Institute for Advanced Studies
in Basic Sciences
Gava Zang, Zanjan, Iran

Automatic Image Description Generation Using Deep Multimodal Embeddings

Master Thesis in Computer Science
Department of Computer Science and Information Technology
Institute for Advanced Studies in Basic Sciences

Hadi Abdi Khojasteh

Supervisors: Ebrahim Ansari
Parvin Razzaghi

September 2019

Abstract

An essential research topic at the heart of visual recognition is automatic captioning of images. This involves designing an algorithm that takes the image as input and generates a natural language description succinctly describing all or parts of the image. Such a system has wide-ranging applications such as annotating images and exploitation natural descriptions to search for images or texts. This changed significantly with the availability of large-scale annotated data, such as the ImageNet dataset and the application of deep learning techniques, specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This has led to the successful application of such deep networks to various other tasks including the task of image captioning and image-text retrieval.

In this thesis, we proposed an end-to-end deep multimodal convolutional-recurrent network for learning both vision and language representations simultaneously to infer image-text similarity. This model is capable of retrieving an image based on a description (search by image) and retrieve a description based on an image query (image annotation). To learn about the joint representations, leverage our newly extracted collection of tweets from Twitter. The main characteristic of our dataset is that it is unmodified leading the text and image to have semantically higher correlation with each other compared with the benchmark datasets in which the descriptions are well-organized. Once the model is trained, by feeding an image (text) to the model, we found the most similar text (image). The retrieved text (image) is used to explore the other similar ones and have shown that by the proposed model and the new data, the previous aligned models can perform better in terms of the evaluation criteria on MS-COCO dataset. The code and collected dataset have been made available publicly.

Keywords: *Machine Learning, Deep Learning, Convolutional Neural Network, Recurrent Neural Network, Multimodal Model, Image-Text Retrieval, Image Captioning*

Acknowledgments

Throughout the writing of this thesis, I have received a great deal of support and assistance.

Foremost, I would like to express my warmest thanks to my advisor, Dr Ebrahim Ansari, who through hundreds of many meetings over four years, moulded me from an eager but mostly confused student to a early career researcher. Ebrahim's zeal for perfection and passion are infectious. We experienced together all the ups and downs of routine work, the shared happiness of success and also the depression of failure when intrusively everything went wrong. Furthermore, I would like to thank my co-advisor, Dr Parvin Razzaghi for her continuous support and many fruitful discussions. She has helped me greatly during all these years and has always been very kind and nice to me. It would never have been possible for me to take this work to completion without their incredible support and encouragement.

Besides, I would like to thank all present and former colleagues, and members of the IASBS AI Lab, for their assistance, and providing an excellent friendly working atmosphere. My sincere thanks are due to my graduate comrades, Akbar Karimi and Alireza Abbas Alipour, for their invaluable feedback on my research which helped me to accomplish this work. I appreciate their generous help, inquisitiveness and moral support extended to me.

I am also very grateful to all those at the Department of Computer Science, who were always so helpful and provided me with their assistance throughout my study.

I would also like to say a heartfelt thank you to my family for their great support, encouragement and patience through the past seven years.

Finally, thanks to my friends Nasser, Karim, Ebrahim, Roozbeh, Mahsa, Hadi, Mehdi and who helped me in ways unknown to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A picture is worth a thousand words. But how many words can the machine say?

Our impression of an old English proverb

The famous English adage states, A picture is worth a thousand words. Translated into technical terms, this is meant to convey that there is a lot the viewer can learn or infer from a single still image and that enumerating all the information encoded in an image can take up to even a thousand words. This is evident in the widespread use of images in all types of communications, from journals to Telegram chats and Instagram posts. Humans are very good at processing images and videos and gathering all this encoded information, but the computers still struggle to make sense of the simplest ones. One could say it is still easier for computers to see, analyze, search and even understand a thousand words of the big article than a single casual image.

The use of multimedia on the internet has increased dramatically in the recent years, due to easy access to cameras in smart phones. For example, more than 700 million photos shared every day on Telegram [Telegram], about 95 million photos and videos are shared on Instagram every day [Instagram] and about 500 hours of video is shared on YouTube every minute [Youtube]. Rapidly growing amount of visual data being created due to this phenomenon presents both an enormous challenge and an opportunity to build smarter computer algorithms to understand and summarize the data. Hence, an automatic understanding of visual media is an interesting and important problem in many aspects of computer vision (CV) and artificial intelligence (AI).

An essential research topic at the heart of visual recognition is automatic captioning of images. This involves designing an algorithm that takes the image as input and generates

a natural language caption concisely describing the all or part of the image. Impressively solving the above problem requires the machine to be able to identify the salient objects in the image, recognize their attributes, extract the relationships between these objects, understand the event and also to correctly recognize the scene. Since the caption generation requires both image feature extraction and natural language generation modules, this task is considerably more durable, for example, than the well-studied image classification tasks, which have been a main focus in the computer vision field. Indeed, a sentence should capture not only the objects contained in a picture, but it also must express how these objects relate to each other as well as their relationship and the activities that they do.

Until a few years ago, the task of reliably identifying even a single object in an image across diverse and large-scale datasets was hard. This changed significantly with the availability of large-scale annotated data, such as the ImageNet dataset [Deng et al., 2009], and the application of deep learning techniques, specifically convolutional neural networks (CNN). It has been discovered that image classification networks that are trained on the large ImageNet dataset, also generalize very well and can be used as generic image feature extraction for different tasks [Yosinski et al., 2014]. This has led to successful application of such deep networks to various other tasks in computer vision including the task of image captioning.

In this thesis, we proposed a novel method which examines the task of automatic image description generation. We defined the problem more precisely and list out the common building blocks of a visual captioning pipeline.

1.1 Problem Statement

Our task is to generate a caption given an input image. We are going to focus on methods to generate single sentence captions only. Thus a caption can be precisely defined to be a sentence, S, which is a sequence of words $(w_0, w_1, ..., w_{L-1})$ with L being the length of the sentence. We are trying to learn the distribution, $P(S|I)$, where I is the input image. This can be written as

$$P(S|I) = P(w_0, w_1, ..., w_{L-1}|I). \tag{1.1}$$

The probability distribution $P(S|I)$ can be modeled using two staged approach, as has been popular in the image captioning literature. In the first stage the visual input I is mapped onto one or more feature vectors I_f. This process is deterministic and the feature vectors extracted are of fixed size for every input. We can explore different methods for extracting feature vectors

a brown bear standing on a rock in forest area.

Figure 1.1: A sample image-caption pair from the MS-COCO dataset.

from images and analyze their performance quantitatively and qualitatively on the automatic captioning task. For example, for images, features extracted from CNNs trained on ImageNet [Deng et al., 2009] for single image classification, explicit object detector features, and features constructed from object localization networks.

The next stage is the language model, which takes the visual feature vector I_f as input and learns a probability distribution over sentences. Since the sentences are sequences of words, they lend themselves naturally to be modeled using sequential models such as recurrent neural networks (RNN) or its variant called Long-Short Term Memory (LSTM) networks [Hochreiter and Schmidhuber, 1997]. We implemented RNN and make our code for the model publicly available [1].

Evaluating an image captioning system is also arguable, since we have to compare the generated caption against a few different reference captions. We also use multiple automatic evaluation metrics adopted from the field of machine translation research. In addition, we present human evaluation results obtained by automatic captioning challenge.

We analyze the results of the experiments and discuss the strengths and weakness of our solution. After this analysis, discuss a few successful new directions the research on vision and language is heading.

[1] https://github.com/hkhojasteh/Char-RNN-OpenCV

1.2 Structure of the Thesis

The rest of the thesis is organized as follows: We first group automatic image description models into the three categories and provide a comprehensive overview of the models in each category in Chapter 2. In Chapter 4, we introduce our proposed model and newly collected dataset. We then examine the available multimodal image datasets used for training and testing description generation models in Chapter 3. In addition, we review evaluation metrics that have been used to gauge the quality of generated descriptions in Chapter 3.

Chapter 2

Related Work

Given that automatic image description is such an interesting task, and it is driven by the existence of mature natural language processing (NLP) and computer vision (CV) methods and the availability of relevant datasets, a large image description literature has appeared over the last years.

The problem of generating natural language descriptions from visual data has long been studied in computer vision, but mainly for video [Gerber and Nagel, 1996, Yao et al., 2010]. Traditionally, this has led to complex systems composed of structured formal language model combined with a visual primitive recognizer, which are further converted to natural language via rule-based systems. Such systems are heavily hand-designed, relatively fragile and have been demonstrated only on limited domains, e.g. traffic scenes or sport.

We sort the prevailing literature into three categories supported the image description models used. The first group of models follows the classical pipeline: they first predict or detect the image content in terms of objects, attributes, scene types, and actions, based on a set of visual features. Then, these models use this content to drive a linguistic system that outputs a picture description. We will term these approaches *direct generation models*.

