

Copyright
by
Meng Zhang
2019

**The Report Committee for Meng Zhang
Certifies that this is the approved version of the following Report:**

**Image Captioning Algorithms for Images Taken by People with Visual
Impairments**

**APPROVED BY
SUPERVISING COMMITTEE:**

Danna Gurari, Supervisor

Ken Fleischmann

**Image Captioning Algorithms for Images Taken by People with Visual
Impairments**

by

Meng Zhang

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN INFORMATION STUDIES

The University of Texas at Austin

May 2019

Acknowledgements

I want to first and foremost thank Dr. Danna Gurari, supervisor of my project, for kindly and warmly providing critical guidance and assistance in all aspects of this project. I couldn't have finished this project without her. Moreover, her fantastic teaching in the class, Introduction to Machine Learning, sparked my interests in the field of machine learning and computer vision and motivated me to face challenging academic problems.

I also want to thank Dr. Ken Fleischmann for being the reader of my report and providing feedback.

I want to thank Yinan Zhao for helping and guiding me in specific skills related to training and evaluating the deep learning models, as well as fixing thousands of trivial but annoying bugs. Thanks to Nilavra Bhattacharya for the help in visualizing the VizWiz dataset and providing ground truth captions of it.

Last but not least, this project is sponsored by the Microsoft Ability Initiative. Thanks to Microsoft for providing the funds to make this project possible.

Abstract

Image Captioning Algorithms for Images Taken by People with Visual Impairments

Meng Zhang, B.E., M.S.Info.St.

The University of Texas at Austin, 2019

Supervisor: Danna Gurari

People with visual impairments regularly encounter the challenge that their visual impairments expose them to a time-consuming, or even impossible, task: what content is presented in an image without assistance. One method to address this problem is image captioning with machine learning. With the help of image captioning algorithms together with artificial intelligence speech system, people who are blind can instantly learn what is in an image, since such systems can automatically generate text captions. In this work, we analyze the new VizWiz dataset and compare it to the MSCOCO dataset, which is widely used for evaluating the performance of image captioning algorithms. We also implement and evaluate two state-of-the-art image caption models with accuracy, runtime, and resource analysis. Hopefully, our research will help the improvement of image captioning algorithms which focus on fulfilling the everyday needs of people with visual impairments.

Table of Contents

| | |
|---|------|
| Acknowledgments | iv |
| Abstract | v |
| List of Tables | viii |
| List of Figures | ix |
| Chapter 1. Introduction | 1 |
| Chapter 2. Related Works | 3 |
| 2.1 Image Captioning Services | 3 |
| 2.2 Image Captioning Algorithms | 5 |
| 2.3 Image Captioning Datasets | 6 |
| 2.4 Evaluation Metrics for image captioning | 7 |
| Chapter 3. Image Captioning Algorithms | 8 |
| 3.1 Up-Down-Captioner [4] | 9 |
| 3.2 Recurrent-Fusion-Network [19] | 10 |
| Chapter 4. Datasets | 12 |
| 4.1 MS COCO dataset [22] | 12 |
| 4.2 VizWiz dataset [14] | 14 |
| Chapter 5. Evaluation Metrics | 16 |
| 5.1 BLEU [26] | 16 |
| 5.2 ROUGE [21] | 17 |
| 5.3 METEOR [6] | 18 |
| 5.4 CIDEr [33] | 18 |
| 5.5 SPICE [2] | 19 |

| | |
|--|-----------|
| Chapter 6. Experiments and Results | 21 |
| 6.1 Runtime Analysis | 21 |
| 6.1.1 Runtime Analysis for VizWiz Dataset | 22 |
| 6.1.2 Runtime Analysis for MSCOCO Dataset | 23 |
| 6.2 Accuracy Analysis | 25 |
| 6.2.1 Accuracy Analysis of the Up-Down model [4] | 26 |
| 6.2.2 Accuracy Analysis of the Recurrent Fusion model [19] | 28 |
| 6.3 Resource Analysis on the Recurrent Fusion model [19] | 29 |
| Chapter 7. Discussion | 31 |
| 7.1 Defects of Current Image Captioning Research | 31 |
| 7.2 Reflections and Future Works | 33 |
| Chapter 8. Conclusion | 35 |
| Bibliography | 36 |

List of Tables

| | | |
|-----|---|----|
| 6.1 | Evaluation Scores of the Up-Down Model | 29 |
| 6.2 | Evaluation Scores of the Recurrent Fusion Model | 29 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | Facebook’s Image Captioning Service. Users can see the automatically-generated image descriptions(words and phrases) and/or over-write them for their uploaded images for people with visual impairments. | 4 |
| 2.2 | Twitter’s Image Captioning Service. Users can add descriptions for their uploaded images for people with visual impairments. | 5 |
| 4.1 | An example of an image taken from MS COCO [22]. With this image, five human-made sentences are provided that describing the image. | 13 |
| 4.2 | An example of VizWiz [14] images. Together with each image are several human-generated captions and quality problems. We can tell that even for humans it is hard to recognize the content of the image. | 15 |
| 5.1 | This figure is taken from [2]. It is included to illustrate how SPICE works. For a given image, SPICE will generate a scene graph based on the reference sentences as shown in the right side of the figure. | 20 |
| 6.1 | The histogram of the runtime distribution on the VizWiz [14, 13] dataset. | 23 |
| 6.2 | The correlation analysis on the VizWiz [14, 13] dataset. | 24 |
| 6.3 | The histogram of the runtime distribution per pixel on the VizWiz [14, 13] dataset. | 25 |
| 6.4 | The histogram of the runtime distribution on the MSCOCO [22] dataset. | 26 |
| 6.5 | The correlation analysis on the MSCOCO [22] dataset. | 27 |
| 6.6 | The histogram of the runtime distribution per pixel on the MSCOCO [22] dataset. | 28 |

Chapter 1

Introduction

Images are widely used in our daily lives. Compared to text descriptions, images are capable of containing and conveying more complex, detailed information. However, for some people, it is hard or even impossible to understand the content of images. Specifically, people with visual impairments will frequently encounter the challenge that their disability limits their capability of learning what content is present in an image without assistance. This problem will not only cause daily inconvenience but also will sometimes be serious enough to threaten their lives (for example, failure to recognize medicines and the instructions of some dangerous tools).

One way to address this problem is to first transform images to the form of text describing the content of the images. Then with the help of some text-to-speech systems [9, 11, 18], people with visual impairments can hear and learn about the images. The first processing step is known as image captioning or image annotation. Traditionally this work is done manually, which is time-consuming and expensive. Recently, with the advances in machine learning and computer vision, it has become a popular research topic. [4, 19, 14, 22, 20, 30, 24]. According to Wikipedia, automatic image annotation is the process

by which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image. [35]

In recent years, some machine learning challenges and competitions related to image captioning have been created, of which one of the most famous is the MSCOCO Captioning Challenge [23]. However, automatic image captioning with machine learning is not easy. First, a huge-scale image dataset must be collected for training the captioning models. Then state-of-the-art algorithms should be developed so that the captioning system will not only provide high-quality image captions but also run quickly so that users can get an immediate response when they upload an image to the system.

