluation track.

publication & chapter structure

The methods presented in this chapter have appeared in ICMR 2016 [21] and ACMMM 2016 [20]. The rest of the chapter is organized as follows: Section 5.1 introduces the motivating problem. Section 5.2 talks about the datasets used in the experiments. Section 5.3 and Section 5.4 talks about the architectures used for sentiment extraction and word matching respectively. Section 5.5 explains the captioning models in detail. Section 5.6 discusses the results and Section 5.7 summarizes the chapter providing few interesting directions towards the future.

5.1 PROBLEM DEFINITION

sentiment presence in captions

Significant advances have been made in generating descriptive image captions [40, 109–111]. However, the focus has been on generating factual descriptive image captions similar to the MSCOCO [112] dataset. Here, the priority has been to generate a caption which is an objective representation of the image. Although such methods provide rich textual descriptions of images, these descriptions might not be representative for natural image captioning. Humans often tend to associate a sentiment with an image and express that in the caption (as discussed in Chapter 3). An analysis of YFCC100M also showed that 54.9 of captions contained a sentiment component in them. Therefore, it is important to include the sentiment dimension in machine generated captions. A lack of captioning datasets with sentiment also adds to the challenges in generating sentiment captions. On the one hand, the popular dataset, MSCOCO, although rich in descriptions has very little sentiment information in captions. On the other hand, large real-world datasets, such as YFCC100M, displays a huge variety of captioning styles but these titles can only be considered to be weak labels [113]. Here, captions can be descriptive, emotional or mention information that is not visible in the image. In order to tackle these issues and to make best use of the available datasets, we propose two paths to generate sentiment captions:

1. Capture the general sentiment of the image and express it in the caption.
2. Qualify the relevant nouns in an objective caption to increase the sentiment quotient of the caption.

methods of sentiment injection

There are different methods to induce sentiment into captions. One is the use of emojis[1]. Here, users include short symbolic facial expressions constructed through keyboard symbols to express their emotions. Another is by using relevant adjectives inside captions. This work focuses on the latter and describes the methods to achieve it. Authors in [114] show that by combining adjectives and nouns into an Adjective-Noun Pair can express the visual contents in the image. Hence we assume that incorporating adjectives into machine generated captions is one feasible way of adding an emotional component to the caption.

In the following sections we analyze and take steps towards addressing our following motivating problem:

Can we generate captions with a sentiment dimension?

5.2 DATASET

structured and non-structured training captions

We use two datasets for training each of our captioning models: YFCC100M and MSCOCO. The first dataset, YFCC100M, contains user captioned Flickr images. Users' captions/tags often do not provide the appropriate data required to train classifiers and generate graphical language models. For example, the images often contain camera generated captions, generic titles, single word captions, locations as reported in [78]. Therefore it was important to extract the relevant Image-Caption pairs useful for model training. We filtered the image captions as described in Chapter 3. After applying the filter, we end up with a training set consisting of 9.6 million images and a validation set consisting of 1.2 million images (where the split into train and validation set is based on the user identifier present in the image meta-data). The second dataset, MSCOCO contains objective descriptions of images. These descriptions were generated by humans in a controlled environment using a fixed vocabulary. The intention is to inject sentiment into neutral captions of MSCOCO to increase its sentiment component.

[1] http://instagram-engineering.tumblr.com/post/ 117889701472/emojineering-part-1-machine-learning-for-emoji, last accessed 17.06.2019

We use the 2016 version of this dataset containing 1,026,459 training images along with 40,504 validation and 40,775 test images.

In order to extract sentiment from images and to extend our vocabulary, we use two popular neural network models: *DeepSentiBank* and *Word2Vec*. The following sections provide a brief overview of these networks.

5.3 DEEPSENTIBANK

To generate captions with sentiment, the first step is to capture the emotional and visual contents from the image. The work [114], introduced Adjective Noun Pair (ANP) concepts able to describe images beyond visual content (e.g., "dog") by capturing positive or negative polarity (e.g., "cute dog" or "scary dog"). The resulting set of ANPs as trained by a deep convolution neural network is called *DeepSentiBank* [115]. The underlying dataset is called *Visual Sentiment Ontology*, wherein images are paired with an Adjective-Noun pairs. The ontology contains 2,089 Adjective-Noun pairs and is constructed from user-generated tags of Flickr images. The final dataset contains 867,919 images. The network has eight main layers with five convolutional and three fully-connected layers. The output of the last fully-connect layer is connected to a 2,089-way softmax which produces a class distribution over 2,089 ANPs. This pairing of adjectives and nouns does also provide an insight into the general emotion associated with an analyzed image. Figure 5.1 shows an illustration of DeepSentiBank (DSB) architecture. Processing the image with DSB gives us a feature vector where each element corresponds to one ANP from a list of 2,089 ANPs. These 2,089 ANPs contain 231 distinct adjectives and 424 nouns. However, this number is insufficient in order for the task of generating image captions. Therefore, we need a mechanism to extend our vocabulary by adding similar and semantically relevant words. In the following section, we shall discuss a neural network which can reveal meaningful associations between words in order to extend our vocabulary.

visual sentiment ontology

5.4 WORD2VEC

Word2Vec [116] is a neural network that can provide meaningful vector representation for words. Converting natural language words to mathematical vectors make them computer understandable. More importantly, these vector representation posses semantically meaningful relationships and can be used to establish word associations. For example, words "cat" and "feline" are closer together than "cat" and "dog". These meaningful relationships can act as a basis for applications that involve search, recommendations, sentiment analysis etc. The vectors which are generated through a neural networks are called

word similarity with neural network

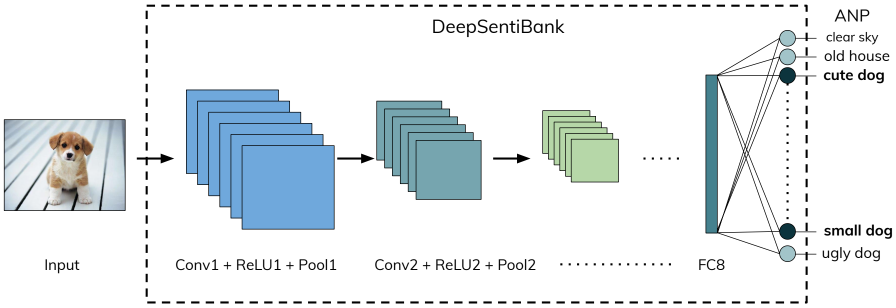


Figure 5.1: The architecture of DSB consists of a CNN with five conv blocks. There are three FC layers at the end. The last layer (FC8) is connected to a 2,089-softmax layer. For a given input image, the last layer provides the probability distribution over the 2,089 ANP label space.

as *embeddings*. Embeddings can also be combined within mathematical operators like addition and subtraction to find interesting word relationships (e.g., “France” + “Capital” = “Paris”, “Library” - “Books” = “Hall”).

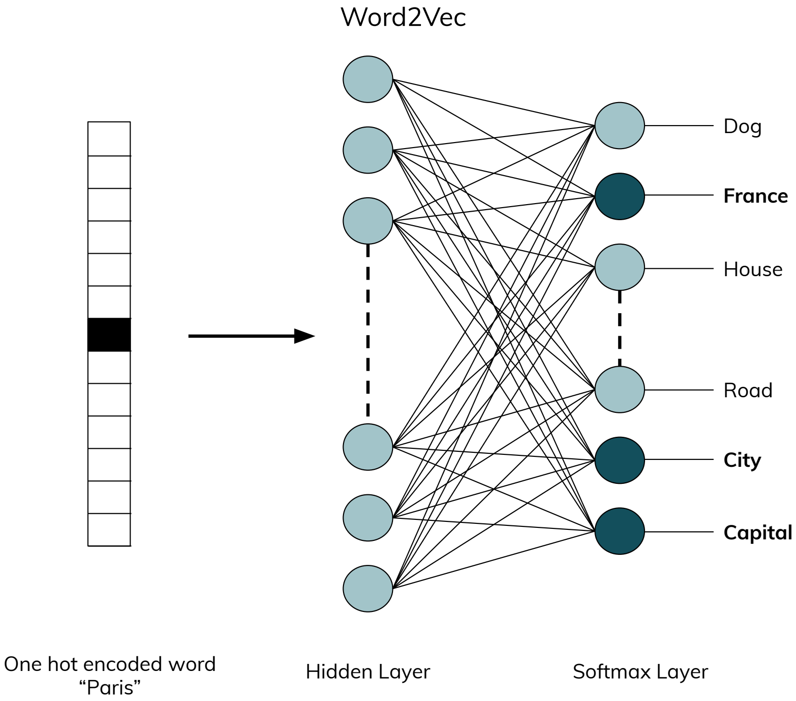


Figure 5.2: Word2Vec architecture consists of a two layer neural network trained to predict semantically relevant words for a given input word. The input word is encoded as a one-hot vector and the last layer provides the probability distribution over label space of output words. In the example above, for input word *Paris*, the words *France*, *City* and *Capital* have the highest scores.

The underlying architecture is a neural network that is trained on a corpus to predict neighboring words (in the corpus) given an input word. This is done by creating bigrams based on a *context-size* around a word. The context-size defines a window around a word (to its left and right) with the word as its center. For example, given three as a context-size, three words to the left and three words to the right of a given word are used to generate bigrams. The network is then trained with the first word of the bigram as input and the second word as target. After the training, the hidden layer of the network acts as the embedding of an input word. Figure 5.2 shows an illustration of this architecture. Word2Vec has been used here to extend the size of the vocabulary by generating meaningful variations. This also helps in creating sentences of different styles yet same meaning/emotion. We shall discuss the caption generation models in the sections below.

context prediction training

5.5 SENTIMENT GENERATION MODELS

In this section, we describe the two models used for generation of image captions with sentiments. The first model, called as the CAST network, is a graphical model. Here a graphical structure is used to capture the language structure and paths within the graph are used to generate captions. The second, the STEM model, is a CNN and LSTM based architecture (see Encoder-Decoder Methods Chapter 2, Section 2.4) which is trained end-to-end.

5.5.1 *Concept And Syntax Transition (CAST) Network*

The CAST network is a multi-directed graph where each node in the network represents a *concept* (i.e. noun, adjective, verb or adverb) connected to other concepts. It is generated in the following steps:

graphical model

1. **Nodes**: For each content word with occurrence count greater than forty in the training titles, we create a node. This leads to a vocabulary of over 21,000 words. Additionally, we add a START and an END node.

2. **Similarity edges**: Similarity edges connect nodes which have similar meaning or are semantically similar. In order to detect such similarities, we train a Word2Vec model on all training sentences. We use a dimension of two hundred for the vectors. For each node we add similarity edges to all nodes that have a Word2Vec similarity above a fixed threshold (0.5 in this case). This accounts for the possibility of replacing words by semantically similar words in the sentence generation process.

word2vec matching

3. **Syntax edge**: Syntax edges try to capture the language grammar of the underlying training set. For each sentence in the training captions, we check how content words are connected. For

each such connection that does not use another content word, a directed edge is created between the corresponding concept nodes. The connecting string (usually consisting of propositions, articles, etc.) is used as edge label and the total number of connection occurrences is annotated as edge weight. These edges contain information about the syntax of the language and are used to connect different concepts.

graph traversals

After the CAST network has been created from the training data, generating a sentence from a set of concepts is reduced to the problem of finding a path from the start to the end node through a set of activated nodes. Computing a list of such paths is done in a heuristic way and this list is then ranked by considering the weights of the included edges. The path with the highest score is then converted to a sentence in a straight-forward way, where similarity edges are used to substitute words. An illustration of a simple CAST network can be found in Figure 5.3. We build our graph (including Word2Vec model) on YFCC100M, showing that our method works well with noisy real-world data without any sophisticated preprocessing[2].

5.5.2 *Template-based Approach*

fallback option

In case, the CAST network fails to find a suitable sentence for the image, we rely on a template based approach to solve this case. The idea of this approach is to use different templates of the kind “ADJ HUMAN with PROPERTY doing VERB on EVENT in LOCATION" to form sentences from a set of visual concepts that have been tagged by according category and are detected in the image.

For this we need:

- **Category tags**: We manually assigned category tags (e.g., “HUMAN" or “LOCATION") to all nouns that occur in any ANPs.
- **Templates**: A few (5) templates based on these category tags were created manually. From that we automatically generated different template variations by removing parts of the template. (e.g., the variations of the above template would include “ADJ HUMAN doing VERB in LOCATION" and “HUMAN with PROPERTY on EVENT in LOCATION".)

On the ACMMM Grand Challenge test set, we found that less than 10% of images required the template based approach as a fallback option.

Finally, the presented system follows a pipeline approach consisting of the following steps:

[2] In principle it would be possible to train the visual concept detector on the YFCC100M data as well.

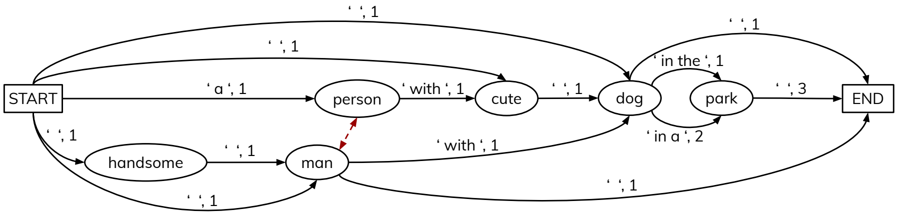


Figure 5.3: Example of a CAST network generated from the titles "handsome man", "a person with cute dog", "dog in a park", "dog in the park" and "man with dog in a park". The dashed line indicates a similarity edge (and is in this toy example not generated from the given sentences). If the red nodes denote the activated concepts, the resulting sentence would be "person with cute dog in a park". (Substituting *man* by *person* because a similarity edge was traversed.)

pipeline for CAST

1. **Sentiment extraction**: We process the image with DSB to extract concepts including emotional cues. Given ANP scores from DSB, we consider all ANPs that have a score above a fixed threshold to create sentences from all suitable template variations (If no score exceeds the threshold we take the ANP with highest confidence and return it as caption.)