The second group presents the problem as a retrieval problem. That is, to create a description for a new image, they search for images in a database that is similar to the new image. Then they create the descriptions for the new image based on a description of a collection of similar images that are retrieved. The new image is described by reusing the similar image retrieved description (transfer), or by combining a new description based on a description of a set of similar images. Retrieval-based models can be divided according to the type of approach they use to represent images and compute similarity. The first subgroup uses a *visual space* to

retrieve images, while the second subgroup of models uses a *multimodal space* that represents text and images jointly. For an overview of the models, and which category they fall into, see Table 3.3. In the following subsections, we describe a comprehensive overview of state-of-the-art approaches to description generation.

2.1 Description Generation from Visual Input

The overall approach of studies in this group is to first predict the most likely meaning of a given image by analyzing its visual content, and then produces a sentence that reflects this. All models in this category achieve this using the following pipeline architecture:

1. Computer vision methods for classifying the scene, to detect the objects present in the picture, to predict their proselytises and the relationships that hold between them, and to recognize the actions taking place.

2. This is followed by a generation step that turns the detector outputs into words or sentences. These are then combined to produce a natural language description of the image, using techniques from sentence generation (e.g., templates, n-grams, grammar rules).

The approaches reviewed in this section perform an explicit mapping from images to descriptions, which differentiates them from the studies described in Section 2.2 and 2.3, which incorporate implicit vision and language models. An illustration of a model is shown in Figure 2.1. An explicit pipeline architecture, while designed to the problem at hand, constrains the generated descriptions, scenes, objects, attributes, and actions.

Approaches to description generation is different in two main dimensions: (a) which image representations they derive descriptions from, and (b) how they address the sentence generation problem.

Recent advances in object recognition and detection moreover as attribute recognition has been wont to drive natural language generation systems, though these are restricted in their expressivity. In terms of the representations used, existing models have conceptualized images in a number of different ways, relying on spatial relationships [Farhadi et al., 2010] by using a detector to infer triplet of scene elements which is converted to text using templates, corpus-based relationships [Yang et al., 2011], or spatial and visual attributes [Kulkarni et al., 2013]. Another group of works using an abstract image representation in the form of meaning tuples which capture different aspects of an image: the objects detected, the attributes of those detections, the spatial relations between them, and the scene type [Farhadi et al., 2010, Kulkarni

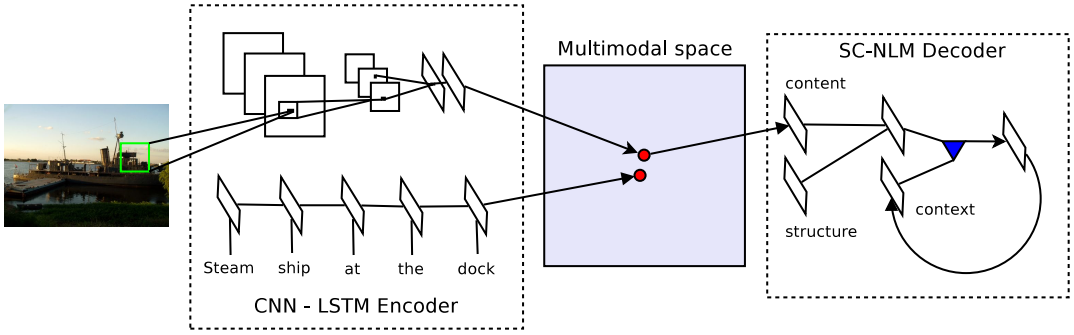


Figure 2.1: The encoder-decoder model for description generation proposed by Kiros et al. [2014].

et al., 2013, Li et al., 2011, Mitchell et al., 2012, Yang et al., 2011]. Some of these approaches have been able to describe images in the wild, but they are heavily hand-designed and rigid when it comes to text generation and cannot use in general purpose. [Yatskar et al., 2014] proposed to generate descriptions of denselylabeled images, which contain objects, attribute, action, and scene annotations. Similar to Fang et al. [2015] work, which does not rely on prior labeling of objects, attributes, etc.

Existing approaches conjointly vary on the second dimension, in however they approach the sentence generation downside. There are approaches that use n-gram-based language models. Examples include the works by Kulkarni et al. [2013] and Li et al. [2011], which both generate descriptions using n-gram language models trained on a subset of Wikipedia. These approaches first determine the attributes and relationships between regions in an image as a regionprepositionregion triples. The n-gram language model is then used to create an image description that is fluent, given the language model. The approach of Fang et al. [2015] is similar, but uses a maximum entropy language model instead of an n-gram model to generate descriptions.

Other approaches have used additional lingually refined approaches to generation. Mitchell et al. [2012] over-generate syntactically well-formed sentence fragments and then recombine these using a tree-substitution grammar. A related approach has been pursued by Kuznetsova et al. [2014], where tree-fragments are learnt from a training set of existing descriptions and then these fragments are combined at test time to form new descriptions. Another linguistically expressive model has been proposed by Ortiz et al. [2015]. The authors model image description as MT over VDRsentence pairs and perform express content selection and surface realization.

Recent image description work using recurrent neural networks (RNNs) it can also be considered as relying on language modeling. A classical RNN is a language model: it obtains the

probability of generating a given word in a string, given the words generated until now. In an image description domain, the RNN is trained to generate the next word given not only the previous characters, but also a set of image features. We will return to this in more detail in Section 2.3.

Retrieving images and rating their descriptions can be done in two ways: either from a visual space or from a multimodal space that combines textual and visual information space. In the following sections, we will survey work that follows these two approaches.

2.2 Description Retrieval in Visual Space

The studies in this group cause the matter of automatically generating the description of an image by retrieving pictures almost like the query image (i.e., the new image to be described); this is illustrated in Figure 2.2. In other words, these systems exploit similarity within the visual space to transfer descriptions to the query pictures. Compared to models that generate descriptions directly (Section 2.1), retrieval models usually need a large quantity of training data so as to produce relevant descriptions.

In terms of their algorithmic elements, visual retrieval approaches usually follow a pipeline of three main steps:

1. Represent the given query image by specific visual feature vector.
2. Retrieve a candidate set of images from the training set supported a similarity measure within the feature space used.
3. Re-rank the descriptions of the candidate images by further creating use of visual and/or textual info contained within the retrieval set, or as an alternative combine fragments of the candidate descriptions according to certain rules or schemes.

One of the first model to follow this approach was the Im2Text model [Ordonez et al., 2011]. GIST [Oliva and Torralba, 2001] and Tiny Image [Torralba et al., 2008] descriptors are used to represent the query image and to determine the visually similar images within the retrieval step. Most of the retrieval-based models contemplate the results of this step as a baseline. For the re-ranking step, a spread of scene classifiers and detectors (e.g., object, stuff, pedestrian, action detectors) specific to the entities mentioned in the candidate descriptions are first applied to the images to better capture their visual content, and the images are represented by means of these detector and classifier responses. Finally, the re-ranking is carried out via a trained classifier over these semantic features.

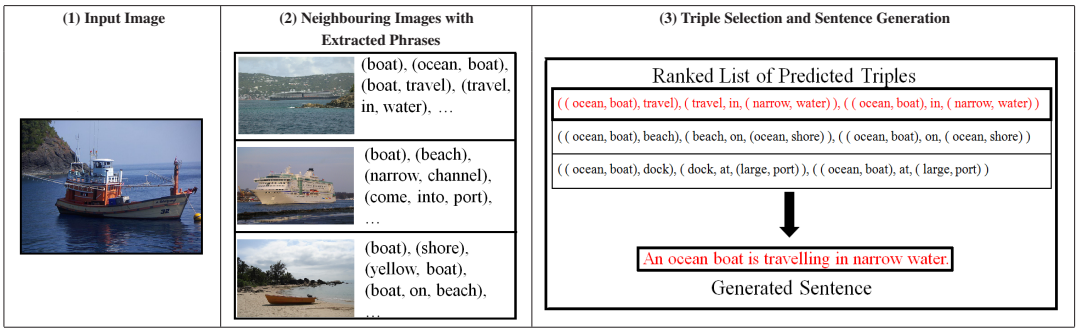


Figure 2.2: The image description generation system based on most similar images and extracted phrases proposed by Kuznetsova et al. [2012].

Some work has addressed the problem of ranking descriptions for a given image [Hodosh et al., 2013, Ordonez et al., 2011]. Such approaches supported the concept of co-embedding of pictures and text within the same vector space. For an image query, descriptions are retrieved which close to the image in the embedding space. Most closely, neural networks are used to co-embed pictures and sentences along [Socher et al., 2014] or even image crops and subsentences [Karpathy et al., 2014] however do not attempt to generate novel descriptions. In general, the on top of approaches cannot describe antecedently unseen compositions of objects, even though the individual objects might have been observed in the training data. Moreover, they avoid addressing the matter of evaluating however sensible a generated description is.

The re-ranking step of Mason and Charniak [2014] examines only textual information and the final output description is determined by using extractive summarization techniques. Later studies make use of CNNs to compute the image features [Devlin et al., 2015]. Phrase-based approaches to synthesise the output were first used by Kuznetsova et al. [2012]. Similar detectors and classifiers used in the reranking step of the IM2TEXT are applied on a query input picture to represent and extract its semantic content. Then a separate image retrieval step for each visual entity in the query image is carried out to collect related phrases from the retrieved sentences. For instance, if a dog is detected in the given image, then the retrieval process returns the sentences referring to visually similar dogs in the training set. Finally, a description is combined from a selection of the retrieved sentences, considering factors such as word order or redundancy.