In this work, I explore how machine learning algorithms perform on a new image captioning dataset created using images taken by people with visual impairments. In Section 2, I will discuss some recent studies about image captioning, including image captioning services, algorithms, datasets and evaluation metrics. In Section 3, I will describe two state-of-the-art image captioning algorithms that we will evaluate. In Section 4, I will describe two image captioning datasets which will be used in our experiments. In Section 5, I will describe some standard evaluation metrics for evaluating the performance of image captioning algorithms. In Section 6, I will describe how we perform runtime, resource and accuracy analysis for the two algorithms on the two datasets, and then analyze the results. In Section 7, I will discuss some defections of current image captioning algorithms and datasets, and provide some potential directions of our further works.

Chapter 2

Related Works

2.1 Image Captioning Services

Publicly-available image captioning services support customers to upload their images and receive the text description of the images automatically. Typically, a customer can upload one image each time via the computer or the web link, and the image will be taken as an input to the captioning model used by the service to predict the description to show the result to the customer. Currently, one typical image caption service for developers is Microsoft CaptionBot [1]. Some image captioning services for users are provided by Twitter [2] and Facebook [3]. On twitter, users can add captioning for their uploaded images for people with visual impairments. Figure 2.1 shows the interface for users to add an image description when uploading their images. On Facebook, an API called Automatic Alt-text [39] can automatically generate captions for users when uploading their images, and they can overwrite the captions if feeling unsatisfied with the quality of the automatically generated image descriptions. Figure 2.2 shows the interface for users to create an image description for their

¹The website of Microsoft CaptionBot: <https://www.captionbot.ai/>

²<https://help.twitter.com/en/using-twitter/picture-descriptions>

³<https://www.facebook.com/help/216219865403298?helpref=faq-content>

images.



Figure 2.1: Facebook’s Image Captioning Service. Users can see the automatically-generated image descriptions(words and phrases) and/or over-write them for their uploaded images for people with visual impairments.

These services, mostly stand from the view of users to help people with visual impairments, but not directly from the views of people with visual impairments. In my work, I try to address this problem by providing image captioning services directly aiming to people with visual impairments so that

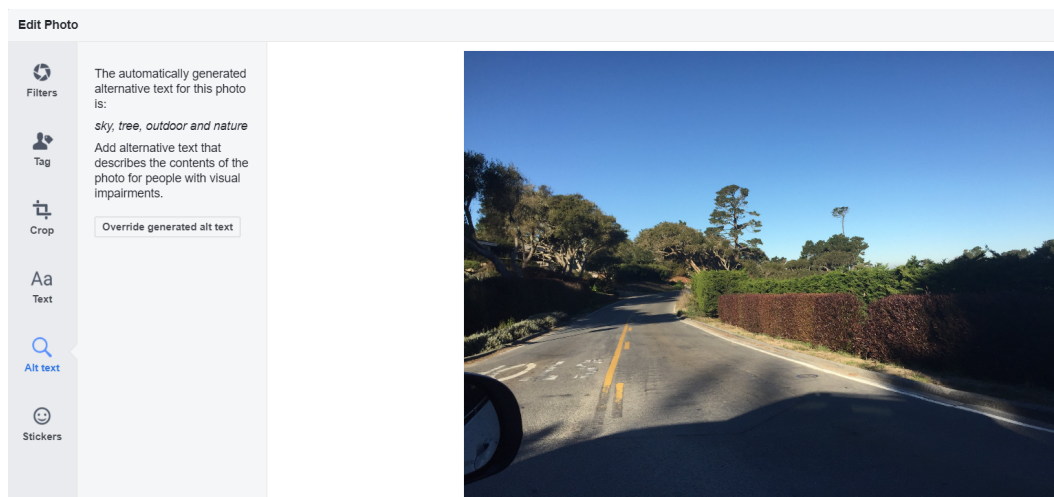


Figure 2.2: Twitter’s Image Captioning Service. Users can add descriptions for their uploaded images for people with visual impairments.

they can get immediate response to the content of the images they want to learn about.

2.2 Image Captioning Algorithms

Image captioning services rely on image captioning algorithms to predict the content of the images. In recent years, most of the state-of-the-art computer vision algorithms [4, 19, 38, 42, 3] use deep learning methods. Images are encoded and extracted into some feature vectors, from which any objects shown in the imaged can be recognized. Then the feature vectors will be decoded to words describing the objects and then arranged in some order as sentences to describe the images. In this work, we will focus on two state-of-the-art algorithms for the MS COCO Challenge [22], the Up-Down Captioner

[4] and the Recurrent Fusion Network [19]. The details will be discussed in Chapter 3. The former researches, however, trained their models with images datasets with high-quality images. In this work, I will study the performance of these two image captioning algorithms for images taken by people with visual impairments, to find out if these algorithms fit well with those special type of images and if the algorithms need specific optimization.

2.3 Image Captioning Datasets

Image captioning datasets are datasets collected for the purpose of training and evaluating algorithms for image captioning. Typically, such datasets will include many images, and for each image, there will be one or more human-made text sentences or phrases to describe the content of the image, i.e. annotations. Taking these annotations as the ground truth, researchers can train their model to predict the automated annotation and evaluate it with the ground truth.

In recent years several image captioning datasets are collected. MS COCO [22] is one of the most famous datasets for image captioning. It contains more than 330,000 images, and for each labeled image, there are five human-made annotations to describe the content of the image. Some other image captioning datasets are Pascal Sentences [27], Flickr8K [27], Flickr30K [44] and Conceptual Captions [31]. In this work, I will focus on a new VizWiz dataset, with new ground truth captions collected. This work is special since the VizWiz is one of the first image datasets that are totally composed of

images taken by people with visual impairments.

2.4 Evaluation Metrics for image captioning

Evaluating the performance of image captioning is a challenge. Unlike image classification, for which we can simply calculate the accuracy, which is the portion where the predicted categories of the images are the same with the true categories, for image captioning, two sentences may both precisely describe one image while being quite different. For example, when describing the scene of a plane flying in the sky, one may use the words “plane”, “fly”, while another will use “jet”, “in flight”.

Currently, the five most-widely-used standard evaluation metrics for image captioning are BLEU [26], ROUGE [21], METEOR [6], CIDEr [33] and SPICE [2], in the order of the published date. All of these five standard evaluation metrics set some formulas to quantify the performance of how similar the predicted caption is to the ground truth description. The details of these evaluation metrics will be demonstrated and discussed in Chapter 5.1 Image Captioning Accuracy Metrics. In this work, I will research on the evaluation scores of the automatically generated captions using the two image captioning algorithms on VizWiz, and find out which metrics fit well with images taken by people with visual impairments, as well as what elements should be added into evaluating the performance of the captions with regard to those specific type of images.

Chapter 3

Image Captioning Algorithms

Image captioning is the task of providing text description that describes an image that can include the objects, people, scenes, activities, etc. Manual image captioning relies on human perception and knowledge. It is time-consuming to hire people for such tasks and costs lots of money. In recent years, machine learning algorithms for image captioning are on the rise. For these algorithms, the input is one image, and the output is a phrase or complete sentence describing the corresponding images.