2. **CAST network**: Generate ranked sentences from the detected concepts.

3. **Ranking**: We rank all resulting sentences based on a scalar rating score that is computed for each sentence individually, using for the computation the DSB scores of all ANPs that are present in the sentence.

4. **Templates**: If the rankings from the network are below a threshold, we use a template-based approach to create sentences.

Algorithm 1 and Figure 5.4 show the overview of the steps involved in the complete pipeline.

5.5.3 *Show and Tell with Emotions (STEM)*

neutral captions to sentiment

The STEM model for sentiment caption generation is an encoder-decoder model. It generates sentiment in captions in two steps. In the first step, it generates a neutral caption related to the image. In the second, it injects adjectives into the neutral captions at suitable locations to enhance the sentiment component of the same. The following sections describe the two steps in detail.

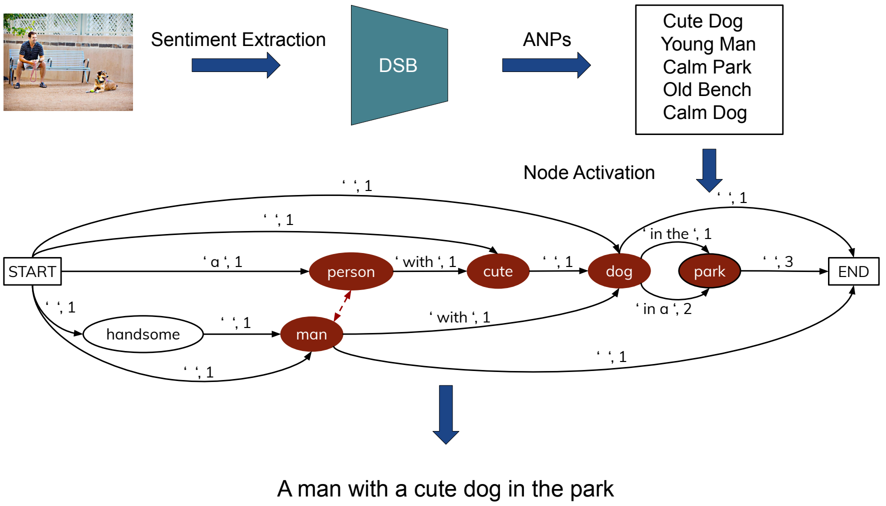


Figure 5.4: The pipeline of CAST network consists of three main steps. First the sentiment of the image is extracted with DSB to get the relevant ANPs of the image. Second, The nodes relevant to the adjectives and nouns are activated in the CAST graph. Third, relevant paths through the graphs are ranked and the path with the highest rank is provided as the caption.

Algorithm 1: CAST Network pipeline.

```
Input: Image, $I$, CAST Graph, $G$
Output: Sentiment Caption, $C_s$
Data: YFCC100M
/* Step1: Sentiment Extraction                                  */
Extract sentiment of $I$ with DeepSentiBank through ANPs,
  $AN = \{a_1, n_1, a_2, n_2, \cdots a_k, n_k\}$
Extract nouns from $C_n$, $N = \{n_1, n_2, \cdots, n_l\}$
/* Step2: CAST:Sentiment Caption Generation                     */
foreach ($adjective : a, noun : n_a$) in $AN$ do
    Mark relevant nodes in $G$ that correspond to $a$ and $n_a$
Find paths $P = \{p_1, p_2, \cdots, p_n\}$ in $G$ to include the maximum of
  marked nodes
foreach path $p$ in $P$ do
    Generate Caption by listing node and edge labels in $p$
    Append $p$ to $C_l$
```

5.5.3.1 *Neutral Caption Generation*

encoder-decoder model

The neutral caption for the image is generated in a similar fashion of [11], wherein a CNN is used to convert the image into a feature vector. We used a VGG network [25], to embed the image into a feature vector. This feature vector acts as the initial hidden state for a LSTM network (with a hidden dimension of 256). The LSTM is unrolled over time and generates one word at each time step to form a caption of the image. We use the MSCOCO dataset to train this architecture.

5.5.3.2 *Sentiment Injection*

word2vec and DSB for matching

Having generated a neutral caption, we would like to inject sentiment into it. In order to do so, we have to identify suitable adjectives for the image and positions inside the caption. In the first step, we extract the sentiment with relevant ANPs from the image using DSB. Then relevant nouns inside the neutral caption are identified with a part-of-speech tagger (we use the NLTK [90] library here). We match the nouns inside the neutral caption against the nouns in extracted ANPs. The matching is done using Word2Vec and a match-value higher than 0.5 is considered as relevant. In the last step, the adjectives from the matched ANPs are inserted before the matched nouns in the neutral caption. Algorithm 2 and Figure 5.5 show the overview of the steps involved in the pipeline.

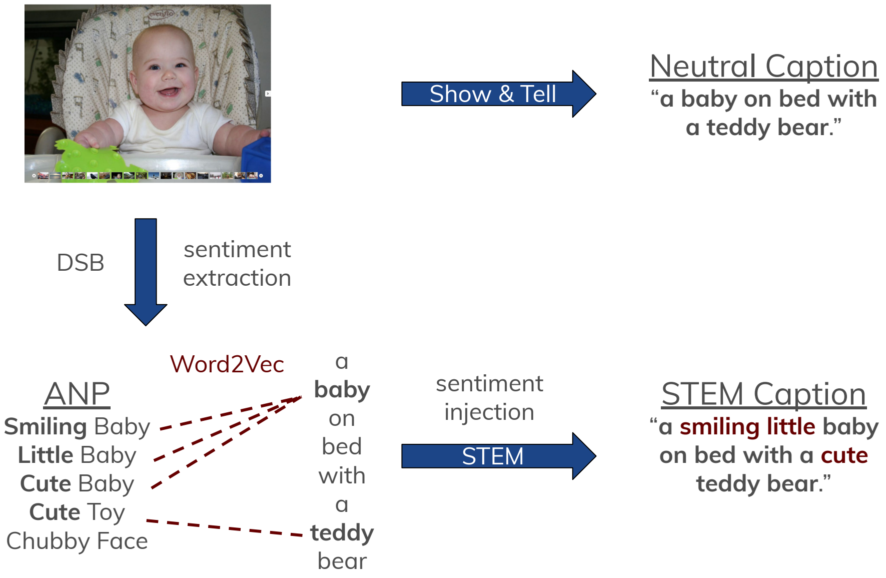


Figure 5.5: In the STEM model, first a neutral caption is generated for the given image through an encoder-decoder model. The sentiment of the given image is extracted through DSB and the relevant ANPs are found. The nouns of the ANPs are matched against the neutral caption through Word2Vec. In the last step, the adjectives are injected into the neutral caption at the relevant positions improving the sentiment quotient of the neutral caption.

5.6 RESULTS

The results were evaluated through two methods. In the first evaluation, the individuals were asked to choose between two captions (ground truth and generated ones). The second was done to evaluate

Algorithm 2: STEM Model

```
Input: Image, $I$
Output: Sentiment Caption, $C_s$
Data: MSCOCO
/* Step1: Neutral Caption Generation */
Train encoder-decoder model for generating a neutral caption, $C_n$
/* Step2: Sentiment Extraction */
Extract sentiment of $I$ with DeepSentiBank through ANPs, $AN = \{a_1 n_1, a_2 n_2, \cdots a_k, n_k\}$
Extract nouns from $C_n$, $N = \{n_1, n_2, \cdots, n_l\}$
/* Step3: Sentiment Injection */
foreach $noun, n$ in $N$ do
    foreach $adjective, a, noun, n_a$ in $AN$ do
        if $Word2Vec(n, n_a) > 0.5$ then
            /* Find most similar nouns between neutral caption and ANPs */
            insert $a$ in $C_n$ before $n$
```

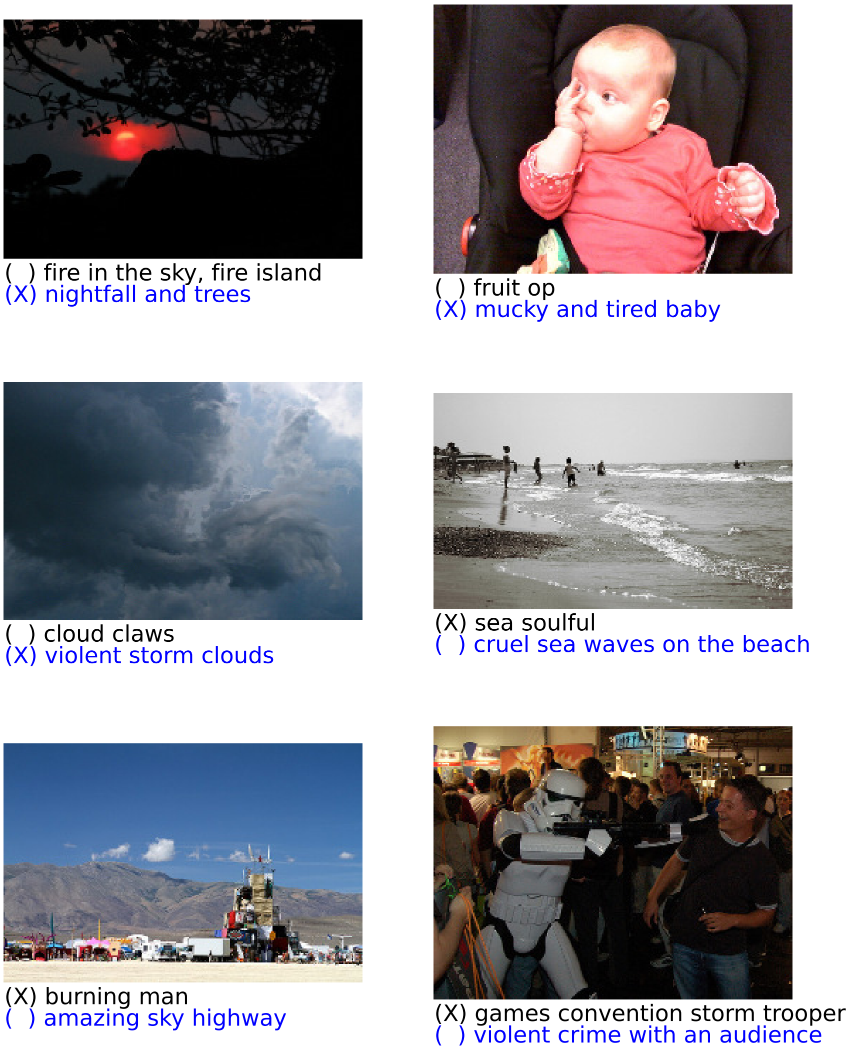
the quality of the generated caption through the ACMMM Grand Challenge. We discuss both of the evaluation methods in the following sections.

5.6.1 *Human Evaluation (CAST)*

forced choice

In order to evaluate the *humanness* factor of the generated caption, we selected 200 random images from our test set. These images were assigned two captions: The original caption present in the YFCC100M dataset and the caption generated by our method. Without informing the individual about the source of the two captions, we asked human subjects to choose one among the two captions which they thought had a higher emotional content and were generated by a human. To compensate for the subjective bias in human evaluation, each image was shown to three different individuals and the opinion of the majority was decided as the final result for that image.

We report that 31.5% of the captions generated by our method were reported as more human-like in comparison to the original caption by at least two subjects. In 62.5% of images at least one subject chose our caption over the original one. These results are encouraging and the generated captions often read naturally, conveying emotions. The creativity and subjectivity that is displayed in some of the captions is very entertaining. Figure 5.6 shows a small selection of titles generated by our method.



( ) fire in the sky, fire island
(X) nightfall and trees

( ) fruit op
(X) mucky and tired baby

( ) cloud claws
(X) violent storm clouds

(X) sea soulful
( ) cruel sea waves on the beach

(X) burning man
( ) amazing sky highway

(X) games convention storm trooper
( ) violent crime with an audience

Figure 5.6: Qualitative results of our approach for images of the YFCC100M dataset. The captions in black are the ground truth titles, in blue we have captions produced by the combination of DSB and CAST. The “X" marks indicate which caption the majority of people in our evaluation experiment believed to be created by a human.

5.6.2 *ACMMM Grand Challenge Human Evaluation*

ACMMM human evaluation track

The results generated by the models were submitted to the ACMMM Grand Challenge 2016. In this challenge, participants were asked to train their models on a training set released by the committee. The results were evaluated by human experts assigned by the committee on a set of test images. The committee evaluated the results on the following criterion on a scale of zero to one:

- [score 0]: The caption has very little relation to the content or simply does not make sense.

expert evaluation criterion

- [score 0.25]: The caption seems reasonable, but other metadata associated with the photo (e.g., location where the photo was taken) makes the caption incorrect.
- [score 0.5]: The caption correctly identifies some of the content or scene (whether explicitly or implicitly visible)
- [score 1.0]: The caption correctly identifies most of the content and emotion (whether explicitly or implicitly visible)

Our models scored on an average 0.932 and were selected for the challenge as they stood within the top-3 submissions in the contest. These results are encouraging and point towards the effectiveness of sentiment injection through adjectives.

5.7 CONCLUSIONS

We started the chapter with the following question:

> *Can we generate captions with a sentiment dimension?*

In order to answer this question, we proposed two models to generated captions with sentiments. The first, CAST network, was a graphical model trained on YFCC100M. The second, STEM, was a neural network based model for sentiment injection. Both models used DSB as fixed visual sentiment extractor and Word2Vec for word matching. We also proved the effectiveness of the models through human evaluations. The models were also selected at the ACMMM Grand Challenge 2016 as they stood among the top-3 entries in the contest. In general, CAST provides a new possibility for the challenging task of generating sentences from an arbitrary sets of words. With the graphical structure one also can follow and influence all the steps from detected concepts to final sentence, giving much more control and making the model more explainable. By using word similarity in the sentence generation process, they display a high degree of creativity, effectively extending the vocabulary. They do all this in a simple and transparent way which makes it easy to find the source of mistakes, allowing for systematic improvements or customization of the system in the future.