Yagcioglu et al. [2015] proposed an average query expansion approach which is based on compositional distributed semantics. To represent images, they use features extracted from the recently proposed VGG-CNN neural network [Simonyan and Zisserman, 2014]. These features

are the activations of the last layer of a deep artificial neural network trained on ImageNet [Deng et al., 2009], which have been proven to be effective in many computer vision problems. Then, the original query is expanded as the average of the distributed representations of retrieved descriptions, weighted by their similarity to the input image.

The approach of Devlin et al. [2015] also uses CNN activations as the global image descriptor and carries our k-nearest neighbour retrieval to determine the images that are visually similar to the query image from the training set. Like Mason and Charniak [2014] and Yagcioglu et al. [2015] approaches, it then selects a description from the candidate descriptions associated with the retrieved images that best describes the images that are similar to the query image. Their approach differs in terms of how they represent the similarity between description and how they select the best candidate over the whole set. Specifically, they propose to compute the description similarity based on the n-gram overlap F-score between the descriptions. Vinyals et al. [2015] combine DNN for image classification with recurrent networks for sequence modelling, to create a single network that generates descriptions of images by train networks in the context of this single end-to-end network.

2.3 Description Retrieval in Multimodal Space

The third group of studies casts image description generation once more as a retrieval problem, however from a multimodal space [Hodosh et al., 2013, Karpathy et al., 2014, Socher et al., 2014]. The intuition behind these models is illustrated in Figure 2.3, and therefore the overall approach are often characterised as follows:

1. Learn a common multimodal space for the visual and textual data using a training set of imagesentence pairs.

2. Given a query, use the joint representation space to perform cross-modal (image description) retrieval.

In contrast to the retrieval models that work on a visual space (Section 2.2), wherever unimodal image retrieval is followed by ranking of the retrieved descriptions, here image and description features are projected into a common multimodal space. Then, the multimodal space is is employed to retrieve sentences for a given image. The advantage of this approach is that it allows bi-directional models, i.e., the common space can also be used for the other direction, retrieving the most appropriate image for a query sentence.

Approaches in multimodal retrieval-based systems dissent in the main within the method the common multimodal space is learnt. The seminal paper of Hodosh et al. [2013] makes use of KCCA, a kernelized version of Canonical Correlation Analysis (CCA). A disadvantage of KCCA is that it is only applicable to smaller datasets, as it requires two kernel matrices to be kept in memory during training. This becomes prohibitive for very large datasets. Neural network models are more efficient in constructing a multimodal space, and are now the method of choice. For example, Socher et al. [2014] use Dependency Tree Recursive Neural Network (DT-RNN) for building sentence representations and a nine layer neural network for building image vector representations that are then mapped into a common embedding space.

Karpathy et al. [2014] extended the Sochers multimodal embedding model. Rather than directly mapping entire images and descriptions into a common embedding space, the model embeds more fine-grained units, i.e., fragments of images (objects) and sentences (dependency tree fragments), into a common space. The final model outperforms the DTRNN approach. Other variants of deep neural networks have been used, for example, LongShort Term Memory (LSTM) recurrent neural networks [Kiros et al., 2014], or convolutional network (CNN) to compute the multimodal space.

Models supported retrieval and ranking are limited by the availability of very large datasets with descriptions. A good variety of multimodal models have therefore been developed to not only to rank sentences, but also to generate them, for instance [Karpathy and Fei-Fei, 2015, Kiros et al., 2014, Vinyals et al., 2015].

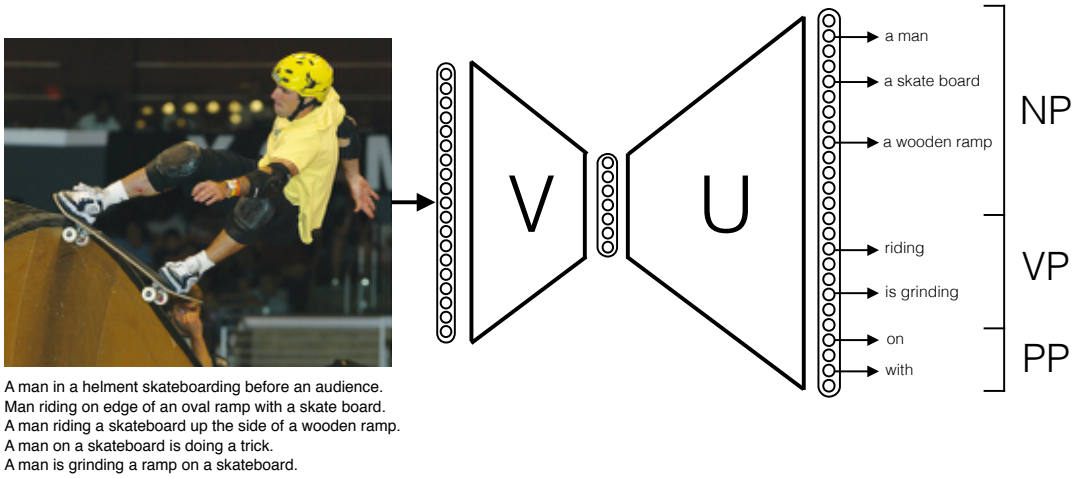


Figure 2.3: Image descriptions as a retrieval task as proposed in Lebret et al. [2015].

Karpathy and Fei-Fei [2015] improve on previous models by proposing a deep visual-semantic alignment model with a simpler network architecture and objective function. Their key insight is to assume that components of the description refer to particular but unknown regions in the image. Their model tries to infer the alignments between segments of sentences and regions of images and is based on CNNs over image regions, bidirectional RNN over sentences and a structured objective that aligns the two modalities. Sentences and image regions are mapped into a common multimodal embedding. The multimodal RNN architecture uses the inferred alignments to learn and generate novel descriptions. Here, the image is used as condition for the first state in the recurrent neural network, which then generates image descriptions.

The closest works are by Kiros et al. [2014] who use a feed-forward neural network, to predict the next word given the image and previous words. This is terribly almost like this proposal, but there are a number of important differences because of using deep neural nets. In addition, some studies try to model in a more explicit fashion the visual anchoring of sentence parts claiming a performance benefit. An explicit word to region alignment is used throughout training by Karpathy and Fei-Fei [2015]. Also Xu et al. [2015] explore attention mechanisms over image regions where while emitting words the system can focus on image parts with hard and soft attention approaches.

The general RNN-based ranking and generation approach is additionally followed by Lebret et al. [2015]. Here, the main innovation is on the linguistic side: they employ a bilinear model to learn a common space of image features and syntactic phrases (noun phrases, verb phrases, and prepositional phrases). A Markov model is then utilized to generate sentences from these phrase embedding. On the visual aspect, usual CNN-based features are used. This leads to an elegant modeling framework, whose performance is broadly comparable to the state-of-the-art.

Description generation systems are difficult to evaluate, therefore the studies reviewed above treat the problem as a retrieval and ranking task [Hodosh et al., 2013, Socher et al., 2014]. While such associate approach has been valuable because it allows comparative evaluation, retrieval and ranking is limited by the availability of existing datasets with sentences. To alleviate this problem, recent models have been developed that are extensions of multimodal spaces; they are able to not only rank sentences, but can also generate them Chen and Zitnick [2015], Donahue et al. [2015], Karpathy and Fei-Fei [2015], Kiros et al. [2014], Lebret et al. [2015], Mao et al. [2014], Vinyals et al. [2015], Xu et al. [2015].

Chapter 3

Datasets and Evaluation

There is a wide range of datasets for investigating image descriptions. Images in this dataset are accompanied by text descriptions and vary in particular aspects such as size, description format, and how to collect descriptions from each other. Here, common approaches to collecting datasets, the datasets themselves and evaluation measures are reviewed to compare generated descriptions with ground-truth texts. The datasets are summarized in Table 3.1, and examples of images and descriptions are given in Figure 3.1. It provides a introductory comparison of some of the existing language and vision datasets. It is not restricted to automatic image description, and it reports some straightforward statistics and quality metrics like perplexity, syntactical quality, and abstract to concrete word ratios.

3.1 Image-Description Datasets

The rapid progress in automatic image captioning in the recent years has also been driven by the availability of large-scale datasets to train and test such models on. These captioning datasets have images with one or more associated reference captions. The reference captions can be collected with large-scale human annotation using crowd sourcing tools such as Amazon Mechanical Turk or they can be mined from other related sources.