Recently, most of the state-of-the-art works [4, 19, 15, 8, 5, 40, 29, 41, 25, 24, 43] implement an encode-decode framework to address this problem. The basic method is to use a convolutional neural network (CNN) as the encoder to encode input images so that image feature vectors can be found and extracted. Then a recurrent neural network (RNN) is used as the decoder will take the image feature vectors as the input, create words corresponding to the image features, and then compose the words into a meaningful sentence to describe the image.

We implement two state-of-the-art image captioning algorithms and the details of each work are shown below.

3.1 Up-Down-Captioner [4]

This method combines a bottom-up mechanism based on fast R-CNN [28] in conjunction with ResNet-101 CNN [16] to propose regions from images and extract image feature vectors. A top-down mechanism composed of two LSTM [17] layers calculates the attention of each object recognized, determines feature weightings for them, and then forms a sequence of words. This is why it is called Up-Down-Captioner.

In the bottom-up attention model, Faster R-CNN is used to identify instances of objects by localizing them with bounding boxes. The Faster R-CNN is initialized with ResNet-101 [30] and is pretrained on the Visual Genome dataset [20]. For each given image I , on each spatial location, a Region Proposal Network will predict object box proposals of multiple scales. Then the top box proposals are selected using greedy non-maximum suppression (NMS) with an intersection-over-union (IoU) threshold. NMS is used to make sure there is only one particular object recognized in a region that may contain multiple detected boxes of the same object overlapping with each other. In performing NMS, this helps to avoid getting redundant objects. Then the proposals are passed to region of interest (ROI) pooling to extract feature maps and batched together to output the softmax distribution over the class labels. The final output with non-maximum suppression for each object class uses an IoU threshold and all regions where any class detection probability exceeds a confidence threshold are selected and formed as the image feature vectors.

In the top-down caption model, two LSTM layers are used with the

standard implementation [10]. The first is a top-down attention LSTM and the second is a language LSTM. At each step, the normalized attention weights for each of the k image features is generated. At the Top-Down Attention LSTM layer, it takes the outputs from the bottom-up attention model as the image features. For each time step, the input vector consists of the previous output of the language LSTM, the mean-pooled image features and an encoding of the previously generated word. Given these inputs, the top-down attention LSTM can calculate a weight of each image feature vector. Then in the language LSTM layer, it takes the image features with the outputs of the top-down attention LSTM as the input. At each time step, the conditional distribution over possible output words is calculated with the softmax of the learned weights and biases. The model seeks to minimize the standard cross entropy loss for optimizing the model.

According to the evaluation results, compared to other works [29, 37, 41, 25, 24, 43], this method has competitive performance in terms of identifying objects, object attributes and also the relationships between objects when trained with cross entropy loss without using ensemble methods.

3.2 Recurrent-Fusion-Network [19]

This method also uses a similar encoder-decoder framework [8, 32] where images are encoded by a CNN and then translated into natural language with an RNN.

First, multiple pre-trained CNN models are employed as the encoder to

extract several sets of feature vectors of the input images, respectively. Then a LSTM is deployed as the decoder to transform the representation of the images into a natural language description. This architecture adds a recurrent fusion network (RFNet) right before the decoder LSTM, which consists of multiple components of various encoders and extracts complementary information from them which are formed into one set of thought vectors. A thought vector is a vector containing hundreds of numeric values which represent how each thought relates to other thoughts. [36] It is composed by two stages.

In Fusion Stage I, it takes several sets of annotation vectors as inputs, which are generated using multiple CNN models. For each set, it will calculate a corresponding thought vectors. Then these vectors will be aggregated into one set of thought vectors which contains all of the components and passed into the second stage.

In Fusion Stage II, it will review and compress the aggregated thought vectors to select only one set of thought vectors. In this way, the thought vectors can provide more information than directly input to the feature vectors from the CNN architecture to the LSTM decoder. After the final decoding, the annotation will be generated.

The most novel contribution of this algorithm is that they applied multiple encoders, and proposed a recurrent fusion network (RFNet) including interactions among the outputs of various encoding CNN models and generate new compact and informative representations for the decoder. This work performed state-of-the-art on the MS COCO test server [22].

Chapter 4

Datasets

In this section, we will discuss the details of the image datasets we use in our experiments. We use datasets proposed for general and specific purposes. The MS COCO dataset is used for training and evaluation since this dataset is a standard in the field of image annotation and is widely used in most of the state-of-the-art image captioning algorithms. We will also evaluate our image captioning algorithms with the VizWiz dataset, which is developed by us specifically for captioning images taken by people with visual impairments.

We will describe each dataset below.

4.1 MS COCO dataset [22]

MS COCO is a large-scale image dataset for object recognition and image captioning with more than 200,000 labeled images. Most of the images contain complex scenes with multiple objects. Each image was labeled with five distinct human-made sentences in English that describe the image. The five sentences were provided as the ground truth for image captioning. An example of a MS COCO image is shown in Figure 4.1¹. The 2014 version

¹Source of this example image: <http://cocodataset.org/#explore?id=208408>

of the MS COCO dataset, i.e. 2014 train/validation/test images² is used by most of our baseline algorithms for training and/or evaluation. The training, validation and test set contains 83,000, 41,000 and 41,000 images respectively.

a person walking in the rain on the sidewalk.
a man walks down the strip as it rains.
a person walking through the rain with an umbrella.
a person walking in the rain while holding an umbrella.
a european street scene with a person and vehicles.



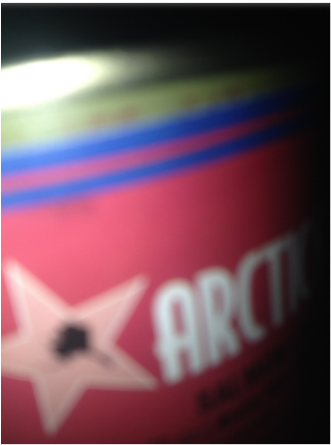
Figure 4.1: An example of an image taken from MS COCO [22]. With this image, five human-made sentences are provided that describing the image.

²<http://cocodataset.org/#download>

4.2 VizWiz dataset [14]

This dataset is proposed for a challenge to develop algorithms to help people with visual impairments overcome their daily visual difficulty like recognizing objects. What makes this dataset unique is that images in this dataset are taken by people who are blind, thus typically the images will be highly blurred as the picture is improperly focused, or just capture an incomplete part of the objects they want to include in the picture since people who are blind cannot see and verify the quality of the photos they make. This is helpful to develop optimized algorithms to specifically recognize photos taken by people who have visual impairments and annotate the content.

Curentntly, VizWiz has two versions, VizWiz v1 [14] and v2 [13]. VizWiz v1 contains 20,000/3173/8000 images for training/validation/test, respectively. VizWiz v2 contains 8088 images. For each image, there are five human-created text descriptions describing the content of the image. Additionally, eaach image is also marked with the quality of the image for some possible problems including blur, light, framing, etc. One example of a VizWiz images is shown in Figure 4.2. As we can see, a typical image taken by people with visual impairments can be blurred, too bright or too dark, incomplete, or in strange view. Unlike other datasets like MS COCO, due to possible low quality of the images, some people may not be able to recognize and describe the image, so not all captions describe objects, scenes, etc.