To improve the existing models and to get the caption quality closer to human levels we are planning to extend our work by incorporating the following points: The grammar of the whole sentence needs to be given more weight. We are currently working on an additional ranking mechanism to take that into account. Also, so far the confidences of the detector are only respected in the thresholding and then discarded. We plan to either modify the network traversing algorithm such that it also respects the concept scores or respect the scores in the final ranking of the proposed sentences. We also want to use additional concept detectors to get more different sentences from the network and optimize the whole network on more data.

At this point, we have addressed the properties of subjectivity and sentiment for image captioning. In the following chapters (Chapter 6) we talk about generative models and their application in generating variations in image captions.

6 CONTROLLED VARIATIONS IN IMAGE CAPTIONS

This chapter presents the algorithms proposed to generate variations in image captions. An image can be described in different ways which have similar meanings. We propose a method to control the properties of these variations e.g., sentiment. We take a generative approach to solve this problem using CGAN. The modified CGAN architecture contains a Generator and two Discriminators. An external signal called a *context* is provided to control the property of variations. This context is encoded as an input to the Generator (as a part of its latent space). Reinforcement Learning algorithm of *Policy Gradients* is used to train this architecture to alleviate the challenges posed by vanishing gradients and long term dependencies. Using a two step training process combined with Policy Gradients, we show that this architecture can be trained in a stable manner. The architecture is also evaluated quantitatively on the state-of-the-art image caption dataset and qualitatively using a crowd-sourcing platform. Our results, along with human evaluation prove that we competitively succeed in the task of creating variations and sentiment in image captions.

publication & chapter structure

The methods and models presented in this chapter have appeared in ICANN 2019 [22] .The rest of the chapter is organized as follows: Section 6.1 introduces the motivating problem of this chapter. Section 6.2 discusses the generative architecture and adversarial training. Section 6.3 and Section 6.4 gives a detailed description of the model architecture and training respectively. Section 6.5 describes the dataset used to train the model. Section 6.6 provides the quantitative and qualitative results. Section 6.7 concludes the chapter with a summary.

6.1 PROBLEM DEFINITION

multiple caption-variations for single image

Humans often use a wide variety of captions while describing images. Individuals use a variety of styles while writing captions owing to their different social backgrounds, cultures etc. However, the variation dimension is often neglected in state-of-the-art image captioning models where the intention is to generate a caption which is as close to the ground-truth as possible. To change this paradigm, Generative algorithms have been used in creating variation in captions [55]. However, these variations are often generated randomly without any control from the user. Therefore, it would it beneficial if certain properties (e.g., sentiment) of the generated captions can also be controlled through an external context. In such a case, the model can not only generate variable captions but the user can also control its properties by varying the context. For example, if the context is the sentiment, then the user can preset the sentiment of the generated variations through an additional binary variable (for the rest of the chapter we shall consider the context as the sentiment information). This requires a fusion of the context(sentiment information here) with the input.

Generative models are inherently difficult to train on languages for two main reasons:

difficulty in training

1. The process of generating language is a sequential-sampling procedure which is non-differentiable, making the direct application of backpropagation difficult
2. Long dependencies formed as a result of longer sentences mean that training them with back propagation shall suffer from vanishing/exploding gradients.

Given these challenges, our motivating question then becomes:

> *Can machines generate captions with variations and control its properties?.*

sentiment as a control property

To begin our approach, we must find a *property* which can easily be verified in the variations. There are different properties that can be used here. For example, the style of the caption, language, linguistic properties, sentiment. We choose the property of sentiment for this experiment as it is relatively easier to both qualitatively and quantitatively verify this property. Additionally, we found that for the property of sentiment we could also compare our results against state-of-art models.

6.2 GENERATIVE MODEL

Image captioning frameworks generally follow an encoder-decoder architecture [11, 15]. The input image is encoded into a n-dimensional space using a CNN. The encoded image acts as the initial state for the decoder which is a LSTM to generate a text sequence. The network is trained using a *maximum likelihood* loss (e.g., Cross Entropy Loss). This ensures that the network generates captions which are as close to the ground truth as possible. Deviating from this convention, we can have a generative architecture.

GAN for captioning

Generative models have shown to be effective at approximating unknown distributions. The most successful among generative models, called GAN [52] has proved to be highly efficient at tasks like image generation, image completion etc. [117]. A typical GAN includes a generator network which, given a noise vector z, generates data items and a discriminator network which evaluates these items (if generated or real). Together, they perform a *min-max* game, where the generators objective is to generate data which can fool the discriminator and the discriminators objective is to accurately distinguish the generated data from real. A variant of GAN, called CGAN [117] follows an architecture where generator and discriminators are conditioned on an external input. Our proposed method takes inspiration from the CGAN architecture where sentiment acts as the external condition (context).

Reinforcement Learning for training

As mentioned in the above section, training a GAN for languages is a non trivial task. Therefore, we need to search other algorithms (other than backpropagation) to stabilize the training. The authors of [55] have shown that reinforcement learning algorithms like Policy Gradients and Monte-Carlo rollouts can be used to mitigate these effects in order to train a GAN for caption generation. Our final model takes as input, an image and a binary variable (indicating the desired positive or negative sentiment of the caption) to generate captions accordingly. Figure 6.1 and Table 6.2 show the basic architecture of the model and few examples of generated captions respectively.

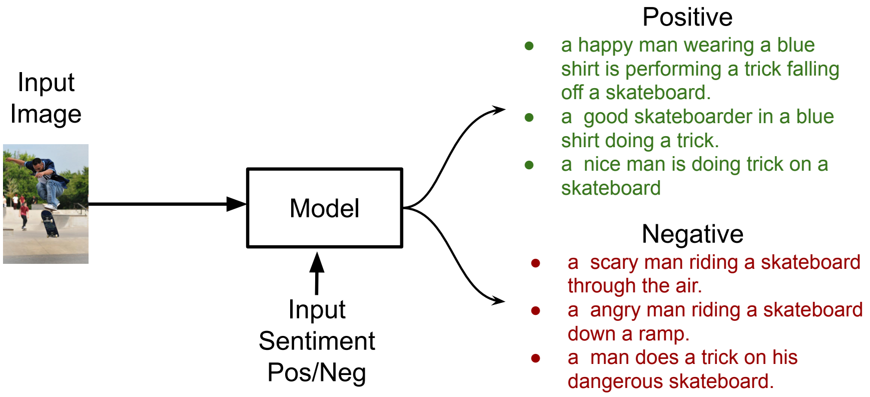

Input Image
Model
Input Sentiment Pos/Neg
Positive
a happy man wearing a blue shirt is performing a trick falling off a skateboard.
a good skateboarder in a blue shirt doing a trick.
a nice man is doing trick on a skateboard
Negative
a scary man riding a skateboard through the air.
a angry man riding a skateboard down a ramp.
a man does a trick on his dangerous skateboard.

Figure 6.1: Basic overview of our model. The input to the model is the image and a binary vector indicating the required sentiment (positive/negative) of the output captions. The model generates the caption which has the input sentiment and multiple variations.

In the following sections, we shall look into this architecture in detail and its training methods.

6.3 ARCHITECTURE

one generator, two discriminators

Our architecture is similar to a CGAN [55] but with one generator: G and two discriminators: D_r, D_s (as opposed to one generator and one discriminator). The generator, G, given an image, generates a sequence which is evaluated by the discriminators. The objective of D_r and D_s is to accurately judge the relevance and the sentiment of the generated caption respectively. Figure 6.2 shows an illustration of the proposed architecture.

Our training also differs from the adversarial approach [52]. Briefly put, our training contains two phases. In the first phase, we train both the generator and discriminator. After the first phase, the discriminator weights are frozen and they now act as reward agents. In the second phase, the rewards produced by the discriminators (for the generated captions) act as a feedback to further train the generator. Training using this reinforcement technique is called Policy Gradients. We shall look into the individual components of the network in detail and the encoding scheme used for the input.

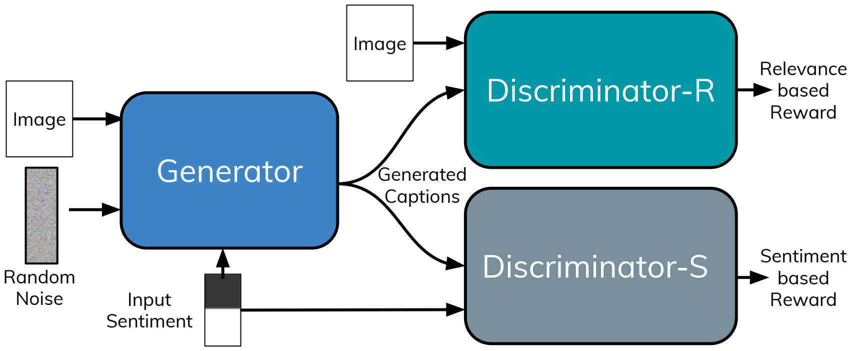


Figure 6.2: Detailed architecture of our model. The Generator takes the image and a binary sentiment vector as input. Discriminator-R uses the same image to evaluate the quality of the generated caption. Discriminator-S uses the input sentiment vector and the generated caption while evaluating its reward.

6.3.1 *Generator*

LSTM generator

The generator G, is a single layer LSTM network (hidden dimension h_g) which takes an image along with a noise vector $z \in \mathbb{R}^m$ as input and generates a caption by sampling discretely from the output. The input image is first converted into a feature vector, $f \in \mathbb{R}^n$ using the last fully connected layer of a pretrained CNN. The objective of the generator is to generate captions which are relevant to the image and have a positive/negative sentiment based on the encoding of input noise vector z. Figure 6.3 shows an illustration of the generator.

6.3.2 *Noise Encoding*

context encoding in noise vector

We want to provide an input to G for the intended sentiment of the caption. To achieve this, we used the noise variable z. We split z into two parts: a 512-dimensional vector sampled from $\mathcal{N}(0, 1)$ and a 512-dimensional latent code vector which is assigned values based on the sentiment in the ground truth caption, namely:

- if the ground truth caption had a positive sentiment, then the first 256 dimensions in latent code were set to 1 and rest as 0

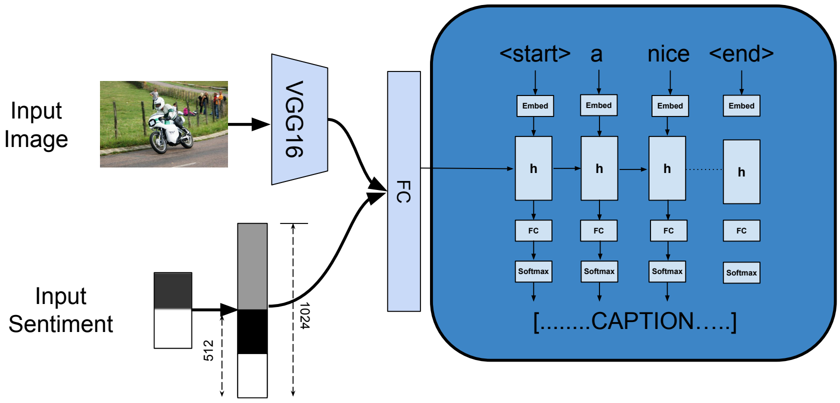


Figure 6.3: The generator is an LSTM network. It takes two inputs: the encoded image from a VGG16 network and an encoded noise 1,024-vector z. These inputs are combined together into a FC layer and provided as the starting state of the LSTM

- if the ground truth caption had a negative sentiment, then first 256 dimensions in latent code were set to 0 and rest as 1.

Figure 6.4 shows an illustration of this encoding scheme.

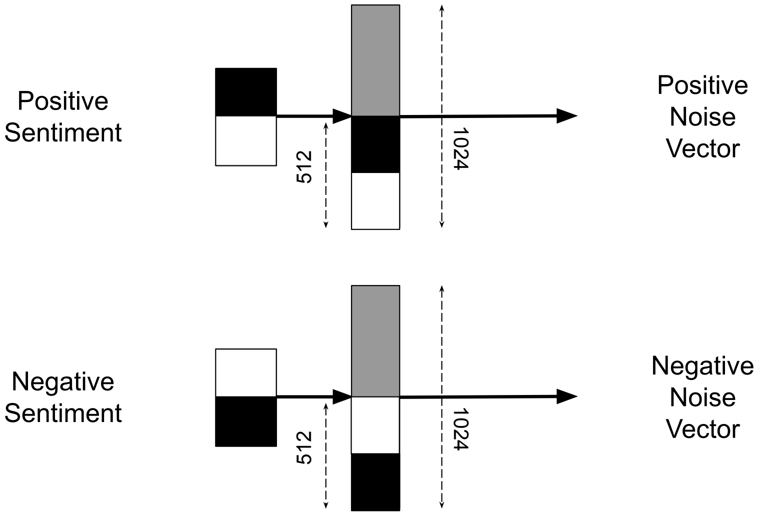


Figure 6.4: The binary input of sentiment is encoded into a 1,024 noise vector z. The first 512 values are chosen from $\mathcal{N}(0,1)$, indicated by the grey color. The second 512 values are encoded based on the chosen sentiment

6.3.3 *Discriminator-1*

caption relevance discriminator

The first discriminator D_r, is a LSTM network (hidden dimension h_d), which given an image and a caption, distinguishes between the captions generated by G from the ones present in the training set. D_r also takes into account the semantic relevance of the generated caption given the input image and the true caption of the input image. Figure 6.5 shows the brief illustration of D_r. The objective function to train the discriminator is an extended version used by [55]. For D_r with parameters η, given an image I, the objective function (Equation 6.1) and reward (Equation 6.2) can be formulated as:

$$\begin{aligned} L_{D_r}(I;\eta) = \mathbb{E}_{S_r \sim S_T} \log R_{D_r}(I, S_r) + \alpha \cdot \mathbb{E}_{S_g \sim S_G} \log(1 - R_{D_r}(I, S_g)) \\ + \beta \cdot \mathbb{E}_{S_n \in S_N} \log(1 - R_{D_r}(I, S_n)), \end{aligned} \tag{6.1}$$

$$R_{D_r} = \sigma(f_\theta(I) \cdot h_\eta(S)), \tag{6.2}$$

where η represents parameters of D_r, θ represents the parameters of the CNN, f and h are embedding functions of image and caption respectively, $< \cdot >$ is the dot product, S_T is the true caption for I from the training set, S_G is a generated caption from G for I and S_N is a "irrelevant-caption" from the training set that does not belong to I. α and β are balancing coefficients.