The Pascal1K sentence dataset [Rashtchian et al., 2010] is a one of the early datasets for image captioning which is commonly used as a benchmark for evaluating the quality of description generation systems. This medium-scale dataset, consists of 1,000 images five human-annotated captions generated by humans on Amazon Mechanical Turk (AMT) service that were selected from the Pascal 2008 object recognition dataset [Everingham et al., 2010] and includes

	Images	Texts	Judgments	Objects
Pascal1K [Rashtchian et al., 2010]	1,000	5	No	Partial
VLT2K [Elliott and Keller, 2013]	2,424	3	Partial	Partial
Flickr8K [Rashtchian et al., 2010]	8,108	5	Yes	No
Flickr30K [Young et al., 2014]	31,783	5	No	No
Abstract Scenes [Zitnick and Parikh, 2013]	10,000	6	No	Complete
IAPR-TC12 [Grubinger et al., 2006]	20,000	1-5	No	Segmented
MS COCO [Lin et al., 2014]	164,062	5	Collected	Partial
BBC News [Feng and Lapata, 2012]	3,361	1	No	No
SBU1M Captions [Ordonez et al., 2011]	1,000,000	1	Collected	No
Déjà Image-Captions [Chen et al., 2015]	4,000,000	Varies	No	No

Table 3.1: Image datasets for the sentence generation models. We have split the summary into image description datasets (top) and caption datasets (bottom).

objects from different visual classes, such as humans, animals, and vehicles. The Visual and Linguistic Treebank [Elliott and Keller, 2013] makes use of images from the Pascal 2010 action recognition dataset. It augments these images with three, two sentence explanations per image. These sentences were collected on AMT with specific instructions to articulate the main action depicted in the image and the actors involved (first sentence), while also mentioning the most important background objects (second sentence). For a subset of 341 images of the Visual and Linguistic Treebank, object annotation is available (polygons around all objects mentioned in the descriptions).

Flickr8k [Rashtchian et al., 2010] and its extended version Flickr30k [Young et al., 2014] are relatively much larger datasets, comprising approximately 8,000 and 30,000 images, respectively. The images in these two datasets were selected through user queries for specific objects and actions. They also have five human-written captions for each image which were collected from AMT workers similar to the Pascal1K dataset.

The Abstract Scenes dataset [Zitnick and Parikh, 2013] consists of 10,000 clip-art images and their descriptions. The images were created through AMT, where workers were asked to place a fixed vocabulary of 80 clip-art objects into a scene of their choosing. The descriptions were then sourced for these worker-created scenes. The authors presented these descriptions in two different ways. While the first group contains a single description for each image, the second group contains two alternative descriptions in each image. Each of these two descriptions consists of three simple sentences with each sentence that describes the various aspects of the scene. The main advantage of this dataset is that it creates an opportunity to explore image description generation without the need for the need for automatic object recognition, thus

avoiding the associated noise. Another version of this dataset has been created that contains 50,000 different scene images with more realistic human models and with five single sentence description.

The IAPR-TC12 dataset introduced by Grubinger et al. [2006] is one of the primitive multimodal datasets and contains 20,000 images with descriptions. The images were originally retrieved via search engines such as Google, Bing and Yahoo, and the descriptions were produced in multiple languages (predominantly English and German). Each image is accompanied by one to five descriptions, where each description refers to a different aspect of the image, where applicable. The dataset also contains complete pixel-level segmentation of the objects. Currently, the most popular and largest dataset for image captioning is the Microsoft Common Objects in Context (MS-COCO) collection [Lin et al., 2014] with over 200,000 images and at least five human-written captions per image. Images in this dataset are annotated for 80 object categories, Which means that enclosed boxes around all items in one of these categories are available for all images. There exists also an associated MS-COCO evaluation server, where researchers can upload their captions on the blind test dataset and compare the performance of their system to the state-of-the-art methods on a public leaderboard.

NYU dataset [Silberman et al., 2012] contains 1,449 indoor scenes with 3D object segmentation. This dataset has been augmented with five descriptions per image by Lin et al. [2015].

3.2 Image-Caption Datasets

Image descriptions express what can be seen in the image, i.e., they refer to the objects, actions, and attributes depicted, the type of scene, and so on. Captions, on the opposite hand, are generally texts related to pictures that verbalize data that can't be seen within the image. A caption provides personal, cultural, or historical context for the image. Images shared through social networking or photo-sharing websites can be accompanied by descriptions or captions, or a mixtures of both types of text.

The BBC News dataset [Feng and Lapata, 2012] was one of the earliest collections of images and co-occurring texts. Feng and Lapata harvested 3,361 news articles from the British Broadcasting Corporation News website, with the constraint that the article includes an image and a caption.

The SBU1M Captions dataset [Ordonez et al., 2011] differs from the previous datasets in that it is a web-scale dataset containing approximately one million captioned images. It is compiled from information available on Flickr with user-provided image descriptions.

1. a close up of a small cactus in a pot
2. A close-up of three terra cotta pots with cacti growing in them.
3. A line of potted cactus plants.
4. A row of three potted cactus.
5. The trio of cactus take in some sun.

(a) Pascal1K

1. There are several people in chairs and a small child watching one of them play a trumpet
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

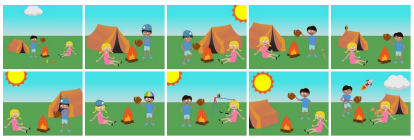
(b) VLT2K

1. A brown dog is shaking off the snow.
2. A brown dog squats in a deep pile of snow.
3. A brown long-haired dog plays in the snow
4. The brown dog is playing in the snow.
5. The brown dog is playing in the white snow.

(c) Flickr8K

1. a yellow building with white columns in the background
2. two palm trees in front of the house
3. cars are parking in front of the house
4. a woman and a child are walking over the square

(d) IAPR-TC12



1. Mike is going to burn Jenny's baseball glove in the campfire which makes Jenny very sad. Jenny starts to cry as Mike holds it over the fire.

(e) Abstract Scenes

1. a man standing in the sun under an umbrella.
2. some people and a woman with a black umbrella and a camera
3. a man holding an umbrella while operating a camera.
4. a man holding an umbrella behind a video camera.
5. person with umbrella being recorded with other people around

(d) IAPR-TC12

Figure 3.1: Example images and descriptions from the benchmark image-text datasets.

The Déjà Image-Captions dataset [Chen et al., 2015] contains 4,000,000 images with 180,000 near-identical captions crawled from Flickr. The image captions are normalized through lemmatization and stop word removal to create a corpus of the near-identical texts.

3.3 Evaluation Metrics

The evaluation of natural language generation (NLG) systems is not trivial, due to the non-unique nature of the solution space. An image can be correctly described with a wide variety of captions differing not only in the syntactic structure, but also in the semantic content. We can see an example of this in Figure 3.2, where a sample image from MS-COCO training set is shown with the corresponding ground truth captions. We see that each caption focuses on different aspects of the image, from the *red jacket* to *buildings in distance*, but all the captions are equally valid.

a man in a red jacket walks down a walking trail next to a park.

a man with a hat on walking near a park and a bench.

a man in red is walking down the long street.

view down a city walkway and street, with grass, pedestrians, trees, cars on street and parked on side of street, a bench, and some buildings in distance.

Figure 3.2: A sample image from the MS-COCO training set with associated ground truth captions. Here we see a clear case where different captions focus at least partially on different aspects of the image.

A good method of evaluation is to compare the machine-generated captions with multiple human-annotated reference captions and use automatic measures. It is worth mentioning the reference captions only represent few samples from the space of all valid captions for the image. Having a large number of reference captions makes it more likely that the solution space is better covered by them and thereby leading to more reliable evaluation.

In other hand, In image retrieval the aim for an model is to correctly indicate which class an image belongs to, or at least make sure that the top images shown to the user are relevant. We can distinguish four cases when an algorithm assigns a description to an image, which are shown in Table 3.2. The most important case is only the *true positive* one, since the user

		Ground truth	
		Class of Interest	Class Not of Interest
Assigned Label	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Table 3.2: Correct and incorrect image-text assignment.

is interested in being shown relevant images or sentences. Furthermore, because the number of data retrieved is limited, the *false positive* case influences the number of correctly labeled relevant images that are presented, since one or more of the shown images may actually be incorrectly labeled as relevant.

Nevertheless, image captioning literature has borrowed three evaluation metrics popular in machine translation, namely BLEU [Papineni et al., 2002], ROUGE-L [Lin and Hovy, 2003] and METEOR [Denkowski and Lavie, 2014]. The models are also able to use measures from information retrieval, such as median rank (mRank), precision at k (S@k), or recall at k (R@k) to evaluate the descriptions they return, additionally to the other text similarity measures. Other metrics popular in image captioning evaluation are the CIDEr [Vedantam et al., 2015] and SPICE [Anderson et al., 2016] metric, specifically for this task. This evaluations reported high correlation with human judgments for imagesentence based ranking evaluations.

3.3.1 Precision

This measure evaluates the performance of an algorithm in returning the relevant images or sentences. If we use the terminology of true/false and positives/negatives, and we assume that the retrieval system only returns us images that it thinks belong to the related descriptions, then we can express precision as follows:

$$precision = \frac{|true\,positives|}{|true\,positives| + |false\,positives|}. \tag{3.1}$$

However, if we assume that the model return us a ranking of images and we only look at a few of them, then the following formula expresses precision:

$$precision = \frac{|true\,positives|}{total\ number\ of\ images\ looked\ at}. \tag{3.2}$$

The number of images looked at thus far is commonly referred to as the *scope*. When precision values are compared at a particular scope value, the performance measure is called the *precision*

rate, and researchers often specify such a value as for instance *p@10*, which in this case means the precision value when the scope equals 10.