IMG: VizWiz_test_000000020038. jpg

Crowdsourced captions (Step #1):

- Quality issues are too severe to recognize visual content.**
WorkerID: A12VXVR5FPANU0 English: Native
- Quality issues are too severe to recognize visual content.
WorkerID: A1QT7BQZV0BSY0 English: Native
- A rounded tin with red and blue background colored label with a star as its logo.**
WorkerID: A2UQUDK22T6PED English: Native
- Quality issues are too severe to recognize visual content.
WorkerID: A379EUP7SXCA2N English: Native
- Quality issues are too severe to recognize visual content.
WorkerID: A33BAUWN5T7E0Q English: Native

Image Quality Issues (Step #2):

| | | | | | | | |
|------|------------|----------|-------------|---------|----------|-------|----------|
| 5 | 0 | 2 | 0 | 4 | 0 | 0 | 0 |
| Blur | Too Bright | Too Dark | Obstruction | Framing | Rotation | Other | No issue |

Figure 4.2: An example of VizWiz [14] images. Together with each image are several human-generated captions and quality problems. We can tell that even for humans it is hard to recognize the content of the image.

Chapter 5

Evaluation Metrics

To evaluate the quality of the generated image captions, we use five standard image captioning evaluation metrics. They are BLUE [26], ROUGE [21], METEOR [6], CIDEr [33] and SPICE [2]. We will describe each metric below.

5.1 BLEU [26]

BLEU is one of the most classic image captioning evaluation metrics. BLEU is based on the precision measure, comparing n-grams of the candidate captions with the reference translation and counting the number of matches. The core component of the formula of BLEU is as follows:

$$P_n = \frac{\sum_{C \in \{candidates\}} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C' \in \{candidates\}} \sum_{ngram' \in C'} Count(ngram')} \quad (5.1)$$

where n is the number of n-grams (a word group consisting n words), $Count$ is the number of all n-grams in each caption, and $clip$ is the number of n-grams that have a match between the candidate and reference captions. If using multiple n-grams, then BLEU will calculate the geometric average of all the precision values and return the final score. For example, given a reference sentence “The president speaks to the public” and a candidate sentence “The

president speaks in public”, the BLEU-1 score will be 0.8, since there are four matched unigrams between the reference and candidate sentences out of the five words in the candidate sentence.

This evaluation is only based on precision so lacks consideration of recall. There’s also no consideration of word stemming, word order, or synonyms. Since BLEU uses geometric average, so the final score will be zero if one precision of n-gram is zero, which is another defection of this metric.

5.2 ROUGE [21]

This metric is similar to BLEU, except for it is recall-based. The equation of the core component of ROUGE is shown below:

$$ROUGE = \frac{\sum_{S \in \{RefSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{RefSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (5.2)$$

ROUGE compares n-grams of the candidate captions with the reference translation and count the number of matches, and then calculate the recall score. For the same example as demonstrated in BLEU, the 1-gram ROUGE score will be 0.67, since there are four matched unigrams between the reference and candidate sentences out of six in total the count of words in the reference sentence.

ROUGE shares the same pros and cons as the BLEU does.

5.3 METEOR [6]

METEOR combines unigram precision, unigram recall, as well as a penalty for sentence fragmentation. The formula for METEOR score is shown below:

$$Score = \frac{10PR}{R + 9P} * (1 - penalty) \quad (5.3)$$

where P is the unigram precision calculated like the BLEU score and R is the unigram recall calculated like the ROUGE score. The *Penalty* is a proportion of the number of “chunks” (word segments) where in each chunk all matched unigrams are in adjacent positions, out of the number of matched unigrams. In other words, the more fragments of unigrams that the candidate caption has compared to the reference captions, the higher the penalty will be. The equation is as follows:

$$penalty = 0.5 * (\frac{\#chunks}{\#matched_unigrams})^3 \quad (5.4)$$

Still using the same example, the unigram precision is 0.8, and the unigram recall is 0.67, thus the score without penalty is 0.68. As for the penalty, there are two matched chunks – “The president speak” and “public”, and so four matched unigrams in total. Thus the penalty equals 0.0625, and the final score is 0.6375.

5.4 CIDEr [33]

The unique feature for this evaluation metric is that it introduces a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-

gram. The formula of CIDEr score of n-gram is shown as:

$$CIDEr_n(C_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|} \quad (5.5)$$

Where C is the candidate sentence and S are the reference sentences. $g^n(c_i)$ is a vector corresponding to TF-IDF scores for all n-grams, and $\|g^n(c_i)\|$ is the magnitude of the vector. So as for S_{ij} . Then the CIDEr score will be the uniform weight average of all CIDEr scores of n-grams.

5.5 SPICE [2]

Rather than calculate based on n-gram overlap, SPICE uses semantic propositional content (a scene graph) to assess the quality of image captions. For a set of reference caption sentences, SPICE will first generate a scene graph, which lists all objects recognized, recognized attributes (e.g. colors, shapes) of these objects, and possible relationships between objects like actions and belonging. Then SPICE will calculate the score based on the how well the candidate sentence matches the scene graph. SPICE shows the best correlation with human judgement compared with the previously discussed n-gram based metrics. Figure 5.1 shows an example of how the scene graph is generated for SPICE evaluation.



"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

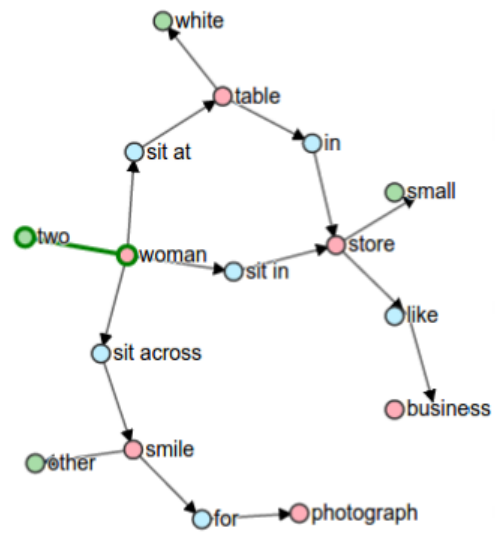


Figure 5.1: This figure is taken from [2]. It is included to illustrate how SPICE works. For a given image, SPICE will generate a scene graph based on the reference sentences as shown in the right side of the figure.

Chapter 6

Experiments and Results

Some experiments are designed to test and evaluate the performance of the two algorithms. Runtime analysis conducted evaluate how fast the algorithms are when generating captions for a given image. Accuracy analysis conducted to evaluate the quality of the automatically generated captions. Finally resource analysis conducted to evaluate how much resources (CPU, disk memory, etc) are needed.