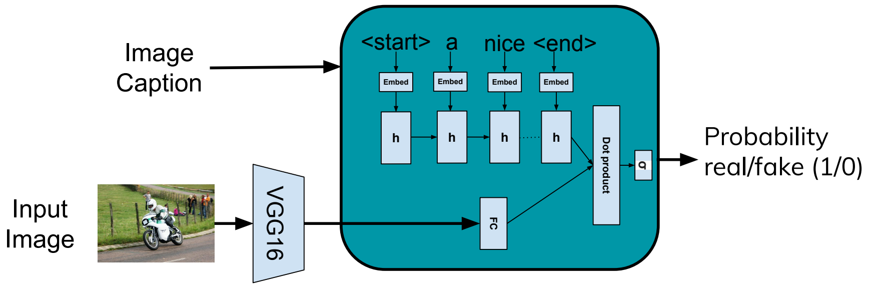


Figure 6.5: The first discriminator is an LSTM network. It takes two inputs: The caption generated by generator and the encoded image from VGG16. It processes these inputs to provide an estimate of the relevance of the caption, given the image.

6.3.4 *Discriminator-2*

caption sentiment discriminator

The second discriminator D_s, takes the generated caption from G, the input sentiment vector and assigns a reward for each of the tokens generated by the generator. Figure 6.6 shows the brief illustration of D_r. Our experiments showed that a pre-trained sentiment classifier can also be used with our modified objective function[1]. D_s provides a high reward if the computed sentiment is the same as the expected sentiment and punishes G for deviations. Thus, the reward from D_s can be defined as follows:

$$R_{D_s}(S, \omega) = \mathbb{E}_{S \sim S_G}[\delta_{wp} \log f_p(S) + \delta_{wn} \log f_n(S)], \tag{6.3}$$

[1] We used sentiment classifier provided by TextBlob (https://textblob.readthedocs.io/en/dev), which provides a sentiment value in [-1,1]

$$f_p(t) = \begin{cases} 1, & s(t) > 0.5, \\ 0.8, & 0 \leq s(t) \leq 0.5, \\ 0.1, & s(t) < 0, \end{cases} \qquad f_n(t) = \begin{cases} 1, & s(t) < -0.5, \\ 0.8, & -0.5 \leq s(t) \leq 0, \\ 0.1, & s(t) > 0, \end{cases}$$

where $\omega \in \{p, n\}$ is the input sentiment, $s(t)$ is sentiment value of token t assigned by D_s and δ is the Kronecker delta. It should be noted that, after the discriminators are trained, their role is to provide a reward to each token in the caption generated by G.

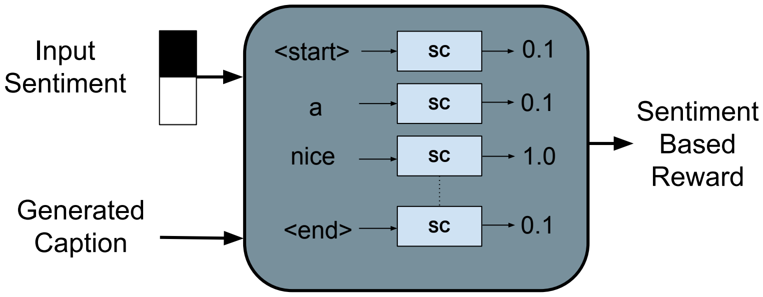


Figure 6.6: The second discriminator is a sentiment classifier network. It takes two inputs: The generated caption from generator and the binary sentiment variable. Given these two inputs, it provides a high reward if the generated caption has the intended sentiment (as indicated by the binary variable). A low reward, otherwise.

6.4 TRAINING

We divide the architecture training into two phases. In the first phase, the generator and discriminators are trained. After the first phase, the generator is able to generate words which are relevant to the image (without any specific language structure). For the second phase, the discriminators are frozen and the generator is further trained (via policy gradients) to incorporate sentiment and language variations in the caption. We found that this method of training increased the stability of the model and prevented the model from the "helvetica scenario" or mode collapse [52].

two phase training

6.4.1 *Phase 1*

The generator G in this setup was pre-trained with maximum likelihood estimation technique for e_g epochs. This pre-training was done in order to stabilize the gradients. We reach a stage where the generator starts generating some relevant words related to the image. The discriminator D_r was then trained using this generator for e_r epochs (with the loss function in Equation 6.1). Although, it can be argued that this is not truly an alternating adversarial training, we did not

adversarial training

find any significant difference in variations of captions with alternating adversarial training. Moreover, with our approach it reduced the training time (in terms of number of required epochs) of Generator. The noise variable z was set to a 1024 sampling from $\mathcal{N}(0, 1)$.

6.4.2 *Phase 2*

policy gradients

We used a policy gradient approach to further train G wherein the discriminators D_r and D_s act as reward agents. This means that G needs to generate captions through which it can maximize the rewards given by D_r and D_s. Therefore, the Phase-2 loss of G can be formulated as follows:

$$L_G(I) = \mathbb{E}_{S_g \in S_G}[-\gamma_1 \cdot R_{D_r} - \gamma_2 \cdot R_{D_s} + \gamma_3 \cdot \Omega(S_g, S_t)], \qquad (6.4)$$

where, $\gamma_1, \gamma_2, \gamma_3 \in (0, 1]$ are the balancing coefficients learned from the validation set and Ω is a regularizing term used to prevent the discriminator from collapsing to trivial patterns. We found that setting Ω to cross-entropy function (between generated caption S_g and true caption S_t of the image) gave the best results.

Steps of the two-phase training are enumerated in algorithm 3. Having described the architecture and the training process in detail, we move on to the details of the dataset in the next section.

6.5 DATASET: SENTIMENT ENHANCED MSCOCO

adjectives added to MSCOCO

MSCOCO [118] is an image-caption dataset containing 150,000 image-caption pairs in total (train, validation and test). At present, MSCOCO is also the most preferred dataset for state-of-the-art image captioning research [11, 16, 41, 42, 55, 119]. Therefore, we chose to use MSCOCO as this gave us a clear way to compare our results against the state-of-the-art. Although, MSCOCO is the benchmark dataset for captioning models, the captions provided are quite objective and clearly lack the sentiment dimension. A sentiment classification showed us that there were only 29,521 and 26,851 captions with a positive and negative sentiment respectively. The rest, 61,915 were neutral captions. The sentiment captions dataset from [119] (with 998 images) was found to be too small for our training task. To overcome these challenges, we decided to modify each of the nouns present in the MSCOCO dataset with a suitable positive or negative adjective. The intention was to enhance the sentiment value of the training set. Rather than randomly adding positive and negative adjectives, we used the adjectives from *aspects-DB*l [120] to find the list of suitable adjectives for each noun. We used the 2017 train/val split of MSCOCO which consists of 118,287 training images and 5,000 validation images[2]. For each of these im-

[2] MSCOCO does not have ground-truth captions for the test set

Algorithm 3: Two-Phase Training of the architecture. In the first phase (step 1 - 2), we train the generator, G and discriminator, D_r. In the second phase (step 3-13), we switch to Reinforcement Learning and train only G using policy gradients. In the second phase D_r and D_s act as reward agents for policy gradients.

```
   Data: Sentiment Enhanced MSCOCO Training set $S$
   /* Phase-1 Training                                          */
1  Train Generator $G$ with MLE on $S$
2  Train Discriminator $D_r$ using samples generated by $G$ and $S$
     with loss from Equation 6.1 ($L_{D_r}$)
   /* Phase-2 Training                                          */
3  for 100 epochs do
4      for Each Image in Dataset do
           /* Policy Gradient Training with $D_s$               */
5          Generate a sequence with $G$ with positive latent code
             vector.
6          Get reward for each token from the $Ds$
7          Generate a sequence with G with Noise Variable with
             negative latent code vector.
8          Get reward for each token from the $D_s$
           /* Policy Gradient Training with $D_r$               */
9          Sample noise vector $z$ w.r.t. the sentiment in
             ground-truth caption
10         Generate a sequence with $z$
11         Get reward for each token with $D_r$ using
             Monte-Carlo-Rollouts
           /* MLE Regularization                                */
12         Calculate Cross Entropy Loss for ground truth caption
             with $z$
13         Calculate total loss $G$ parameters
```

ages, there are 5 captions in the dataset. Following the sentiment enhancement, we processed each of these captions similar to [55]:

pre-processing

1. Remove all the non-alphabetic characters apart from comma.
2. Convert all the words to lower-case.
3. Add a START ($< start >$) and END ($< end >$) token at the beginning and end of each caption.
4. Remove all the words with the frequency of less than 5 in training and validation set combined.

This gave us the vocabulary size of 10,496 words. All the words that were not in the vocabulary were replaced with a token $< unk >$. We used the maximum sequence length of 16 and thus truncated all captions up until this length and padded the shortened sequences with token $< pad >$. After the modification, the sentiment-enhanced MSCOCO contained 50,303 positive and 67,981 negative image-caption pairs.

6.5.1 *Hyperparameters*

empirically determined parameters

For Deep Learning approaches, in most cases, hyperparameters are arrived at in an empirical fashion. Although, there have been different *rule-of-thumb* approaches, they still need to be tweaked through experiments to suit ones need. Therefore, all deep learning works In this section we describe the set of parameters which were empirically determined based on the validation set. The hidden dimensions of the generator and discriminator LSTM networks, h_g, h_d were both set to 512. The VGG16 network was used as the feature extractor for images with the feature vector $f \in \mathbb{R}^{4096}$. The noise vector z was from R^{1024}. The coefficients for Equation 6.1, α, β were set to 1. The coefficients for Equation 6.4 , $\gamma_1, \gamma_2, \gamma_3$ were set to $1, 1, 0.5$ respectively. For the first phase training, epochs e_g and e_r were $50, 30$ respectively. For the second phase, e_g was 100.

6.6 RESULTS

Since our work addresses the dimensions of sentiment and variability, the results were evaluated both quantitatively as well as qualitatively. Quantitative evaluation usually involves reporting conventional scores of BLEU [64], METEOR [66], ROUGE [65] and CIDER [67] against the ground truth. Qualitative evaluation uses human subjects to evaluate the generated captions for sentiment and grammar.

	Metric	SnT[11]	SCap[119]	Ours		
				c=1	c=5	c=10
Positive Captions	BLEU-1	0.620	0.567	0.547	**0.621**	**0.656**
	BLEU-2	0.437	0.365	0.346	0.406	**0.439**
	BLEU-3	**0.306**	0.240	0.220	0.267	0.295
	BLEU-4	**0.218**	0.164	0.144	0.181	0.202
	METEOR	0.219	0.199	0.185	0.209	**0.221**
	ROUGE_L	0.473	0.443	0.418	0.469	**0.488**
	CIDER	**0.752**	0.545	0.461	0.591	0.631
Negative Captions	BLEU-1	0.620	0.572	0.570	**0.645**	**0.676**
	BLEU-2	0.437	0.367	0.362	0.428	**0.463**
	BLEU-3	0.306	0.246	0.234	0.287	**0.319**
	BLEU-4	0.218	0.164	0.151	0.191	**0.219**
	METEOR	0.219	0.200	0.199	**0.222**	**0.235**
	ROUGE_L	0.473	0.447	0.445	**0.483**	**0.504**
	CIDER	**0.752**	0.516	0.509	0.627	0.688

Table 6.1: Conventional metrics for Show n Tell (SnT), SentiCap (SCap) (for both positive and negative captions) and our model (with 1, 5 and 10 generated captions). Even though our objective is not to maximize conventional scores, we still outperform both objective and sentiment models in most of these scores as we increase the variations. SnT scores are the same for Positive and Negative captions because they generate a neutral caption.

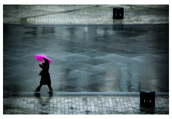
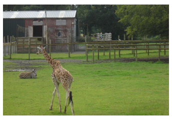
Image	Captions
	+ a proud woman walking down the street holding a colorful umbrella.
	+ a attractive person walking across a street holding a umbrella.
	+ a great person walking with a umbrella on top of a street.
	– a dangerous person walking down the street in the rain.
	– a evil person walking across with a umbrella.
	– a dangerous person walking holding a pink umbrella.
	+ a beautiful giraffe standing on top of a lush green field.
	+ a beautiful giraffe standing near a tree in a field.
	+ a wonderful giraffe in a field with a bird in the background.
	– a sad giraffe standing in a field next to a bush.
	– a sick giraffe standing in a lush green field.
	– a sick giraffe standing in a field next to a tree.
	+ a white and blue great plane is on a runway.
	+ a popular passenger jet is parked on the runway.
	+ a large white great airplane sitting on a runway.
	– a white and blue jet sitting on a wrong runway.
	– a expensive passenger jet is parked on the runway at an airport.
	– a fake airplane that is sitting on a runway.

Table 6.2: Positive and negative captions generated by our model. For each sentiment (+/–), there are three variations shown.