3.3.2 Recall

Recall measure the performance of batch information retrieval models. When the ground-truth of the data is available and results obtained by search algorithm are ranked, we can use this measures to weight the result quality in terms of well-ordered true positives in answer. Recall is used to indicate how complete an algorithm is in returning the relevant images, i.e. what percentage of relevant images we have found at this stage:

$$precision = \frac{|true\,positives|}{|true\,positives| + |false\,negatives|}. \tag{3.3}$$

Here it does not particularly matter how many incorrect images are returned, since the recall performance measure only focuses on the number of relevant images that are found thus far.

3.3.3 F Measure

The F_1 measure, or score, is a weighted harmonic mean that combines precision and recall into a single value by weighting them equally:

$$F_1 = 2\,\frac{recall\,.\,precision}{recall\,+\,precision}. \tag{3.4}$$

Differently weighted versions of this measure also exists that give more emphasis on either precision or recall, but these are not as frequently used as the standard F_1 score.

3.3.4 Mean Average Precision

By averaging the precision values obtained every time a relevant image is encountered you get a good sense of how well a method overall performs:

$$AP = \frac{\sum_{i=1}^{N} precision(i)}{N}, \tag{3.5}$$

where $N = |truepositives| + |falsenegatives|$. By calculating the average precision for multiple queries and averaging all these values a single value, the mean average precision (MAP), is obtained.

3.3.5 BLEU

BLEU [Papineni et al., 2002] is a simple metric which scores captions based on the *n-gram* matches between the candidate and the reference captions. First, occurrence counts of different *n-grams* in the candidate sentence are counted and clipped to their maximum value in any single reference sentence, and then accumulated. Next, a modified precision score is computed by dividing this accumulated score by the total number of *n-grams* in the candidate. This process is repeated for different *n-grams*, accommodating modified precision scores p_n. The BLEU score is given by

$$BLEU_n = BP.\, exp(\sum_{n=1}^{N} w_n log(p_n)), \tag{3.6}$$

where BP is the brevity penalty applied in order to penalize short candidate sentences. This additional term has been required, if we only use precision, degenerate candidates such as the ones containing just single words will always score better than longer sentences.

3.3.6 ROUGE-L

ROUGE-L [Lin and Hovy, 2003] metrics were proposed for evaluating text summaries. A metric based on recall and precision scores of the longest common subsequences (LCS) between the reference and candidate sentences:

$$\begin{aligned} R_{lcs} &= \frac{LCS(Cand,\ Ref)}{Reference\ Length}, \\ P_{lcs} &= \frac{LCS(Cand,\ Ref)}{Candidate\ Length}, \end{aligned} \tag{3.7}$$

where R_{lcs} and P_{lcs} are recall and precision metrics, $LCS(Cand, Ref)$ is the longest common subsequence between the candidate Cand and reference Ref.

The metric looks for common sub-sequences by looking for words which appear in the same order in both the reference and candidate captions. Finally, the ROUGE-L metric is computed as the F_β score with $\beta = P_{lcs}/R_{lcs}$:

$$ROUGE - L = F_\beta = \frac{(1+\beta^2)\, R_{ics} R_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}. \tag{3.8}$$

3.3.7 METEOR

In order to compute the METEOR [Denkowski and Lavie, 2014] metric, the candidate and reference sentences are first aligned, wherein each word in the candidate sentence is matched

to at most one word in the reference. When matching words between the candidate and the reference, apart from the exact match, WordNet synonyms, stemmed token matches and paraphrase matches are also considered, in that order. The alignment is done so as to minimize the number of chunks.

Once the two sentences are aligned, weighted precision and recall are computed on the matched words, with different weights being applied to different kinds of matches. The final METEOR score is computed as the product of a penalty term to penalize the number of chunks in the alignment and the $F - score$ based on the weighted precision and recall:

$$\begin{aligned} Pen &= \gamma.(\frac{ch}{m})^{\Theta} \\ METEOR &= (1 - Pen)\,\frac{P_m R_m}{\alpha P_m + (1-\alpha) R_m}, \end{aligned} \tag{3.9}$$

Where R_m and P_m are the recall and precision metrics, ch is the number of chunks the alignment has, m is the length of the candidate and α, γ and θ are hyper-parameters tuned to maximize the correlation of the metric with human judgment. When there are multiple references, the maximum METEOR score between the candidate and any reference is taken.

3.3.8 CIDEr

CIDEr [Vedantam et al., 2015] metric aims to measure how well the candidate caption matches with the consensus formed by the multiple reference captions. For this purpose, each candidate caption and reference sentence are represented using term frequency inverse document frequency vectors (TF-IDF). TF represents the consensus, by considering frequently occurring terms in the reference captions, and IDF helps down-weight common words which occur across captions for many different images.

$CIDEr_n$ metric is computed by averaging the cosine similarity between the TF-IDF vectors of the candidate caption and all reference captions. Here n is the *n-gram* size considering which TF-IDF vector was formed. Final CIDEr metric is the mean of four $CIDEr_n$ metrics, with $n = 1, 2, 3, 4$. Now modified version of this metric which is called *CIDEr-D* widely used in image captioning.

3.3.9 SPICE

The SPICE [Anderson et al., 2016] score ensures our captions are semantically faithful to the image. Rather than directly comparing a generated sentence to a set of reference sentences in

terms of syntactic agreement, SPICE first parses each of the reference sentences, and then uses them to derive an abstract scene graph representation. The generated sentence is then also parsed, and compared to the graph; this allows for a comparison of the semantic similarity, without paying attention to syntactic factors.

The researchers showed [Anderson et al., 2016] that SPICE is the only existing metric that has a strong correlation with human ratings, and ranks human captions above algorithms submitted to the COCO benchmark. The advantage of this metric is that Places importance on capturing details about objects, attributes and relationships. In addition to these this metric does not check whether the grammar is correct and use an equal weighting of different nouns, attributes, relationships.

3.3.10 Conclusions

In summary, all the metrics discussed here evaluate the suitability of a caption to the visual input, by comparing how well the candidate caption matches the reference captions. They perform better with the increasing number of reference captions. The study [Anderson et al., 2016, Vedantam et al., 2015] found that SPICE, CIDEr and METEOR have the highest correlations to human judgment, followed by ROUGE-L and finally BLEU-4.

The models that approach the description generation problem from a cross-modal retrieval perspective [Gong et al., 2014, Hodosh and Hockenmaier, 2013, Hodosh et al., 2013, Karpathy et al., 2014, Socher et al., 2014, Verma and Jawahar, 2014] are also able to use metrics from information retrieval to evaluate the descriptions they return. This evaluation paradigm was first proposed by Hodosh et al. [2013], who reported high correlation with human judgments for imagesentence based ranking evaluations.

In Table 3.3, we summarize all the image description approaches discussed, and list the datasets and evaluation metrics employed by each of these approaches. It can be seen that systems have converged on the use of large description datasets (Flickr8K/30K, MS COCO) and employ evaluation metrics that perform well in terms of correlation with human judgments (Meteor, CIDEr). However, the utilization of BLEU, despite its limitations, remains widespread; additionally the use of human evaluation is by no means that universal within the literature.

Reference	Approach	Datasets	Measures
Farhadi et al. [2010]	MulRetrieval	Pascal1K	BLEU
Kulkarni et al. [2013]	Generation	Pascal1K	Human, BLEU
Li et al. [2011]	Generation	Pascal1K	Human, BLEU
Ordonez et al. [2011]	VisRetrieval	SBU1M	
Yang et al. [2011]	Generation	Flickr8K/30K, IAPR, COCO	BLEU, ROUGE, Meteor, CIDEr, R@k
Gupta et al. [2012]	VisRetrieval	Pascal1K, IAPR	Human, BLEU, ROUGE
Kuznetsova et al. [2012]	VisRetrieval	SBU1M	Human, BLEU
Mitchell et al. [2012]	Generation	Pascal1K	Human
Elliott and Keller [2013]	Generation	VLT2K	Human, BLEU
Hodosh et al. [2013]	MulRetrieval	Pascal1K, Flickr8K	Human, BLEU, ROUGE, mRank, R@k
Gong et al. [2014]	MulRetrieval	SBU1M, Flickr30K	R@k
Karpathy et al. [2014]	MulRetrieval	Flickr8K/30K, COCO	BLEU, Meteor, CIDEr
Kuznetsova et al. [2014]	Generation	SBU1M	Human, BLEU, Meteor
Mason et al. [2014]	VisRetrieval	SBU1M	Human, BLEU
Patterson et al. [2014]	VisRetrieval	SBU1M	BLEU
Socher et al. [2014]	MulRetrieval	Pascal1K	mRank, R@k
Verma and Jawahar [2014]	MulRetrieval	IAPR, SBU1M, Pascal1K	BLEU, ROUGE, P@k
Yatskar et al. [2014]	Generation	Own data	Human, BLEU
Chen and Zitnick [2015]	MulRetrieval	Flickr8K/30K, COCO	BLEU, Meteor, CIDEr, mRank, R@k
Donahue et al. [2015]	MulRetrieval	Flickr30K, COCO	Human, BLEU, mRank, R@k
Devlin et al. [2015]	VisRetrieval	COCO	BLEU, Meteor
Elliott and de Vries [2015]	Generation	VLT2K, Pascal1K	BLEU, Meteor
Fang et al. [2015]	Generation	COCO	Human, ROUGE, BLEU, Meteor, CIDEr
Jia et al. [2015]	Generation	Flickr8K/30K, COCO	BLEU, Meteor
Karpathy et al. [2015]	MulRetrieval	Flickr8K/30K, COCO	BLEU, Meteor, CIDEr, mRank, R@k
Kiros et al. [2014]	MulRetrieval	Flickr8K/30K	R@k
Lebret et al. [2015]	MulRetrieval	Flickr30K, COCO	BLEU, R@k
Lin et al. [2015]	Generation	NYU	ROUGE
Mao et al. [2014]	MulRetrieval	IAPR, COCO, Flickr30K	BLEU, mRank, R@k
Ortiz et al. [2015]	Generation	Abstract Scenes	Human, BLEU, Meteor
Lebret et al. [2014]	MulRetrieval	COCO	BLEU
Ushiku et al. [2015]	Generation	Pascal1K, IAPR, SBU1M, COCO	BLEU
Vinyals et al. [2015]	MulRetrieval	Flickr8K/30K, Pascal1K, SBU1M	BLEU, Meteor, R@k CIDEr, mRank
Xu et al. [2015]	MulRetrieval	Flickr8K/30K, COCO	BLEU, Meteor
Yagcioglu et al. [2015]	VisRetrieval	Flickr8K/30K, COCO	Human, BLEU, Meteor, CIDEr