6.1 Runtime Analysis

Runtime analysis is of great interest in evaluating the performance of an algorithm. Runtime analysis provides an overview of the duration the code runs to complete the task. To test the performance of the Up-down model [4], I separately run two runtime analyses where this algorithm is used to predict the caption results for the MSCOCO [22] and VizWiz [14, 13] datasets separately. The detailed results are shown in below subsections.

6.1.1 Runtime Analysis for VizWiz Dataset

We implement a runtime analysis for the Up-down-captioner [4] model. We generate one caption for each of the images from the VizWiz dataset, including both v1 [14] and v2 [13] versions on a machine which contains four GPUs, each having 11GB memory.

According to our results, we test a total of 39,168 VizWiz images. The average size of the images is 1050*1400 pixels. The average runtime for the algorithm to generate an annotation for an image is 0.44 seconds. A histogram of the distribution of the runtime for all images is shown in Figure 6.1. As we can see, the distribution of the runtime roughly follows a normal distribution.

We also analyze the correlation among runtime and the image length, width, as well as the number of boxes recognized in the image. A box is a small region in the image where there contains recognized objects. The number of boxes in each image will be calculated when captioning the image. The result is as in Figure 6.2. According to the chart, the runtime has a strong positive correlation with image length and width, while having a weak correlation with the number of boxes.

Additionally, to test if the image size has a consistent impact on the runtime, we also implement a runtime analysis per pixel. Removing some outliers, the average runtime per pixel is 0.46 microsecond. A histogram of the distribution of the runtime per pixel is shown in Figure 6.3. The variance of runtime per pixel is $4e-07$, which shows the runtime per pixel is quite

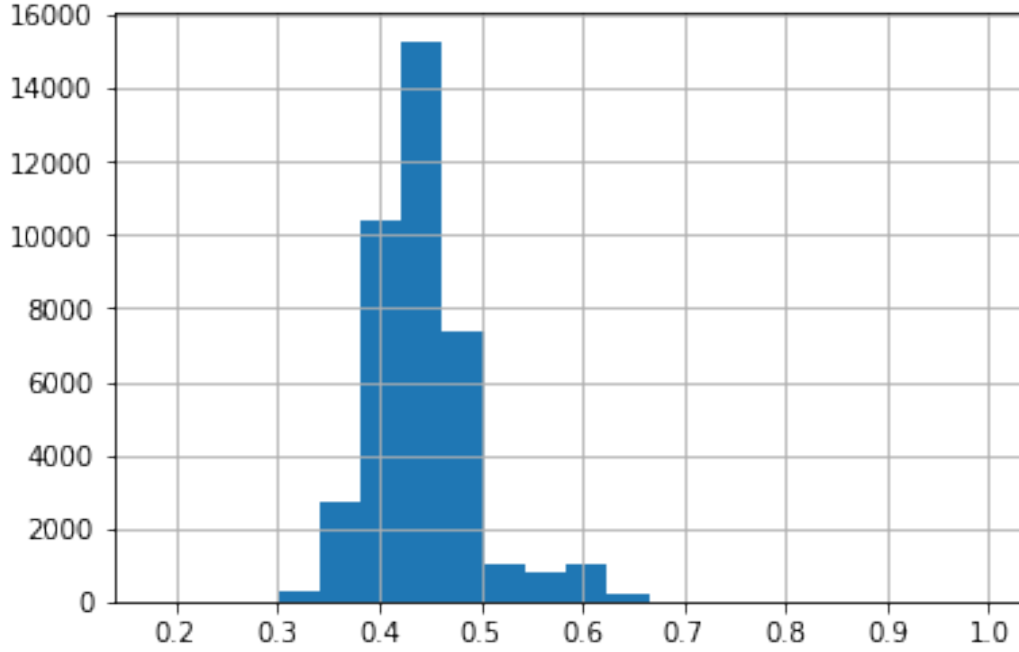


Figure 6.1: The histogram of the runtime distribution on the VizWiz [14, 13] dataset.

consistent.

6.1.2 Runtime Analysis for MSCOCO Dataset

We also implement a runtime analysis with the MSCOCO [22] dataset. The average size of the images is 577*484 pixels. The average runtime is 0.37 second, which is slightly lower than observed for the VizWiz [14, 13] dataset. A histogram of the distribution of the runtime is shown in Figure 6.4. As we can see, the distribution of the runtime roughly follows a normal distribution.

The correlation analysis is shown in Figure 6.5. The result indicates that the runtime has a very weak correlation with either image width, height

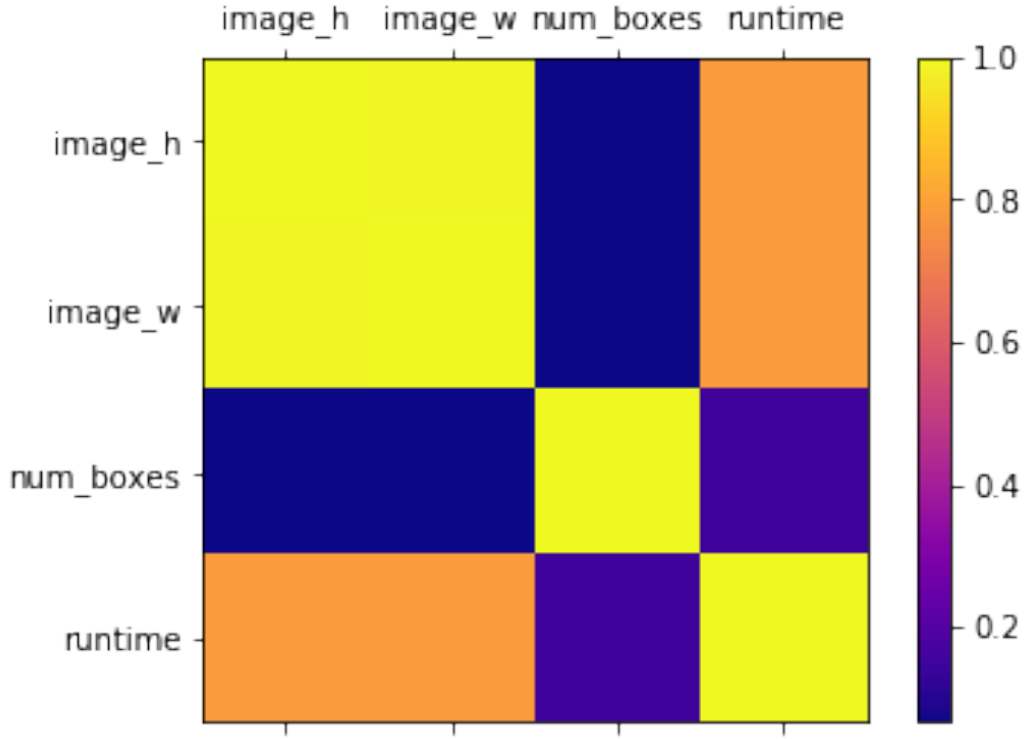


Figure 6.2: The correlation analysis on the VizWiz [14, 13] dataset.

or the number of boxes. A possible reason is in MSCOCO [22], most of the images have similar sizes.