6.6.1 *Quantitative Results*

Classical scores like BLEU, METEOR, ROGUE and CIDER are generally evaluated by matching n-grams between target and the generated captions. Therefore, a higher score would suggest that the generated caption is closer to the target sentence. Even though our models are not trained to emulate the ground truth (in turn maximize the benchmark scores), we would like to report these scores to show that we can still outperform the state-of-the-art, simply by increasing the variations for our captions. To compare against the state-of-the-art for objective captioning, we use the "Show n Tell"[11] model. For comparison against the state-of-the-art for sentiment captioning, we use "SentiCap"[119] model. We use the test set published by [119] which contains 433 positive and 433 negative image-caption pairs. The results show that even though we did not train the model according to the conventional criteria, we competitively outperform the state of art as shown in Table 6.1. As we increase our generated captions (c=1,5,10), we also get some variations which are similar to the ground truth, thereby achieving high value for these scores. Furthermore, we have to use the same underlying vocabulary to generate variations. The nouns present in the ground truth caption, like park, kitchen, man etc. are present in the generated captions/variations as well (although their positions are different) providing a boost to these values.

outperforms the state-of-the-art

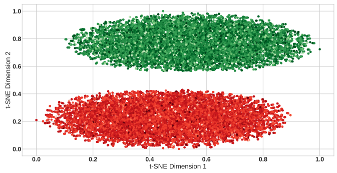
In order to determine whether the z vector truly encodes the intended sentiment, we created 30,000 pairs (15,000 positive/negative each) of encoded z vectors and calculated the sentiment of the generated caption. We then used t-SNE algorithm to visualize these vectors. Figure 6.7 (left) shows the distribution of these vectors. As can be seen, there are two clusters that represent two different sentiment encoded z vectors. The colors indicate the sentiment of the generated caption. Each of the two clusters are dominated by a single output sentiment (positive or negative) as indicated by their color coding. Figure 6.7 (right) shows the confusion matrix w.r.t the sentiment. As seen from the confusion matrix, the overall accuracy of the intended sentiment is 93.19%. These results (visual and confusion matrix) indicate that the encoding scheme is effective and achieves the intended sentiment in the generated caption.

qualitative visualization

6.6.2 *Human Evaluation*

As our task involves generating variable and sentiment captions, a fair evaluation is only possible through humans. The evaluation should include the following judgments about:

1. The validity of the caption (given the image).
2. The sentiment of the caption.



		Generated	
		Pos	Neg
Target	Pos	13,498	1,502
	Neg	541	14,459

Figure 6.7: The plot(left) shows the t-SNE projection of the z vectors onto a 2-D space. Each of the two clusters formed by z vectors are dominated by a single sentiment (of generated caption). The right side shows the confusion matrix w.r.t the sentiment (expected vs generated sentiment).

sentiment and validity check

To conduct the human evaluation of the generated captions, we randomly sampled 200 *held-out* images from the validation set of MSCOCO. We generated 3 positive and 3 negative captions per image. Each image-caption pair was evaluated by 3 subjects and a majority vote decided the final answer. Through the entire experiment, we collected 3,600 responses from human subjects. Each subject was shown an image and 6 captions (3 positive, 3 negative) but in a random order. For each caption, given the image, subjects were then asked to answer the following questions:

1. Is this a valid caption for the given image?
2. What is the general sentiment of the caption: positive, neutral, negative?

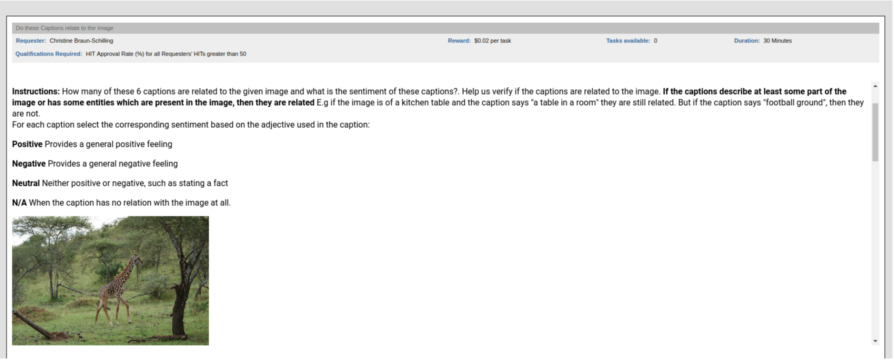
We used the popular crowd sourcing platform, Amazon Mechanical Turk[3] to conduct our human evaluation. Figure 6.8 shows the screen capture of the task. To ensure annotation quality, we set up the following:

1. A validation set which would be run to ensure that annotators did not cheat on the task.
2. Only contributors with a minimum rating of 75% were allowed to participate.

controlled variations possible

From the 3,600 responses collected, 77.7% of the generated captions were voted valid and having the intended sentiment. This indicates that the captions (and the variations) from our model were of high quality (semantically relevant) and had the intended sentiment. It implies that it is possible to control properties of generated variations through external contexts with a GAN. Table 6.2 shows few examples that we used for this task. In 10.3% of the cases, the subjects voted for a "neutral" sentiment. An analysis of such cases showed that it was because of the generated adjective not being strong enough to convey the sentiment.

[3] www.mturk.com



Do these Captions relate to the Image

Requester: Christine Braun-Schilling Reward: $0.02 per task Tasks available: 0 Duration: 30 Minutes

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than 50

Instructions: How many of these 6 captions are related to the given image and what is the sentiment of these captions?. Help us verify if the captions are related to the image. **If the captions describe at least some part of the image or has some entities which are present in the image, then they are related** E.g if the image is of a kitchen table and the caption says "a table in a room" they are still related. But if the caption says "football ground", then they are not.
For each caption select the corresponding sentiment based on the adjective used in the caption:

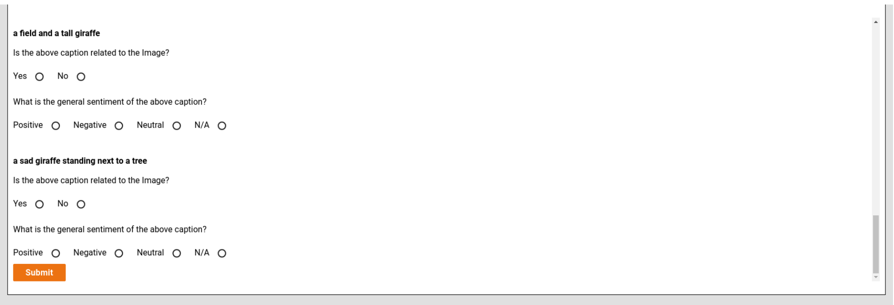
Positive Provides a general positive feeling

Negative Provides a general negative feeling

Neutral Neither positive or negative, such as stating a fact

N/A When the caption has no relation with the image at all.

(a) The first part contains the questionnaire shown to the user including the image.



a field and a tall giraffe

Is the above caption related to the Image?

Yes ○ No ○

What is the general sentiment of the above caption?

Positive ○ Negative ○ Neutral ○ N/A ○

a sad giraffe standing next to a tree

Is the above caption related to the Image?

Yes ○ No ○

What is the general sentiment of the above caption?

Positive ○ Negative ○ Neutral ○ N/A ○

(b) The second part contains the caption variations and corresponding questions

Figure 6.8: Screen capture of the task given for user evaluation. Shown in two parts here to accommodate the large length.

6.7 CONCLUSIONS

We started with the motivating question:

> *"Can machines generate captions with variations and control its properties?"*.

To answer this question, we needed to fix a property to demonstrate our results. We chose sentiment as a property to be controlled by the model. The choice of sentiment as a property allowed us to easily verify our results quantitatively against the state-of-the-art. We then took a generative approach to combine sentiment and variation in a single model. Our architecture was similar to a GAN, but with one generator and two discriminators. The first discriminator assured the relevance of caption and the second checked for the sentiment. In order to encode sentiment information in the input of the GAN, we used the noise variable z. We designed an encoding scheme for z such that the model can correlate the encoding with the intended sentiment. We also found that it was non-trivial to train a generative model for language because of the non-differentiable nature of language. To alleviate the training hurdles of a normal GAN with languages, we used the policy gradients algorithm from Reinforcement Learning. A two-phase training approach was introduced to stabilize the training process. We evaluated our model both quantitatively and qualitatively. We showed that our model competitively outperforms the two state of the art models (for objective and sentiment captions) for image captioning. To further evaluate the results, we also performed a human evaluation and showed that 77.7% of the generated captions are valid with intended sentiments. Our results imply that it is possible to generate variable-sentiment captions with good degree of accuracy.

7 APPLICATION: REAL-TIME CAPTIONING SYSTEM

Proof-of-Concept systems (demo) not only help to showcase ones research but also make it easier to explain ones experiments/models. Such visual systems also provide ideas to further improve the research direction and carve future directions. Additionally, developing such systems is chance to test the limitations of deep learning frameworks. We describe a real-time captioning demo, named *Captittude*, deployed in the web (The system can be accessed at `http://www.madm.eu/demo2/caption/`). It contains both the CAST and STEM captioning models. Given an image, these models process them in parallel and provide their captions. Figure 7.1 shows a brief overview of the system.

importance of demos

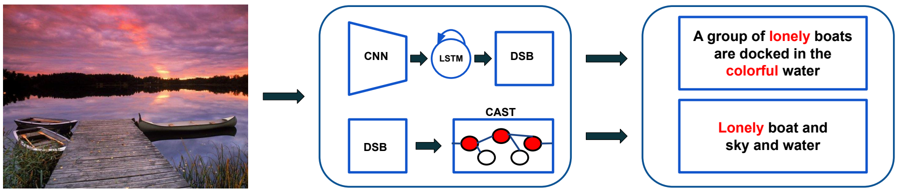

CNN
LSTM
DSB
CAST
DSB
A group of lonely boats are docked in the colorful water
Lonely boat and sky and water

Figure 7.1: Brief illustration of Captittude pipeline. The input image is sent to the server, where CAST and STEM models process them. Each of the captioning models then return a caption for the image.

7.1 ARCHITECTURE

The system contains a python based backend and bootstrap based front end user interface. It also provides a user upload facility with the help of DropZone. With this facility, viewers can upload pictures into the system and get captions in return. We describe the python and bootstrap interfaces in the following section.

7.1.1 *Backend*

The backend architecture of Captittude is based on python. Python provides seamless integration with web services through the *CherryPy* module. The deep learning models like DeepSentiBank, CAST and STEM have been implemented in *Tensorflow*. CherryPy[1] is a python module that provides a minimalist web framework. This allows developers to create web related demos quickly. Therefore, this was also the first choice for Captittude development as well. In order to implement the

python based

[1] https://docs.cherrypy.org/en/latest/

deep learning part of the captioning models, we used the popular deep learning framework, Tensorflow[2]. Tensorflow (released by Google) has been used by researchers in developing neural network models for their experiments. The optimized GPU versions have been helpful in reducing the training time of neural network models.

7.1.2 *Frontend*

ease of navigation

The user interface and the landing page have been designed with BootStrap library (javascript) providing an enhanced browsing experience. Bootstrap is an open-source toolkit for web development and provides easy prototyping of applications. The user upload functionality has been developed using DropZone javascript library. This allows users to drag-n-drop images into the field for automatic seamless upload. Figure 7.2 shows a brief overview of the architecture.

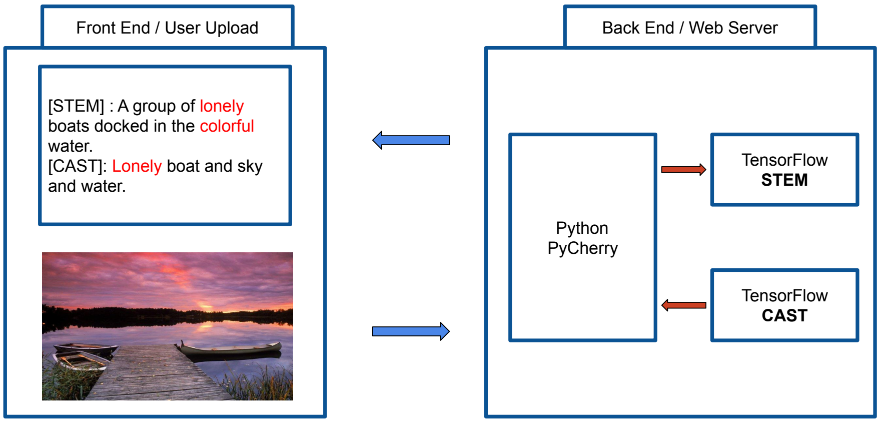

Front End / User Upload
[STEM] : A group of lonely boats docked in the colorful water.
[CAST]: Lonely boat and sky and water.
Back End / Web Server
Python
PyCherry
TensorFlow
STEM
TensorFlow
CAST

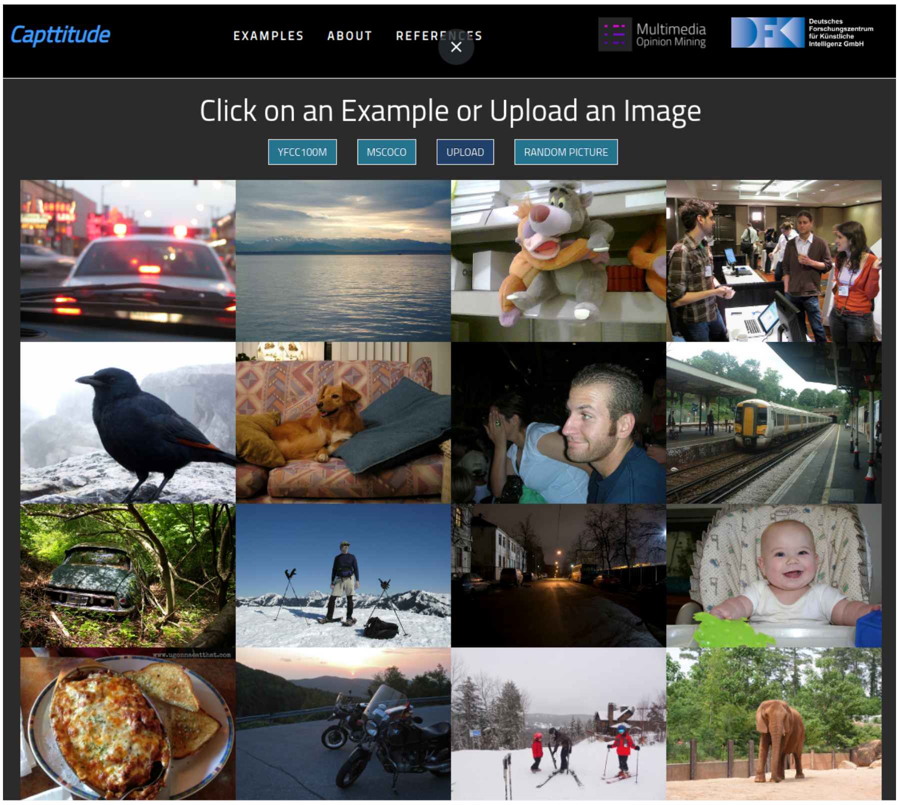
Figure 7.2: Overview of internal architecture. The frontend is a bootstrap page allowing users to upload images. Backend consists CherryPy server communicating with Tensorflow models.