Table 3.3: An overview of the approaches, datasets, and evaluation measures organised in chronological order. We have categorized the literature into approaches that directly generate a description of an image (Generation), approaches that retrieve images via visual similarity and transfer their description to the new image (VisRetrieval), and approaches that frame the task as retrieving descriptions and images from a multimodal space (MulRetrieval) (more details in Chapter 2).

Chapter 4

Proposed Model

This chapter considers the task of matching images and sentences by learning a visual-textual embedding space for cross-modal retrieval. Finding such a space is a challenging task since the features and representations of text and image are not comparable. In this work, we introduce an end-to-end deep multimodal convolutional-recurrent network for learning both vision and language representations simultaneously to infer image-text similarity. The model learns which pairs are a match (positive) and which ones are a mismatch (negative) using a hinge-based triplet ranking. To learn about the joint representations, we leverage our newly extracted collection of tweets from Twitter. The main characteristic of our dataset is that the images and tweets are not standardized the same as the benchmarks. Furthermore, there can be a higher semantic correlation between the pictures and tweets contrary to benchmarks in which the descriptions are well-organized. Experimental results on MS-COCO benchmark dataset show that our model outperforms certain methods presented previously and has competitive performance compared to the state-of-the-art. The code and dataset has been made available publicly.

4.1 Preface

The baseline image captioning model consists of two parts: a language model and the image feature extraction stage. Image model consists of various techniques to extract descriptors of the visual contents of the input image and represent that as one or more vectors with fixed length. The language model then uses these feature vectors and generates a suitable caption to describe the image regions. Both of these steps are performed simultaneously.

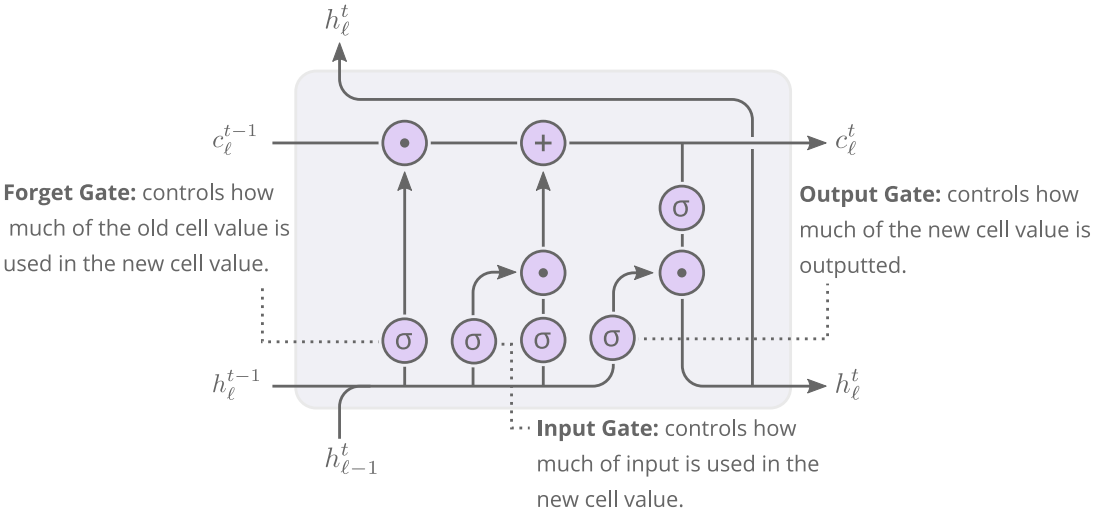


Figure 4.1: A single LSTM cell that allows for long-term memorization by gateing its update, thereby solving the vanishing gradient problem.

4.1.1 Text representation

We leverage conditional language model which takes as input the visual features and input text. The Long-Short Term Memory (LSTM) network [Hochreiter and Schmidhuber, 1997] architecture has been a popular choice to model the probability of a sentence S, given an image feature I, as $P(S|I)$.

The LSTM (or RNN) model has been chosen as the language model based on two basic requirements the image captioning problem imposes. Firstly, the language model needs to handle sentences of arbitrary length and LSTMs are able to do this by design. Secondly, during the training with gradient descent methods, the error signal and its gradients need to propagate a long way back in time without exploding, and LSTMs use shared parameters for this criterion.

The block diagram of a single LSTM cell is shown in Figure 4.1. LSTM cell consists of a memory cell h, whose value at any time step t is influenced by the current vectorial input $x(t)$, the previous output $y(t-1)$ and the previous cell state $h(t-1)$. The update to the memory cell value h is controlled using the input gate o. The gates are implemented with non-linearities $\sigma(.)$ to keep them completely differentiable.

The input and forget gates of the LSTM cells have the ability to keep the content of the memory cell over long periods, which makes it easier to learn long sequences. This process is

formalized in the equation system:

$$\begin{aligned} i(t) &= \sigma(W_{ix}x(t-1) + W_{iy}y(t-1)) \\ o(t) &= \sigma(W_{ox}x(t-1) + W_{oy}y(t-1)) \\ f(t) &= \sigma(W_{fx}x(t-1) + W_{fy}y(t-1)) \\ y(t) &= o(t).h(t) \\ h(t) &= f(t).h(t-1) + i(t).tanh(W_{hx}x(t) + W_{hy}y(t-1)), \end{aligned} \tag{4.1}$$

where W_{XX} are the network weights learned during the training phase.

The baseline language model consists of an LSTM cell with a softmax layer at its output. The softmax outputs the probability distribution over the model's vocabulary as:

$$P(W_t|W_{t-1}, ..., W_0, V) = softmax(D_y(t)), \tag{4.2}$$

where D is the decoder matrix which maps the vector $y(t)$, with the same dimensions as the number of LSTM units, to the output vocabulary size.

The visual features V are fed into the LSTM through an embedding matrix W_{ix} at the zeroth time step as the input $x(0)$. We refer to this feature input as the *init* feature since it initializes the hidden state of the LSTM. In the subsequent time steps t, a start symbol followed by the word embeddings for each word in the reference description (during training) or the previously generated word (during testing) are fed through the same input line, as $x(t)$.

4.1.2 Image representation

Finding good image feature vector representations for the input pairs is a very important task for the successful design of a captioning model. Such a feature representation should be compressed, but also able to encode all the information relevant for the task. For the automatic image description task, the feature vector should capture all the objects in the image, their most essential properties such as size, color, their absolute position and relative location to each other, along with the type of the scene these objects are located in.

Activation values extracted from the deep Convolutional Neural Network (CNN) layers are the primary features used to represent images in the model in this thesis. The CNN features are able to encode a rich variety of information, including object type, scene context, etc., a seen from its performance in the model. However, this representation is still very dense and probably inefficient for the language model to be able to extract the information it needs to

assign correct descriptions.

CNNs have in recent years become practically ubiquitous in most image understanding tasks for all tasks related to image classification and understanding. It is shown in [Donahue et al., 2014] and [Sharif Razavian et al., 2014] that activations of the fully-connected layers of a CNN trained for image classification task act as a general feature representation of the image and can be successfully used to solve other tasks as well. In line with this, image features are extracted here from different CNN architectures pre-trained on two large datasets namely, ImageNet [Deng et al., 2009] and MIT Places [Zhou et al., 2014], originally aimed for object and scene classification, respectively.

Although CNNs have been used as early as the nineties to solve character recognition tasks [Le Cun et al., 1997], their current widespread application is due to much more recent work, once a deep CNNs was used to beat state-of-the-art in the ImageNet image classification challenge [Krizhevsky et al., 2012]. The CNNs used are based on the widely used ResNet [He et al., 2016], GoogLeNet [Szegedy et al., 2015], and VGG [Simonyan and Zisserman, 2014] architectures. All of these architectures achieved good results in the object classification challenges.