The histogram of the distribution of runtime per pixel is shown in Figure 6.6. Removing the outliers which have strangely higher runtime, the average runtime per pixel is 1.4 microseconds, which is about three times that seen for the VizWiz [14, 13] dataset. A possible reason for that is images in MSCOCO [22] generally has a more complicated scene than VizWiz [?, 14, 13], and so has more objects to be recognized. The variance of runtime per pixel is $1e-07$, which shows the runtime per pixel is quite consistent.

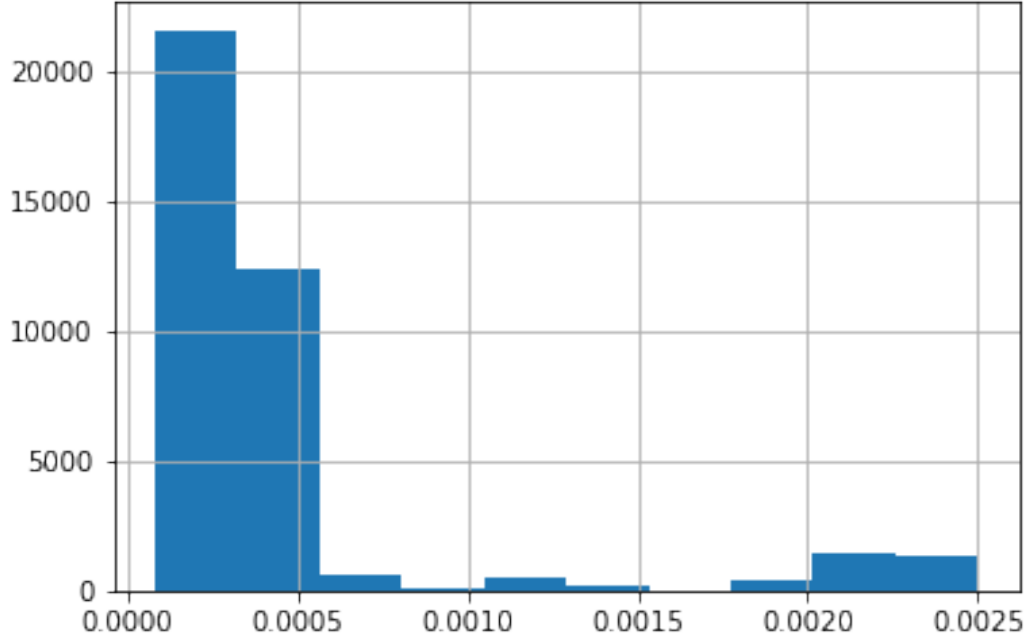


Figure 6.3: The histogram of the runtime distribution per pixel on the VizWiz [14, 13] dataset.

According to the results of the runtime analysis of the Up-down model on both the MSCOCO [22] and VizWiz [14, 13] datasets, this algorithm has consistent performance in runtime. If further optimized, it has the potential of being applied in commercial use when large-scale batch processing of image captioning tasks will be required.

6.2 Accuracy Analysis

The accuracy of image captions is one of the most important metrics when we evaluate the performance of an image captioning model. In this

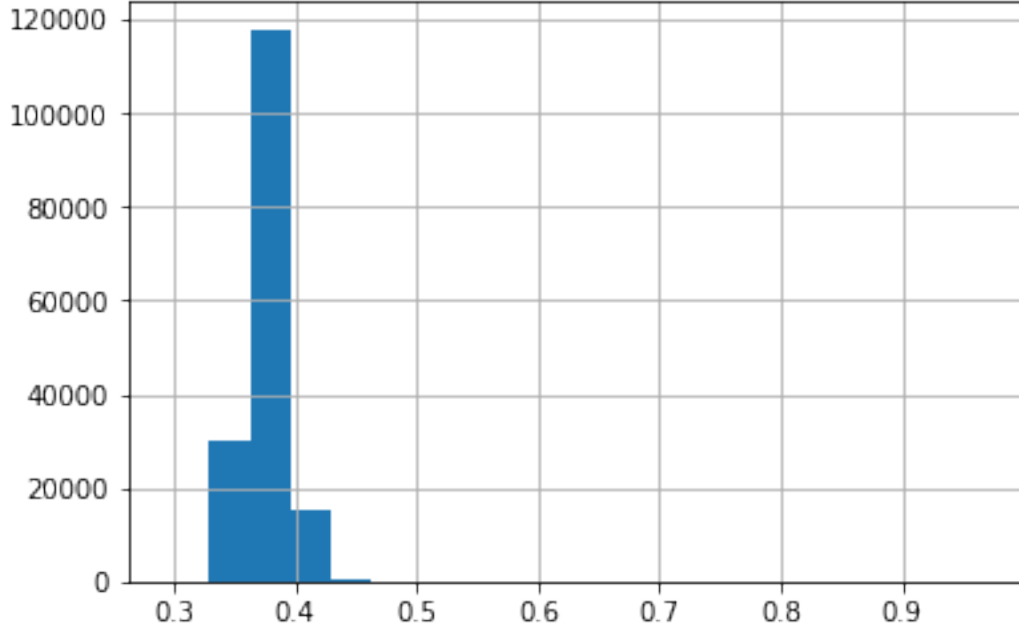


Figure 6.4: The histogram of the runtime distribution on the MSCOCO [22] dataset.

subsection, we evaluate the accuracy of our models using all five standard evaluation metrics discussed above: BLEU [26], ROUGE [21], METEOR [6], CIDEr [33] and SPICE [2]. For each model, we will evaluate those scores on both MSCOCO [22] and VizWiz [14, 13] datasets. The results of the accuracy analysis are shown below in each subsection.

6.2.1 Accuracy Analysis of the Up-Down model [4]

We evaluate the accuracy of this model on the MSCOCO [22] dataset and report the evaluation results together with the scores reported in [43, 29] in Table 5.1. Comparing the scores with those reported in [43, 29], the Up-

| | image_h | image_w | num_boxes | runtime |
|-----------|-----------|-----------|-----------|-----------|
| image_h | 1.000000 | -0.422943 | 0.071638 | -0.017960 |
| image_w | -0.422943 | 1.000000 | 0.062225 | 0.117809 |
| num_boxes | 0.071638 | 0.062225 | 1.000000 | 0.395538 |
| runtime | -0.017960 | 0.117809 | 0.395538 | 1.000000 |