7.2 YFCC100M & MSCOCO EXAMPLES

preloaded data

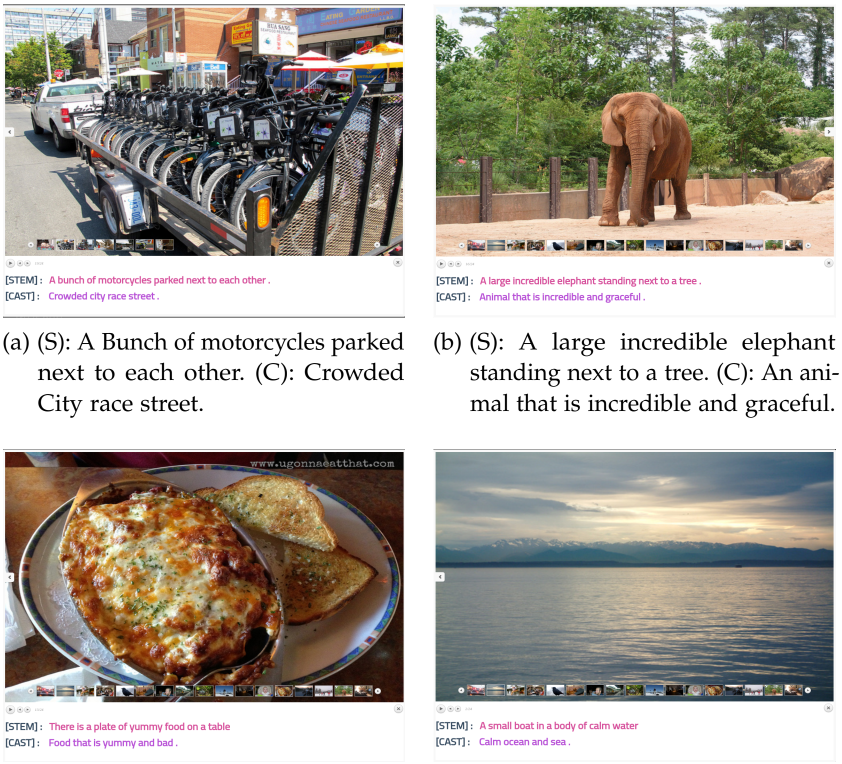
Random YFCC100M and MSCOCO images are preloaded into the system for users to get an overview about the system. Each time a user clicks a preloaded image, the image is sent to the CherrPy server and the captioning models process the image and provide their captions. Figure 7.3 shows a screen capture of the preloaded examples. Users are also allowed to search random samples from each of these datasets.

[2] https://www.tensorflow.org/


Capttitude
EXAMPLES
ABOUT
REFERENCES
Multimedia Opinion Mining
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
YFCC100M
MSCOCO
UPLOAD
RANDOM PICTURE

Click on an Example or Upload an Image

Figure 7.3: Preloaded example images in the system. YFCC100M and MSCOCO examples are loaded for users to browse through them. However, each time a user clicks an image, the system sends it over to the server and the captions are generated on-the-fly



[STEM] : A bunch of motorcycles parked next to each other .
[CAST] : Crowded city race street .

(a) (S): A Bunch of motorcycles parked next to each other. (C): Crowded City race street.

[STEM] : A large incredible elephant standing next to a tree .
[CAST] : Animal that is incredible and graceful .

(b) (S): A large incredible elephant standing next to a tree. (C): An animal that is incredible and graceful.

[STEM] : There is a plate of yummy food on a table
[CAST] : Food that is yummy and bad .

(c) (S): There is a plate of yummy food on the table. (C): Food yummy and bad.

[STEM] : A small boat in a body of calm water
[CAST] : Calm ocean and sea .

(d) (S): A small boat in a body of calm water. (C): Calm ocean and sea.

Figure 7.4: Examples of generated captions by the Captittude system. Both the captions from CAST and STEM are shown denoted by C and S respectively.

7.3 SHOWCASE

presentation & feedback

The demo was also showcased at CeBIT-2017[3], which is the largest international computer expo. We received a lot of positive feedback from the audience at the event. Figure 7.4 shows few of the generated captions from by the system. As a future enhancement, we also plan to integrate GAN models and RL algorithms in the backend. This will allow the system to take the user feedback into account and improve the system over time.

[3] https://en.wikipedia.org/wiki/CEBIT

Part V

SUMMARY AND OUTLOOK

8 SUMMARY

With the increased connectivity across the world brought-in by internet and social media platforms, our communication methods have also become multimodal. We now use both the visual and language medium to express our personalities, ideas, views, opinions etc. The visual contents across the internet/social-media are often accompanied with a textual component called as *caption*. For images, the caption usually provides a description of the image or is connected to the context of an image. There are multiple reasons as to why we need captions. These reasons range from Goffman's theory of *Presentation of Self* to *Effective social communication*. We also pointed out the increased efficiency of human computer interaction, if machines could generate image captions which are *human-like*. We call this field as *Affective Image Captioning*. Such abilities would not only allow machines to assist humans, but also improve its ability to effectively interact with a human. Therefore, we began with the question:

captions needed for communication

automatic caption for better human computer interaction

> Can machines generate human-like image captions?

To understand the major properties that constitute human-like captions, we performed a large-scale human caption study (Chapter 3). This study spanned across the USA, Canada, England, Ireland, Australia, and New Zealand. The data used for this study was sampled from the YFCC100M dataset. Three thousand image-caption pairs were sampled and shown to the annotators. The annotators were asked a set of questions regarding the image-caption pair to evaluate the focus, visibility, intent, and meaning of the caption. The study ran for a span of six months with 298 participants joining at various stages. We performed a statistical and qualitative evaluation of the results. The analysis show us that there are three major properties and few minor properties that constitute the core component of human-like captions. The major properties are: *Subjectivity*, *Sentiment*, and *Variations*. We show that the minor properties of Humor and Sarcasm require prior contextual knowledge and hence are out of the scope of this work. The outcome of this study is published as a dataset, *captions-DB*, available freely to the public. We address each of the major properties with Deep Learning models.

understand human caption properties

In Chapter 4, we brought forth the challenges in capturing **subjectivity** arising from difference in interpretations. We also point out the drawbacks of capturing subjectivity in a holistic manner (e.g., in the form of an adjective-noun pair for the whole image). In order to move away from holistic approaches and to disentangle subjectivity into

fusion techniques for subjectivity

FAV & aspect detection

Tensor Fusion better than XResNet

components, we propose the task of *aspect detection* and the FAV model. In this model, we capture subjectivity in three components of *Focus*, *Aspect*, and *Value*. The focus is the attention point inside the image, typically a noun. The aspect is a dimension of evaluation of the focus, e.g., age, activity etc. Aspects are usually arranged as set of opposing adjectives. The value provides a binary evaluation as to which side of the opposing adjectives do the aspect belong to. To implement the FAV model, we propose different architectures: Logistic regression, XResNet, and Fusion methods. A new method of fusion called as *Tensor Fusion* is also introduced. An evaluation of these architectures on aspect detection and aspect-value prediction show that tensor fusion is a better approach across these tasks. A new dataset, *aspects-DB*, is released as a part of the aspect detection task and is freely available for download.

sentiment injection with neural networks

high scores with human evaluation

Chapter 5 addressed the second of the major properties (**sentiment**). We propose two models for sentiment injection into captions: CAST networks and the STEM model. The CAST network is a graphical model which captures the underlying language structure. We show that, using neural networks to capture the adjectives (inside an image) and using these adjectives (inside the graph) to create a path, we can create captions with sentiment. The chapter also introduces the STEM model which can perform sentiment injection. In this method, we start with a neutral caption and inject adjectives at suitable positions. In order to find these positions, we use the Word2Vec network. To perform an evaluation of these models, we presented 200 random images from YFCC100M to human evaluators and asked them to choose one caption among the two options provided (which they thought were human generated). The results show us that in 62.5% of the cases, at least one subject chose our caption. We also discussed about the *ACMMM Grand Challenge* and their human evaluation methods, where these models scored a high score of 93%.

generative models for controlled variations

outperforms state-of-the-art

real-time demo

The last property of **variation** was discussed in Chapter 6. Here we took a generative approach to create variations with sentiment. A modified version of CGAN is used with one generator and two discriminators. This modified architecture allows fusion of an external signal, *context*, into the GAN. As language training is difficult with a GAN, we took the help of policy gradients algorithm from Reinforcement Learning. To further stabilize the training, we introduce a two-phase training approach wherein the model is trained in an adversarial manner in the first phase and with policy gradients in the second phase. Using quantitative evaluation, we show that our model is able to outperform the state-of-the-art sentiment captioning methods. A human evaluation also proves that 77.7% of the generated variations are valid with the intended sentiment. Chapter 7 discussed a real-time proof-of-concept application of the above mentioned models (developed in Tensorflow). The system can take in user images and provide captions in return.

Summarizing, in the process of answering our research question, we have achieved the following goals we set in Chapter 1:

- HUMAN CAPTION STUDY. We conducted a large-scale human caption study in order to identify the properties of human generated captions
- DEEP LEARNING MODELS. For the major properties of subjectivity, sentiment, and variations we proposed predictive and generative deep learning models. We evaluated them against the state-of-the-art along with human evaluation.

contribution summary

- DATASETS. Two important datasets were also published. First, *captions-DB* was the result of the human caption study and can be used for researching into other aspects e.g., *caption interpretation*. Second, *aspects-DB* was compiled to solve the problem of positive bias in the existing datasets and to introduce subjective attributes for aspect detection.

The datasets published as a part of this work can be downloaded at `http://madm.dfki.de/downloads`.

9 FUTURE OUTLOOK

Having provided a summary in Chapter 8, we shall conclude this work by listing out some of the promising directions for future. The task of automatic image captioning has made few important strides into human-like caption generation. However there is much more to achieve in this direction. Given the vast array of potential applications, we benefit immensely if machines can generate language as humans. The future directions of this work can broadly be divided into four categories.

extend fusion

Searching fusion strategies. An important message from the FAV model was that the way of combining information can drastically change various properties of neural networks. This also affects the sensitivity towards various hyper-parameters and generalization ability. Therefore, an important question that arises is: *what is the best way to fuse information?* This question has become more relevant given the present situation, where researchers often use concatenation as the default fusion strategy and focus more on tuning hyper-parameters. Further experiments need to be designed to arrive at a conclusion of the fusion strategies for different neural network architectures.

adjective rearrangement

Extending the FAV model. We would also like to extend the FAV model to other applications like subjective information retrieval and into the domain of visual question answering. The model shall also benefit if the adjectives are arranged on an increasing *scale of strength*. This means that the value predicted by the FAV model could be used to choose the right adjective from this scale.

Reinforcement Learning methods

Reinforcement learning for GAN. In order to stabilize the GAN training we have used the policy gradients approach from RL. RL offers promising directions for training generative models. For example, algorithms like *actor-critic* algorithms need to be explored to check if they can better train GAN architectures. Furthermore, we have only explored one form of encoding sentiment in this work. This encoding scheme shows us that it is possible for a generative model to take external context into account. Further experiments need to designed in order to arrive at most suitable form of information encoding inside a GAN.

multiple contexts

Extension of contexts. We introduced the single context fusion in a GAN, as a part of generating variations. This context was given in the form of sentiment information. We intend to extend this to multiple contexts. For example, the model could take sentiment, length, style information as contexts and generate captions that conform to all these contexts. However, this requires that we first establish a firm training method for the modified GAN architecture than can take multiple contexts as input.

enhanced applications

Last but not the least, the algorithms developed for affective image captioning in this work have a potential to be developed into interesting applications. *Chatbots* are a popular category of applications that are gaining momentum since the past few years. Brands have started using chatbots as a part of their marketing strategies. Affective captioning chatbots can prove to be beneficial in marketing campaigns across the social media platforms. *Automatic Personal Diary* is another application where the user can upload pictures (from vacations, events) as a batch. The system then arranges and writes captions for these pictures which can later be used for retrieval or shared with colleagues.

Part VI

APPENDIX

A CAPTION EVALUATION METRICS

A.1 BLEU

BLEU [64] is the most popular metric in caption evaluation. It has its origins from the machine translation community. We first define a modified version of precision called *modified n-gram precision* which matches test sentence's n-grams only as many times as they are present in the ground-truth. This modified n-gram precision can be defined as follows:

$$P = exp\,(\sum_{n=1}^{N} w_n\, log(p_n)), \tag{A.1}$$

where $w_n = 1/n$.