4.2 Introduction

The advent of social networks has brought about a plethora of opportunities for everyone to share information online in the forms of text, image, video and so forth. As a result, there is a vast amount of raw data on the Net which could be helpful in dealing with many challenges in natural language processing and image recognition. Matching pictures with their textual descriptions is one of these challenges in which the research interest has been growing [Eisenschtat and Wolf, 2017, Faghri et al., 2017, Lee et al., 2018, Wang and Chan, 2018].

The goal in image-text matching is, given an image, to automatically retrieve a natural language description of this image. In addition, given a caption (textual image description), we want to match it with the most related image found in our dataset as shown in Figure 4.2. The process involves modeling the relationship between images and texts or captions used to describe them. This defines the semantics of a language by grounding it to the visual world.

Many studies have explored the task of cross-modal retrieval on the level of sentence and image regions [Karpathy and Fei-Fei, 2015, Liu et al., 2017, Niu et al., 2017, Wang and Chan, 2018]. Karpathy et al. [2014] work on matching parts of an image objects with phrases by using dependency tree relations for sentence fragments and finding a common space for representing

Figure 4.2: Motivation/Concept Figure: Given an image (caption), the goal in image-text matching is to automatically retrieve the closest textual description (image) for that. Tweets are examples of collected dataset.

fragments. Huang et al. [2017] propose a sm-LSTM where they utilize a multimodal context-modulated global attention scheme and LSTM to predict the salient instance pairs. Recently, many researchers [Donahue et al., 2015, Gu et al., 2018, Huang et al., 2018, Lev et al., 2016, Mao et al., 2014, Yan and Mikolajczyk, 2015, Zheng et al., 2017] introduced a neural network model for image caption retrieval consists of RNNs, CNNs, and additional multimodal layers. Practically, one of the reasons that these deep learning approaches have been on the rise is the availability of abundant information on the Web. The next section describes the proposed model.

4.3 Related work

Many studies have explored the task of cross-modal retrieval on the level of image and sentence fragments. Karpathy et al. [2014] work on matching parts of an image (objects) with sentence fragments (words and phrases). This is done by using dependency tree relations for sentence fragments and finding a common space for representing fragments of the two modalities. Addressing the same task, Niu et al. [2017], propose Hierarchical Multimodal LSTM (HM-LSTM) where the intermediate nodes represent phrases as well as regions of an image and the root is the whole sentence or image. This is against viewing sentences as a chain which makes it difficult to work on the phrase and region level. Significant improvements are reported on MS-COCO

dataset by applying HM-LSTM.

Mao et al. [2014] introduce a multimodal Recurrent Neural Network model for captioning images as well as image caption retrieval. Their network architecture consists of an RNN, a CNN, and one multimodal layer where these two interact. After feature embeddings for words are learned by the RNN and the image representations are produced by the CNN, the multimodal layer connects the two by a one-layer representation. For the same task, Huang et al. [2017] propose a selective multimodal LSTM (sm-LSTM) where they utilize a multimodal context-modulated attention scheme to predict the salient instance pairs from image and sentence. Then, using another multimodal LSTM for measuring the local similarity of the predicted parts, they aggregate the local scores to calculate the global similarity and retrieve the image or text. By considering only the salient parts, the less important parts are ignored which is why Lee et al. [2018] propose stacked crossed attention for image-text matching. Given an image and a sentence, they decided on the importance of an image (text) part by comparing it with each text (image) part.

Karpathy and Fei-Fei [2015] build a neural network to learn a common space embedding for image regions and segments of sentences. Using the learned hidden space, they then introduce a Multimodal Recurrent Neural Network for generating textual descriptions for images. This results in a new dataset which is a collection of descriptions for image regions.

Addressing the task of image and text matching, Ma et al. [2015] propose multimodal Convolutional Neural Networks called m-CNNs. In the proposed architecture, two CNNs have been employed, one for producing image embedding and the other for learning the common representation between image and sentence. By attempting to exploit the matching relations between the two media, they explore their relations on the word, phrase, and sentence level. This allows them to find the equivalent objects to words, regions of the pictures to phrases, and pictures to sentences. The proposed method improves the results on MS-COCO datasets.

Inspired by this work, Wang and Chan [2018] employ two CNNs with a hierarchical attention module which learns the relationships between image parts and the concepts of each level. Because of using dot product in the attention module, their model runs faster compared with the models proposed by Lu et al. [2017] and Xu et al. [2015].

One problem of manually detecting visual concepts, according to [Sun et al., 2015], is that it is not suitable for learning the complexity of the real word visual data due to its lack of adaptability. Therefore, they work on automatically discovering the concepts from images using image descriptions. Testing the idea on the three previously mentioned datasets, they show that there is a significant improvement in the quality of their bidirectional image sentence

retrieval system.

Vendrov et al. [2015] view the relation between image and language as a partial order relation and put forward order-embeddings to address the shortcoming of word embeddings which preserve the distance but not order. The proposed method is generalized in order to be utilized in image-caption retrieval, hypernymy prediction, and natural language inference. They show that order-embeddings work better on the first two than the ladder and result in significant improvements over the existing methods such FV [Klein et al., 2015] and m-CNN [Mao et al., 2014] in the task of image-caption retrieval.

Zheng et al. [2017] introduce a dual-path convolutional neural network in order to find a common image-text space to be used for matching images and pictures. In their model, two CNNs, which have been fine-tuned to enhance the results, are employed to find embeddings for sentences as well as images.

Lee et al. [2018] propose a stacked cross attention model for matching images and sentences. For sentence retrieval, sentence words are attended with respect to image regions to find the similarity between the attended sentence vector and image region vector. The same applies to image retrieval. For each sentence word, the regions in images are attended and their similarity is computed. In the end, a hinge-based triplet ranking loss is employed to find the matches. Their model significantly improves the latest results.

Gu et al. [2018] utilize generative processes in cross-modal feature learning. Their model training is carried out in three phases. First, image and sentences are encoded (i.e. transformed into a common embedding space) using a CNN and two RNNs, respectively. In the second and third phases, the image embedding is utilized to produce a sentence vector and the sentence embedding is employed to produce an image feature vector. Then, the original and the new image vectors are compared using a discriminator.

Huang et al. [2018] propose a semantic-enhanced image-text matching model which learns semantic concepts and their orders. To predict the semantic concepts, they employ a multi-regional multi-label CNN on a given image and to find the order of the concepts, they develop a context-gated sentence generation scheme.

4.4 Model

We introduced an end-to-end multimodal neural network for learning image and text representations simultaneously. The architecture is illustrated in Figure 4.3. It consists of two main subnets, a CNN for input image representation and an LSTM with an embedding to map the

captions into the new space. The purpose of the model is to find a mapping from the text and image to a common space in order to represent them with similar embeddings. In this space, an image (text) will have a similar representation to its text (image) but a different one from other texts (images). Once the model is trained, by feeding an image (text) to the network, we find the most similar text (image). The retrieved text (image) is used to explore the other similar ones.

4.4.1 Image representation

For our initial model, after removing the fully-connected layer from ResNet-50 [Xie et al., 2017] which has been pre-trained on ImageNet [Russakovsky et al., 2015], we treat the remaining layers as a feature extractor. Then, fine-tuning is carried out. The inputs of the network are 224×224 images and the output is a 2048×7×7 feature vector. In order to compare the representations for image and text, they need to be of the same dimensions. Therefore, a dense layer with the size of text domain is added to the end of the network. As well as the rest of the network, this layer, now part of our model, is trained to produce image representations. If we call this vector I, which is a representation of the input image, then f_{img} is a visual descriptor that is the result of forward pass in the network. The forward pass is denoted by $F_{img}(.)$, which is a non-linear function and is defined as $f_{img} = F_{img}(I)$.

Conventionally, the image model is often considered as one part of the neural network and with ImageNet pre-trained weights, it is fine-tuned along with the other parts of the network. However, due to the large number of learnable parameters, it is highly time-consuming which is the reason why precomputed features without training are often used.

Then, we add two 1024 fully-connected layers to transform $f_{img}(.)$ to an image feature vector (v_{img}) computed by $v_{img} = W_{img} f_{img}(.) + b_{img}$.