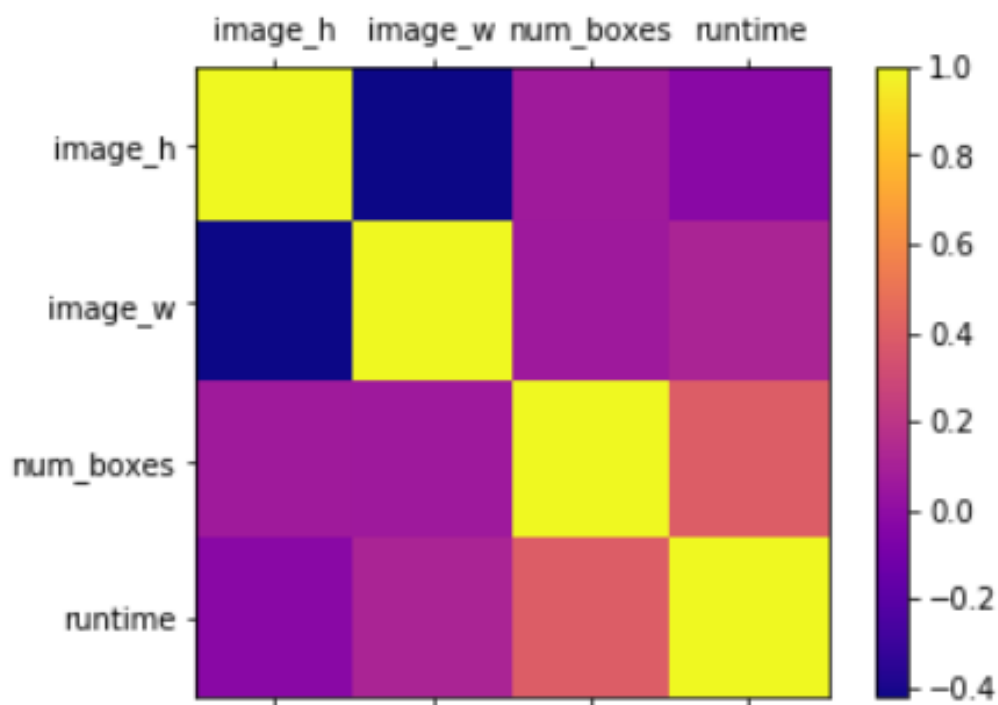


Figure 6.5: The correlation analysis on the MSCOCO [22] dataset.

down model has slightly better scores in terms of nearly all of the evaluation metrics.

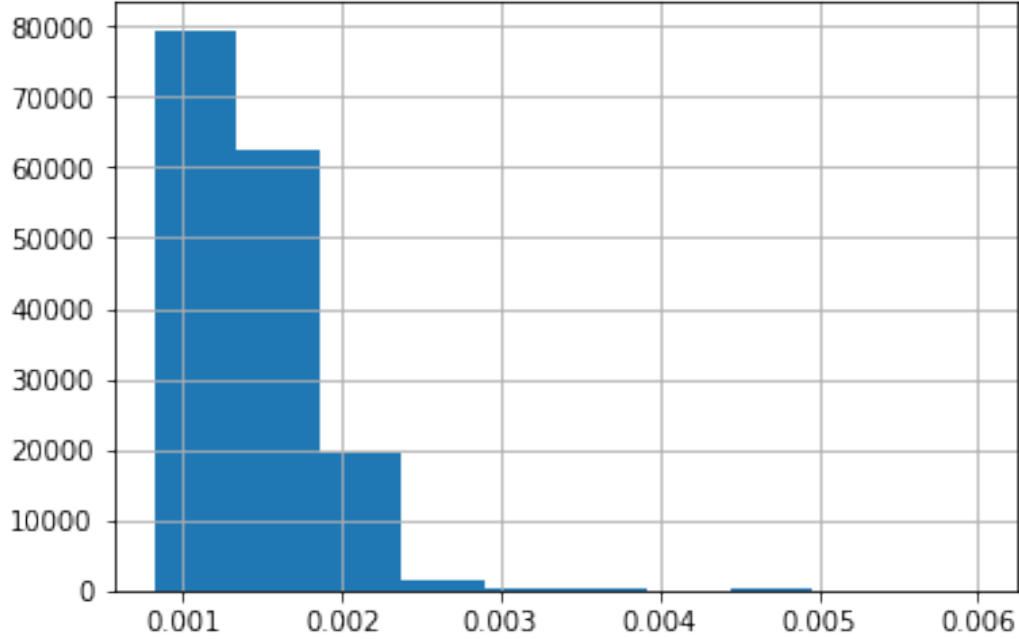


Figure 6.6: The histogram of the runtime distribution per pixel on the MSCOCO [22] dataset.

6.2.2 Accuracy Analysis of the Recurrent Fusion model [19]

We evaluate the accuracy of this model on the MSCOCO [22] dataset as well. The author of this model did not provide a publicly available pre-trained model, so we trained the model with the same dataset split used in [19], and got similar evaluation results as shown in Table 6.2.

Compared with the other methods [29, 43], this method has slightly higher scores. But surprisingly, it has lower scores than the Up-Down model [4], even though it combines and aggregates multiple CNN encoders.

Table 6.1: Evaluation Scores of the Up-Down Model

| Dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | METEOR | CIDEr | SPICE |
|--------------------|--------|--------|--------|--------|-------|--------|-------|-------|
| LSTM-A3* [43] | 0.735 | 0.566 | 0.429 | 0.324 | 0.539 | 0.255 | 0.998 | 0.185 |
| SCST:Att2all* [29] | - | - | - | 0.300 | 0.534 | 0.259 | 0.994 | - |
| UpDown [4] | 0.769 | 0.611 | 0.472 | 0.362 | 0.563 | 0.270 | 1.136 | 0.203 |
| UpDown* [4] | 0.772 | - | - | 0.362 | 0.564 | 0.270 | 1.135 | 0.203 |

Note: Datasets with a “*” indicates that the scores are from the original papers [4, 43, 29], whereas datasets without that sign indicates the scores are from our evaluation.

Table 6.2: Evaluation Scores of the Recurrent Fusion Model

| Dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | METEOR | CIDEr | SPICE |
|--------------------|--------|--------|--------|--------|-------|--------|-------|-------|
| LSTM-A3* [43] | 0.735 | 0.566 | 0.429 | 0.324 | 0.539 | 0.255 | 0.998 | 0.185 |
| SCST:Att2all* [29] | - | - | - | 0.300 | 0.534 | 0.259 | 0.994 | - |
| UpDown* [4] | 0.772 | - | - | 0.362 | 0.564 | 0.270 | 1.135 | 0.203 |
| Recurrent [19] | 0.753 | 0.588 | 0.440 | 0.326 | 0.551 | 0.265 | 1.067 | 0.197 |
| Recurrent [19]* | 0.764 | 0.604 | 0.466 | 0.358 | 0.565 | 0.274 | 1.125 | 0.205 |

Note: Datasets with a “*” indicates that the scores are from the original papers [4, 43, 29, 19], whereas datasets without that sign indicates the scores are from our evaluation.

6.3 Resource Analysis on the Recurrent Fusion model [19]

Similar to the runtime analysis, resource analysis is another kind of assessment of the performance of an algorithm. If running a model will occupy too much memory, it will not be feasible in practice, even if the model will result in accurate predictions.

When implementing [19], we need to first create the flipped version of each image, and then crop both the original images and the flipped images on the top left, top right, bottom left and bottom right corner, respectively, to

enrich the training dataset. As a result, we will basically need ten times the storage space to save these images. Then in the process of feature extraction, we face the same problem, and the experiment shows that running this model needs more than two terabytes of space, which is not really feasible and efficient in practice.

Chapter 7

Discussion

7.1 Defects of Current Image Captioning Research

Although there are a lot of research about image captioning in recent years, there are still significant defects with image captioning services, algorithms, datasets, and evaluation metrics.