Having defined the precision, we now need to introduce the recall for n-grams. But this is tricky, as there can be multiple ground-truth sentences. Intuitively, a longer test sentence will have a higher recall. At the same time our modified precision P penalizes arbitrary repetition of words. Therefore, we introduce recall by penalizing shorter sentences. A new multiplicative term BP for the precision is define here:

$$BP = \begin{cases} 1, & \text{if c > r,} \\ exp\,(1 - \frac{r}{c}), & \text{otherwise,} \end{cases} \tag{A.2}$$

where c is the test sentence length and r is the average length of the ground-truth sentences.

A.2 ROUGE

As the name indicates, ROUGE [65] is based only on recall and is generally used for summary evaluation in machine translation. The ROGUE metric has the following types:

1. ROUGE-N is the recall based on n-grams. For example, ROUGE-1 counts recall based on matched uni-grams. For a given n it counts the total number of n-grams across the ground-truth sentences and counts the fraction of them present in test sentence.

2. ROUGE-L/W/S are based on the Longest Common Subsequence (LCS), weighted LCS and skip-bigram co-occurrence count respectively. Here we use the F-score values instead of recall. For ROUGE-L, the F-score is calculated as follows:

$$P = \frac{LCS(T,G)}{Length_T}, \tag{A.3}$$

$$R = \frac{LCS(T,G)}{Length_G}, \tag{A.4}$$

where T and G are the target and ground-truth sentences respectively. $Length_x$ is the length of x. We calculate F as the weighted harmonic mean of P and R as:

$$F = \frac{(1+b^2)\,RP}{R \ + b^2 P}. \tag{A.5}$$

For ROUGE-W, for calculating the weighted LCS, we track the length of consecutive matches along with the LCS. For ROUGE-S, a skip-bigram is the any pair of words allowing for arbitrary gaps. We calculate the precision and recall on skip-bigrams as mentioned above.

A.3 METEOR

The challenge with BLEU is that the BP value uses lengths which are mean for the entire ground-truth sentences. Therefore, this affects the scores of individual sentences. To address this issue, METEOR [66] proposes modification to the precision and recall functions. It replaces the F-score with a weighted F-score based on unigrams and penalizes incorrect word order.

To calculate the Weighted F-score, we find the largest subset of mapping that can form an alignment between test and ground truth sentences. To achieve this, apart from exact matches, we perform word-stemming and replace words with their synonyms. Having found such an alignment where m is the number of mapped unigrams between the two sentences, we define the precision, recall and F-measure as following:

$$P = \frac{m}{Length_T}, \tag{A.6}$$

$$R = \frac{m}{Length_G}, \tag{A.7}$$

$$F = \frac{PR}{\alpha P + (1-\alpha) R}. \tag{A.8}$$

To encourage the word ordering and penalize arbitrary words, the following penalty function is introduced:

$$f_{penalty} = \gamma \left(\frac{c}{m}\right)^{\beta}, \ 0 \leq \gamma \leq 1, \tag{A.9}$$

c is the number of matching subsets, m is the total number of matches, $\beta and \gamma$ are hyperparameters

The final Meteor score M is calculate as:

$$M = (1 - f_{penalty}) * F. \tag{A.10}$$

A.4 CIDER

The CIDER [67] metric follows the same intuition as BLEU but is specially designed for image captioning (in contrast to machine translation metrics above). For an image I_i, the test sentence c_i is matched against the ground-truth sentences for I_i, represented by $S_i = \{s_{i1}, s_{i2}, \ldots, s_{im}\}$ It starts with stemming all the words in the test and ground-truth captions. Now, a measure of consensus would count how often n-grams in test caption are present in ground-truth. n-grams that are commonly present through all images in the dataset should be given lower weights as they probably are less informative. To achieve this, CIDER performs a Term Frequency Inverse Document Frequency (TF-IDF) for each n-gram. The TF-IDF weighting $g_k(s_{ij})$ for n-gram ω_k is calculated as:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} log\left(\frac{|I|}{\sum_{I_p \in I} min(1, \sum_q h_k(s_{pq}))}\right), \tag{A.11}$$

where $h_k(s_{ij})$ denotes the number of times n-gram ω_k occurs in s_{ij}, Ω is the vocabulary of all n-grams and I is the set of all images in the dataset.

The first term measures the TF of each n-gram ω_k and the second term measure the rarity of ω_k using IDF. The TF gives high weights to frequently occurring n-grams in ground-truth while IDF reduces the weights of n-grams that occur across all images in the dataset. Intuitively IDF gives a word importance score by discounting words which are popular and hence less informative.

Now the CIDER score can be computed as mean cosine similarity between test and ground truth captions:

$$CIDER_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i).g^n(s_{ij})}{|g^n(c_i)||g^n(s_{ij})|}, \tag{A.12}$$

where, $g^n(c_i)$ is a vector formed from $g_k(c_i)$ corresponding to all n-grams of length n and $|.|$ its magnitude.

B

CROWD-SOURCING: LESSONS LEARNED

Verification of subjective aspects is an important part of research when it comes to multimedia research. Although, quantitative metrics exist, these are often limited to well-defined subsets of possible scenarios and can be overcome by tuning hyperparameters. This leads to unreliable results which are not considered while designing the experiments. Therefore, qualitative evaluation are still in need for multimedia topics. However, traditionally, such evaluations are often expensive, require organizational skills and are time-consuming. Crowd-sourcing alleviates the challenges in traditional qualitative evaluation by outsourcing the tasks via Internet to a global worker pool. Although effective, Crowd-Sourcing is not a straight forward approach and brings in its set of challenges. Extra considerations need to be taken in order to gain better results, because of the virtual environment and traditional differences. We list out the pitfalls and best practices [87] in this chapter.

B.1 KEY CHALLENGES IN CROWD-SOURCING

Challenges in Crowd-Sourcing arise as a result of remote-setting of the task, heterogeneity of users, their hardware, environments etc. One has to not only consider the actual user ratings but also the non-standard test equipment like software compatibility of browsers, Internet speed etc.

B.1.1 *Limitations of Crowd-Sourcing*

A Crowd-Sourcing system can, in theory, be used for assessment of any stimuli and any type of subjective methodology. However, in practice, one is faced with several limitations. The main technical constraints are the bandwidth of the users which limit the interactivity of the experiment. Additionally highly interactive experiments also require equally good hardware support with the worker. This is particularly evident in video-captioning where users have to be shown a video clip. Experiments which require interaction among the workers are practically impossible with Crowd-Sourcing. This also applies to experiments which might require olfactory stimuli.

B.1.2 *User Reliability*

There are also cases where on some platforms user ratings are not very reliable. This means that additional tests need to designed in order to filter out low rated users. Technical errors also can occur as these web based experiments, leading to users receiving a different view of test conditions. Such cases can lead to users being flagged off as unreliable, although the answers were valid from users' point of view. Therefore, one needs additional monitoring mechanisms to monitor users' environments. Users may also cheat the system to maximize their incentives while minimizing their efforts and submit low quality work. Crowd-Sourcing platforms, usually provide a gold standard verification mechanism to address this issue. The tasks should be designed in such a way that there is no incentive for user to cheat. This implies that in a task the cheating should take approximately the same time as completing it in the right way.

B.2 BEST PRACTICES

B.2.1 *Popular Platforms*

There are several platforms which are available for researchers to design to conduct evaluations. However, Amazon Mechanical Turk (AMT) and CrowdFlower are the most popular in the community. These platforms differ in the features and fees that they offer. Although, CrowdFlower offers better design features like gold standard integration, it is designed for big corporations to carry out their internal surveys. Therefore, in the past five years, researchers and academic institutions have relied on AMT for their experiments. This also translates to AMT providing pre-designed templates which are suited for most experiments commonly required by the community.

B.2.2 *Incentive Design and User Bonuses*

Incentive Design focuses on improving the quality of results by improving the willingness of users to participate beyond pure financial gains. This usually happens through gamification of tasks helping more users to complete the task in shorter time. Furthermore, users who perform well are encourage further through a bonus payments rewarding their efforts. Most Crowd-Sourcing platforms at present provide the user-bonus mechanism, where researchers can decide to pay extra money to participant users (at a later point in time) after the tasks have been evaluated.

B.2.3 *Demographics*

There are several options to collect demographic data. A survey at the beginning of the experiment can be used to collect this data. However, it should be noted that users may also provide wrong information in such surveys. Therefore, platforms also provide demographic information on the users which can act at an additional validation of the data. Demographics can also be extracted from social networks, if available.

B.2.4 *Two Stage Experiment Design*

Even though some platforms provide an option for gold standard data, these are not implement with sufficient reliability mechanisms. Therefore, it is beneficial to design the experiment in two stages. In the first stage, the users are asked to take a simple test (also collect demographics) which is related to the experiment and hence evaluates their understanding of the tasks. This stage acts as an additional barrier to exclude under-performing users. The second stage, is the actual experiment which is taken only by users who have cleared the first stage.

Part VII

REFERENCES & CV

BIBLIOGRAPHY

[1] Andreas M Kaplan and Michael Haenlein. "Users of the world, unite! The challenges and opportunities of Social Media." In: *Business horizons* 53.1 (2010), pp. 59–68.

[2] VNI Cisco. "Cisco visual networking index: Forecast and trends, 2017–2022." In: *White Paper* 1 (2018).

[3] Erving Goffman et al. *The presentation of self in everyday life*. Harmondsworth London, 1978.

[4] Liam Bullingham and Ana C Vasconcelos. "'The presentation of self in the online world': Goffman and the study of online identities." In: *Journal of information science* 39.1 (2013), pp. 101–112.

[5] J Owyang. *Why automating social media marketing could change Facebook*. 2012.

[6] H Zebida. *TweetAdder: Simply the Fastest Way to Manage Twitter Account*. 2014.

[7] Alok Choudhary, William Hendrix, Kathy Lee, Diana Palsetia, and Wei-Keng Liao. "Social media evolution of the Egyptian revolution." In: *Communications of the ACM* 55.5 (2012), pp. 74–80.

[8] M Sarfraz and SM Ali J Rizvi. "Indoor navigational aid system for the visually impaired." In: *Geometric Modeling and Imaging (GMAI'07)*. IEEE. 2007, pp. 127–132.

[9] Derek Molloy, T McGowan, K Clarke, C McCorkell, and Paul F Whelan. "Application of machine vision technology to the development of aids for the visually impaired." In: *Machine Vision Applications, Architectures, and Systems Integration III*. Vol. 2347. International Society for Optics and Photonics. 1994, pp. 59–69.

[10] Chenyou Fan, Zehua Zhang, and David J Crandall. "Deepdiary: Lifelogging image captioning and summarization." In: *Journal of Visual Communication and Image Representation* 55 (2018), pp. 40–55.

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164.

[12] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3128–3137.

[13] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. “Densecap: Fully convolutional localization networks for dense captioning.” In: *arXiv preprint arXiv:1511.07571* (2015).

[14] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. “From captions to visual concepts and back.” In: *Proceedings of the IEEE Conference on CVPR*. 2015, pp. 1473–1482.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. “Show, attend and tell: Neural image caption generation with visual attention.” In: *International Conference on Machine Learning*. 2015, pp. 2048–2057.

[16] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. “Image captioning with semantic attention.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4651–4659.

[17] Philipp Blandfort*, Tushar Karayil*, Damian Borth, and Andreas Dengel. “Image Captioning in the Wild: How People Caption Images on Flickr.” In: *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*. ACM. 2017, pp. 21–29.

[18] Tushar Karayil*, Philipp Blandfort*, Jörn Hees, and Andreas Dengel. “The Focus-Aspect-Value Model for Explainable Prediction of Subjective Visual Interpretation.” In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. ACM. 2019, pp. 16–24.

[19] Philipp Blandfort*, Tushar Karayil*, Jörn Hees, and Andreas Dengel. “The Focus-Aspect-Value Model for Predicting Subjective Visual Attributes.” In: *International Journal of Multimedia Information Retrieval* (January 2020), pp. 1–14. DOI: 10.1007/s13735-019-00188-5. URL: https://link.springer.com/article/10.1007/s13735-019-00188-5.

[20] Philipp Blandfort*, Tushar Karayil*, Damian Borth, and Andreas Dengel. “Introducing concept and syntax transition networks for image captioning.” In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM. 2016, pp. 385–388. DOI: 10.1145/2911996.2930060. URL: https://dl.acm.org/doi/10.1145/2911996.2930060.

[21] Tushar Karayil, Philipp Blandfort, Damian Borth, and Andreas Dengel. “Generating Affective Captions using Concept And Syntax Transition Networks.” In: *Proceedings of the 2016 ACM on Multimedia Conference*. ACM. 2016, pp. 1111–1115.

[22] Tushar Karayil, Asif Irfan, Federico Raue, Jörn Hees, and Andreas Dengel. "Conditional GANs for Image Captioning with Sentiments." In: *Proceedings of the 28th International Conference Artificial Neural Networks, vol 11730*. Springer. 2019. DOI: 10.1007/978-3-030-30490-4_25. URL: https://link.springer.com/chapter/10.1007/978-3-030-30490-4_25.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[25] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[27] Sepp Hochreiter. "Untersuchungen zu dynamischen neuronalen Netzen." In: *Diploma, Technische Universität München* 91.1 (1991).

[28] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[29] Felix A Gers and Jürgen Schmidhuber. "Recurrent nets that time and count." In: *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*. Vol. 3. IEEE. 2000, pp. 189–194.

[30] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." In: *arXiv preprint arXiv:1406.1078* (2014).

[31] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. "Every Picture Tells a Story: Generating Sentences from Images." In: *ECCV*. 2010.

[32] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. "Composing Simple Image Descriptions using Web-scale N-grams." In: *CoNLL*. 2011.

[33] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. "Baby Talk : Understanding and Generating Image Descriptions." In: 2011.