As described in previous sections, there are some image captioning services provided in the market at present. These services, however, mainly stand from the view of sighted people helping people with visual impairments, and the image captioning services themselves are not very convenient. For example, on Twitter, users have to manually add descriptions to their images, which is time consuming and trivial. For Facebook, although they provide automatically-generated captions for images, the quality of those captions are poor – they are formed by just a bunch of words rather than meaningful sentences. And only a few major objects can be captured and recognized. Such services are more like object detection from images, and cannot fulfill the need for people with visual impairments to learn about the content of the images, especially when the scenes in the images are complex.

There are many image datasets developed for computer vision and im-

age captioning, but quite few image datasets are developed for the purpose of fulfilling the needs of people with visual impairment. Current image datasets are typically large-scale, but the images in the datasets are generally high quality with rich content details, multiple objects, and excellent framing. This is not the typical pattern for images taken by people with visual impairments. Due to their disabilities, most of the photos taken by people with visual impairments are in low quality. The algorithms trained with improper dataset will result in poor performance for predicting on low quality images. Apart from the images themselves, a good dataset for image captioning aimed to people with visual impairments also involves reviewing the quality of the images, like whether or not it is too dim, bright, blurred, or taken from a bad point of view. This may not directly help to improve the performance of image captioning, but can help to research on the patterns of typical images taken by people with visual impairments.

Currently, most image captioning algorithms take BLEU [26], ROUGE [21], METEOR [6], CIDEr [33] and SPICE [2] as the standard evaluation metrics for evaluating their performance. These metrics are classic, and the evaluation scores of these algorithms can be impressive. However, these scores may not truly represent the capabilities for those image captioning algorithms to describe the content of the images taken by people with visual impairments. That is because the standard evaluation metrics have some vulnerabilities. They ignore special needs from people with visual impairments. For example, people with visual impairments will frequently have the need to recognize text

on images, since they cannot read it.

7.2 Reflections and Future Works

This project is one of the first research which creatively studies image captioning for fulfilling the need of people with visual impairments. There are a lot of research need to be improved in potential future works, though.

We have collected the VizWiz [14, 13] dataset specifically developed using images taken by people with visual impairments. However, the dataset still does not contain enough images for training and evaluation. additionally, there is a problem we need to figure out – there are many low quality images in this dataset so that even humans will find it hard to describe them. Whether these images are valuable for the training of image captioning algorithms, and how to generate the ground truth for these images should be dealt with.

In my experiments, I have done the accuracy analysis of the two algorithms on the MSCOCO [22] dataset. It is important to apply the two algorithms on the VizWiz [14, 13] dataset as well. In the future research, I may first use the models trained by the two algorithms on the MSCOCO [22] dataset to predict the captions for the VizWiz [14, 13] dataset to evaluate the performance of the two algorithms on images taken by people with visual impairments. Then I may train the two algorithms directly on the VizWiz [14, 13] dataset to see if the models will have better performance. A new method of image captioning algorithms may be developed based on our analysis of the results of these experiments.

Moreover, I may come up with a new evaluation metric for assessing the accuracy of the image captions generated. As I have discussed, current evaluation metrics [26, 21, 6, 33, 2] have some limitations in evaluating images with low quality. For people with visual impairments, they want the captions to be as detailed and rich in information as possible, and they may want to know any text shown in the images because they cannot read them. I may propose an evaluation metric which adds more weight on the complexity of the description (i.e. how much detailed information can be provided by the caption), as well as the number of words that are recognizing text in the image. A penalty of short length may be added to decrease the evaluation score.

Finally, if possible, I may expect the research about image captioning for people with visual impairments can be applied with some commercial applications, so that this research can actually bring benefit to people with visual impairments.

Chapter 8

Conclusion

For people with visual impairments, it is hard or even impossible to understand the content of images. Image captioning with deep learning is one of the fastest methods to help people with visual impairments learn about images by automatically generating image captions. In this project, I study two state-of-the-art image captioning algorithms, design experiments to train and evaluate them on MSCOCO [22]. I also study a new dataset, VizWiz, which is developed with images taken by visual impairments and annotated with human descriptions. This project acts as a pilot research aimed to come up with possible thoughts about developing better image captioning algorithms, datasets, evaluation metrics and applications in future works.

Bibliography

- [1] Microsoft CaptionBot. <https://www.captionbot.ai>. Accessed: 2019-03-31.
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.
- [3] Peter Anderson, Stephen Gould, and Mark Johnson. Partially-supervised image captioning. In *Advances in Neural Information Processing Systems*, pages 1879–1890, 2018.
- [4] 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 *CVPR*, 2018.
- [5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

- [7] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7995–8003, 2018.
- [8] 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. *arXiv preprint arXiv:1406.1078*, 2014.
- [9] Min Chu, Hu Peng, and Yong Zhao. Front-end architecture for a multi-lingual text-to-speech system, February 24 2009. US Patent 7,496,498.
- [10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [11] Gordon Freedman. Phonetic speech-to-text-to-speech system and method, October 17 2006. US Patent 7,124,082.
- [12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

- [13] Danna Gurari, Qing Li, Chi Lin, Anhong Guo, Yinan Zhao, Abigale Jane Stangl, and Jeffrey P. Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.
- [14] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.
- [15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [16] 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*, pages 770–778, 2016.
- [17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [18] Xuedong Huang, Alex Acero, Jim Adcock, Hsiao-Wuen Hon, John Goldsmith, Jingsong Liu, and Mike Plumpe. Whistler: A trainable text-to-speech system. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP’96*, volume 4, pages 2387–2390. IEEE, 1996.

- [19] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018.
- [20] 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. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.
- [22] 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*, pages 740–755. Springer, 2014.
- [23] 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*, pages 740–755. Springer, 2014.
- [24] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017.

- [25] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.
- [26] 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*, pages 311–318. Association for Computational Linguistics, 2002.
- [27] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon’s mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.
- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [29] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.
- [30] 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. *International journal of computer vision*, 115(3):211–252, 2015.
- [31] 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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2556–2565, 2018.
- [32] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [33] 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*, pages 4566–4575, 2015.
- [34] Wikipedia contributors. Algorithm — Wikipedia, the free encyclopedia. <https://en.wikipedia.org/w/index.php?title=Algorithm&oldid=891502790>, 2019. [Online; accessed 9-April-2019].
- [35] Wikipedia contributors. Automatic image annotation — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Automatic_image_annotation&oldid=887999594, 2019. [Online; accessed 6-April-2019].

- [36] Wikipedia contributors. Thought vector — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Thought_vector&oldid=877033269, 2019. [Online; accessed 28-April-2019].
- [37] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [38] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381, 2018.
- [39] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1180–1192. ACM, 2017.
- [40] 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*, pages 2048–2057, 2015.
- [41] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Ruslan R Salakhutdinov. Review networks for caption generation. In *Advances in Neural Information Processing Systems*, pages 2361–2369, 2016.

- [42] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018.
- [43] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902, 2017.
- [44] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.