[34] Yezhou Yang, Ching Lik Teo, Hal Daumé, and Yiannis Aloimonos. "Corpus-Guided Sentence Generation of Natural Images." In: *EMNLP 2011*. 2011.

[35] Desmond Elliott and Frank Keller. "Image Description using Visual Dependency Representations." In: *EMNLP*. 2013.

[36] Hao Fang et al. "From captions to visual concepts and back." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1473–1482.

[37] Rémi Lebret, Pedro H. O. Pinheiro, and Ronan Collobert. "Simple Image Description Generator via a Linear Phrase-Based Approach." In: *CoRR* abs/1412.8419 (2014).

[38] Micah Hodosh, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics." In: *Journal of Artificial Intelligence Research* (2013), pp. 853–899.

[39] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. "Grounded compositional semantics for finding and describing images with sentences." In: *Transactions of the ACL* 2 (2014), pp. 207–218.

[40] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3128–3137.

[41] Quanzeng You, Hailin Jin, and Jiebo Luo. "Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions." In: *arXiv preprint arXiv:1801.10121* (2018).

[42] Omid Mohamad Nezami, Mark Dras, Stephen Wan, and Cecile Paris. "Senti-Attend: Image Captioning using Sentiment and Attention." In: *arXiv preprint arXiv:1811.09789* (2018).

[43] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks." In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[44] Xinlei Chen and C. Lawrence Zitnick. "Mind's eye: A recurrent visual representation for image caption generation." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 2422–2431.

[45] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models." In: *CoRR* abs/1610.02424 (2016).

[46] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)." In: *CoRR* abs/1412.6632 (2014).

[47] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." In: *arXiv preprint arXiv:1412.2306* (2014).

[48] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), pp. 677–691.

[49] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. "Guiding the Long-Short Term Memory Model for Image Caption Generation." In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 2407–2415.

[50] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image Captioning with Semantic Attention." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4651–4659.

[51] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 6077–6086.

[52] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[53] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient." In: *AAAI*. 2017.

[54] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. "Policy Gradient Methods for Reinforcement Learning with Function Approximation." In: *NIPS*. 1999.

[55] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. "Towards diverse and natural image descriptions via a conditional gan." In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2970–2979.

[56] Shiyang Yan, Fangyu Wu, Jeremy S. Smith, Wenjin Lu, and Bailing Zhang. "Image Captioning Based on a Hierarchical Attention Mechanism and Policy Gradient Optimization." In: *CoRR* abs/1811.05253 (2018).

[57] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. "Self-Critical Sequence Training for Image Captioning." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1179–1195.

[58] Ronald J. Williams. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning." In: *Machine Learning* 8 (2004), pp. 229–256.

[59] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. "Improved Image Captioning via Policy Gradient optimization of SPIDEr." In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 873–881.

[60] Xiang Lin, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M. Hospedales. "Actor-Critic Sequence Training for Image Captioning." In: *CoRR* abs/1706.09601 (2017).

[61] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. "Deep Reinforcement Learning-Based Image Captioning with Embedding Reward." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1151–1159.

[62] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan Ting Hsu, Jianlong Fu, and Min Sun. "Show, Adapt and Tell: Adversarial Training of Cross-Domain Image Captioner." In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 521–530.

[63] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2641–2649.

[64] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a method for automatic evaluation of machine translation." In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 311–318.

[65] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries." In: *Text Summarization Branches Out* (2004).

[66] Michael Denkowski and Alon Lavie. "Meteor universal: Language specific translation evaluation for any target language." In: *Proceedings of the ninth workshop on statistical machine translation*. 2014, pp. 376–380.

[67] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

[68] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. "Describing objects by their attributes." In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE. 2009, pp. 1778–1785.

[69] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." In: *International Journal of Computer Vision* 123.1 (2017), pp. 32–73.

[70] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. "Large-scale visual sentiment ontology and detectors using adjective noun pairs." In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM. 2013, pp. 223–232.

[71] Anush K Moorthy, Pere Obrador, and Nuria Oliver. "Towards computational models of the visual aesthetic appeal of consumer videos." In: *European Conference on Computer Vision*. Springer. 2010, pp. 1–14.

[72] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. "High level describable attributes for predicting aesthetics and interestingness." In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 1657–1664.

[73] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. "Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content." In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM. 2013, pp. 459–460.

[74] Angeliki Lazaridou, Georgiana Dinu, Adam Liska, and Marco Baroni. "From visual attributes to adjectives through decompositional distributional semantics." In: *TACL* 3 (2015), pp. 183–196.

[75] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. "Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs." In: *Proceedings of the 21st ACM International Conference on Multimedia*. MM '13. Barcelona, Spain: ACM, 2013, pp. 223–232. ISBN: 978-1-4503-2404-5. DOI: 10.1145/2502081.2502282. URL: http://doi.acm.org/10.1145/2502081.2502282.

[76] Brendan Jou and Shih-Fu Chang. "Deep Cross Residual Learning for Multitask Visual Recognition." In: *ACM Multimedia*. 2016.

[77] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. "Visual affect around the world: A large-scale multilingual visual sentiment ontology." In: *Proceedings of the 23rd ACM international conference on Multimedia*. ACM. 2015, pp. 159–168.

[78] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. "Real-time analysis and visualization of the YFCC100M dataset." In: *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*. ACM. 2015, pp. 25–30.

[79] Ranjay Krishna et al. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations." In: 2016. URL: https://arxiv.org/abs/1602.07332.

[80] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. "Describing objects by their attributes." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 1778–1785. DOI: 10.1109/CVPR.2009.5206772.

[81] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. "The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding." In: *International Journal of Computer Vision* 108.1-2 (2014), pp. 59–81.

[82] Flickr Group. *Global Flickr Statistics*. http://statsr.net/flickr-stats/. 2016.

[83] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. "YFCC100M: The new data in multimedia research." In: *Communications of the ACM* 59.2 (2016), pp. 64–73.

[84] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context." In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[85] Alireza Koochali, Sebastian Kalkowski, Andreas Dengel, Damian Borth, and Christian Schulze. "Which Languages do People Speak on Flickr?: A Language and Geo-Location Study of the YFCC100m Dataset." In: *Proceedings of the 2016 ACM Workshop on Multimedia COMMONS*. ACM. 2016, pp. 35–42.

[86] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution." In: *SIGIR 2010 workshop on crowdsourcing for search evaluation*. Vol. 2126. 2010, pp. 22–32.

[87] Tobias Hossfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing." In: *IEEE Transactions on Multimedia* 16.2 (2013), pp. 541–558.

[88] Dom Lachowicz. *Enchant Spellcheck library*. https://abiword.github.io/enchant/. 2016.

[89] H. Paul Grice. "Logic and Conversation." In: *Speech acts*. Ed. by Peter Cole. Vol. 3. Syntax and semantics. New York: Academic Press, 1975, pp. 41–58. ISBN: 0127854231.

[90] Edward Loper and Steven Bird. "NLTK: The Natural Language Toolkit." In: *CoRR* cs.CL/0205028 (2002). URL: http://arxiv.org/abs/cs.CL/0205028.

[91] *Subjectivity Definition*. 'https://www.merriam-webster.com/dictionary/subjective'. 2018.

[92] Marie L Smith, Frédéric Gosselin, and Philippe G Schyns. "Measuring internal representations from behavioral and brain data." In: *Current Biology* 22.3 (2012), pp. 191–196.

[93] Claus-Christian Carbon. "Cognitive mechanisms for explaining dynamics of aesthetic appreciation." In: *i-Perception* 2.7 (2011), pp. 708–719.

[94] Heinrich H Bulthoff. "Bayesian decision theory and psychophysics." In: *Perception as Bayesian inference* 123 (1996).

[95] Guido Hesselmann, Christian A Kell, and Andreas Kleinschmidt. "Ongoing activity fluctuations in hMT+ bias the perception of coherent visual motion." In: *Journal of Neuroscience* 28.53 (2008), pp. 14481–14485.

[96] *SUN Attributes*. http://cs.brown.edu/~gmpatter/sunattributes.html. 2018.

[97] Bing Liu and Lei Zhang. "A survey of opinion mining and sentiment analysis." In: *Mining text data*. Springer, 2012, pp. 415–463.

[98] Marco Baroni and Roberto Zamparelli. "Nouns Are Vectors, Adjectives Are Matrices: Representing Adjective-noun Constructions in Semantic Space." In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10. Cambridge, Massachusetts: Association for Computational Linguistics, 2010, pp. 1183–1193. URL: http://dl.acm.org/citation.cfm?id=1870658.1870773.

[99] Birgit Hamp and Helmut Feldweg. "GermaNet - a Lexical-Semantic Net for German." In: *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. 1997, pp. 9–15.

[100] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning." In: *Proceedings of ACL*. 2018.

[101] *Python-NLTK Documentation*. https://www.nltk.org/book/ch05.html. 2018.

[102] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[103] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge." In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[104] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[105] Matthias "Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp" Cimiano. ""Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases"." In: *"Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers"*. "Valencia, Spain": "Association for Computational Linguistics", "2017", "54–64". URL: "http://aclweb.org/anthology/E17-1006".

[106] Emiliano Guevara. "A Regression Model of Adjective-noun Compositionality in Distributional Semantics." In: *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*. GEMS '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 33–37. ISBN: 978-1-932432-82-4. URL: http://dl.acm.org/citation.cfm?id=1870516.1870521.

[107] David "Bamman, Chris Dyer, and Noah A." Smith. ""Distributed Representations of Geographically Situated Language"." In: *"Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) "*. "Baltimore, Maryland": "Association for Computational Linguistics", "2014", "828–834". DOI: "10.3115/v1/P14-2134". URL: "http://www.aclweb.org/anthology/P14-2134".

[108] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. "Mutan: Multimodal tucker fusion for visual question answering." In: *Proc. IEEE Int. Conf. Comp. Vis.* Vol. 3. 2017.

[109] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. “Composing simple image descriptions using web-scale n-grams.” In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. ACL. 2011, pp. 220–228.

[110] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. “Explain images with multimodal recurrent neural networks.” In: *arXiv preprint arXiv:1410.1090* (2014).

[111] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. “Show and tell: A neural image caption generator.” In: *arXiv preprint arXiv:1411.4555* (2014).

[112] X. Chen, H. Fang, TY Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. “Microsoft COCO Captions: Data Collection and Evaluation Server.” In: *arXiv:1504.00325* (2015).

[113] Adrian Ulges, Damian Borth, and Thomas M Breuel. “Visual concept learning from weakly labeled web videos.” In: *Video Search and Mining*. Springer, 2010, pp. 203–232.

[114] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. “Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs.” In: *ACM Int. Conf. on Multimedia (ACM MM)*. 2013.

[115] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. “Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks.” In: *arXiv preprint arXiv:1410.8586* (2014).

[116] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. “Distributed Representations of Words and Phrases and their Compositionality.” In: *CoRR* abs/1310.4546 (2013). URL: http://arxiv.org/abs/1310.4546.

[117] Mehdi Mirza and Simon Osindero. “Conditional generative adversarial nets.” In: *arXiv preprint arXiv:1411.1784* (2014).

[118] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context.” In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[119] Alexander Patrick Mathews, Lexing Xie, and Xuming He. “Senticap: Generating image descriptions with sentiments.” In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[120] Tushar Karayil, Philipp Blandfort, Jörn Hees, and Andreas Dengel. “The Focus-Aspect-Polarity Model for Predicting Subjective Noun Attributes in Images.” In: *International Coference of Multimedia Retrieval* (2019).

ACADEMIC CURRICULUM VITÆ

TUSHAR KARAYIL

EDUCATION

2016-2020 University of Kaiserslautern, Germany *PhD Researcher*

Topic: ***Affective Image Captioning***

Area: Deep Learning, Image Processing, Natural Language Processing

2013-2015 University of Kaiserslautern, Germany *Master of Science*

Thesis: *A segmentation-free approach for printed Devanagari script recognition*

Area: Deep Learning, Image Processing, OCR

2001-2005 National Institute of Technology, India *Bachelor of Technology*

Thesis: *Methods for securing the Linux Kernel*

Area: Operating Systems

WORK EXPERIENCE

2016–2020 PhD Researcher *TU Kaiserslautern*

Topic mining and Trend Analysis in social media.
Development of Google Cloud based solutions.
Deep Learning for Finance.
Deep Learning Image description API.

2010–2012 Senior Software Engineer *Novell Inc.*

Development of Domain Services for Linux.
General Product Support.

2005–2010 Systems Software Engineer *IBM*

Development of Communication Protocols for Unix and Mainframes.
General Product Support.

SELECTED PUBLICATIONS

Sep 2019 Conditional GANs for Image Captioning with Sentiments.

Authors: Tushar Karayil, Asif Irfan, Federico Raue, Jörn Hees, Andreas Dengel *ICANN*

July 2019 The Focus-Aspect-Value Model for Explainable Prediction of Subjective Visual Interpretation.
ICMR Authors: Tushar Karayil, Philipp Blandfort, Jörn Hees, Andreas Dengel

July 2019 Fusion Strategies for Learning User Embeddings with Neural Networks.
IJCNN Authors: Philipp Blandfort, Tushar Karayil, Jörn Hees, Andreas Dengel

August 2017 Image Captioning in the Wild: How People Caption Images on Flickr.
ACMMM Authors: Philipp Blandfort, Tushar Karayil, Damian Borth, Andreas Dengel

August 2016 Generating affective captions using concept and syntax transition networks.
ACMMM Authors: Tushar Karayil, Philipp Blandfort, Damian Borth, Andreas Dengel

TUTOR

2016-2018 Multimedia Analysis and Data Mining

SKILLS

Programming	Python, C, C++
Machine Learning	Statistics, Image Processing, NLP
Cloud Computing	Google-Cloud, AWS
Databases	MySQL, Google BigQuery, Google NDB, MongoDB