4.4.2 Text representation

Each input text (T) is first represented by an $n \times d$ matrix, with n being its length and d being the size of the dictionary. To build the dictionary, stop words and punctuation marks are removed and all the words are stemmed using Porter stemmer. In addition, the removal of the special characters is carried out and the remaining words are all in lowercase format. Each word in the final dictionary is represented by a one-hot d dimensional vector and every word can find an index l in the dictionary. Therefore, for an input sentence T with m words, there is a $d \times m$ matrix as the following:

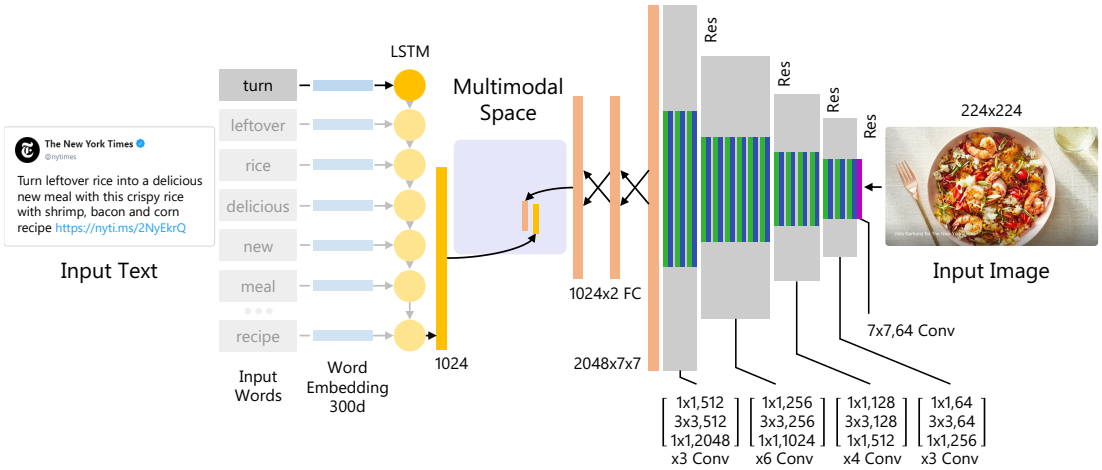


Figure 4.3: Proposed end-to-end multimodal neural network architecture for learning the image and text representations. Image features are extracted by a CNN with 16 residual blocks and text features are extracted by recurrent unit. Then the fully-connected layers join the two domains by feature transformation.

$$T(i,j) = \begin{cases} 1 & j = l_i \\ 0 & otherwise \end{cases}$$

where $1 \leq i \leq m$ and $1 \leq j \leq d$.

Based on this definition, each text should have a fixed length. In this study, since two datasets with various distributions are employed, the length of each sentence is considered a fixed number. In order to meet this criterion, when there are several sentences for one image, we concatenate all the words and build a long description for that image. When the length grows to be more than the expected length, the extra words are removed and when there are fewer words, zero-padding is applied. Therefore, we will have a $70 \times d$ dimension space for the representation of the sentences.

The input of the text representation model is a sequence of integer numbers. In the next step, a word embedding is used to reduce the number of semantically similar words or to remove the words with low frequency, which are non-existent in the dictionary, resulting in a new embedding space. Since the vocabulary size is very large, the reduction is helpful in increasing the networks generalizability. The new embeddings are then fed into an LSTM [Gers et al., 1999] to learn a probability distribution over the above-mentioned sequence in order to predict the next word. The output of the LSTM is not used for word-level labeling. Instead, for the representation of the whole text, only the last hidden state is utilized. Therefore, for the input sentence T, its text descriptor denoted by f_{txt} and using the function $F_{txt}(.)$, is computed

Task	Sentence Retrieval				Image Retrieval				
Method	R@1	R@5	R@10	M r	R@1	R@5	R@10	M r	
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500	1K test images
STV [Kiros et al., 2015]	33.8	67.7	82.1	3	25.9	60.0	74.6	4	
DVSA [Karpathy et al., 2015]	38.4	69.9	80.5	1	27.4	60.2	74.8	3	
GMM-FV [Klein et al., 2015]	39.0	67.0	80.3	3	24.2	59.3	76.0	4	
MM-ENS [Klein et al., 2015]	39.4	67.9	80.9	2	25.1	59.8	76.6	4	
m-RNN [Mao et al., 2014]	41.0	73.0	83.5	2	29.0	42.2	77.0	3	
m-CNN [Ma et al., 2015]	42.8	73.1	84.1	2	32.6	68.6	82.8	3	
HM-LSTM [Niu et al., 2017]	43.9	-	87.8	2	36.1	-	86.7	3	
SPE [Wang et al., 2016]	50.1	79.7	89.2	-	39.6	75.2	86.9	-	
VQA-A [Lin and Parikh, 2016]	50.5	80.1	89.7	-	37.0	70.9	82.9	-	
2WayNet [Eisen. et al., 2017]	55.8	75.2	-	-	39.7	63.3	-	-	
sm-LSTM [Huang et al., 2017]	53.2	83.1	91.5	1	40.7	75.8	87.4	2	
RRF-Net [Liu et al., 2017]	56.4	85.3	91.5	-	43.9	78.1	88.6	-	
VSE++ [Faghri et al., 2017]	64.6	90.0	95.7	1	52.0	84.3	92.0	1	
SCAN [Lee et al., 2018]	72.7	94.8	98.4	-	58.8	88.4	94.8	-	
DeepTwt (ours)	47.5	81.0	91.0	2	48.4	84.3	91.5	2	
GMM-FV [Klein et al., 2015]	17.3	39.0	50.2	10	10.8	28.3	40.1	17	5K test images
DVSA [Karpathy et al., 2015]	16.5	39.2	52.0	9	10.7	29.6	42.2	14	
VQA-A [Lin and Parikh, 2016]	23.5	50.7	63.6	-	16.7	40.5	53.8	-	
VSE++ [Faghri et al., 2017]	41.3	71.1	81.2	2	30.3	59.4	72.4	4	
SCAN [Lee et al., 2018]	50.4	82.2	90.0	-	38.6	69.3	80.4	-	
DeepTwt (ours)	23.8	53.7	67.3	4	25.6	55.1	68.4	3	

Table 4.1: Image and sentence retrieval results on MS-COCO. Sentence Retrieval denotes using an image as query to search for the relevant sentences, and Image Retrieval denotes using a sentence to find the relevant image. R@K is Recall@K (high is good). Med r is the median rank (low is good).

as $f_{txt} = F_{txt}(T)$. The final word feature vector (v_{txt}) is defined by $f_{txt}(.)$.

4.4.3 Alignment Objective

Having an aligned collection of image-text pairs, the goal is to learn the image-text similarity score denoted by $S(T, I)$ which is defined as follows:

$$S(T, I) = -E(v_{txt}, v_{img})$$

where v_{img} and v_{txt} are the same-size image and text representations which have been projected into a partial order visual-semantic embedding space. The penalty paid for every true pair of points that disagree is $E(x, y) = ||max(0, y - x)||^2$.

To compute the training loss, the image and text output vectors (v_{img}, v_{txt}) have been forced to be in the $\mathbb{R}^+$. By merging the image and text embedded models as illustrated in Figure 4.3, we achieve the desired visual-semantic model. To learn an order encoding function, we

considered a hinge-based triplet loss function which encourages positive examples to have zero penalty, and negative examples to have penalty greater than a margin:

$$\sum_{(T,I)}(\sum_{T'}(max\{0, \alpha - S(T,I) + S(T^{'}, I)\} - \sigma^2(T^{'})) \\ + \sum_{I'}(max\{0, \alpha - S(T,I) + S(T, I^{'})\})) - \sigma^2(I^{'}))$$

where $S(T, I)$, the similarity score function, is as described above while $T^{'}$ and $I^{'}$ are inferred from the ground truth by matching contrastive images with each caption and the reverse. $\sigma^2(x)$ is discrete variance written as $\sum_n \frac{x - \mu_c}{|n|}$. For computational efficiency, rather than summing over all the negative samples, we assumed only the negatives in a mini-batch.

4.5 Experiments

4.5.1 Implementation

The proposed method has been implemented with the TensorFlow [Abadi et al., 2016], and Python ran on a machine with GeForce GTX 1080 Ti. For initialization, the GloVe [Pennington et al., 2014] word embeddings, trained on Twitter with 1.2 million vocabulary size, 27 billion tokens and 2 billion tweets, are employed. The training phase starts with an Adam optimizer with learning rate of 0.1 and a batch size of 16 and continues as long as the amount of loss does not change. When it happens, the learning rate is divided by 2. This continues until the learning rate becomes 10^{-7}. Then, the batch size is doubled and the learning rate is reset to 0.1. We repeat this process to optimize the model. During the training, a grid search over all the hyper-parameters is carried out in order to conduct a model selection. For efficiency, the training is performed in batches which allows us to do real-time data augmentation on images in CPU in parallel with training the model in GPU.

4.5.2 Evaluation

Given a sentence (image), all the images (captions) of the test set are retrieved and listed based on their penalty in an increasing order. Then we report the results using *Recall* and *Median Rank*. *Recall* is a metric for assessing how well a system retrieves information to a query. It is computed by dividing the number of relevant retrieved results by the total number of instances. In $R@K$, the top K results are treated as the output and the *Recall* is computed accordingly. *Med r* is the middle number in a sorted sequence of the retrieved instances.

To address this issue, other metrics can be taken into account since the existing measures can be intrinsically problematic. For instance, the retrieval of the exact image (text) is not guaranteed. In these cases, since the exact matches have not been retrieved, its score is considered although similar ones have been matched.

4.5.3 Data Collection and Results

Several other datasets have been published for image-sentence retrieval task [Farhadi et al., 2010, Hu et al., 2017, Ordonez et al., 2011, Rashtchian et al., 2010, Young et al., 2014]. We collect a dataset, as a proof of concept, for evaluating and analyzing our method to better showcase its ability to generalize as well as for demonstrating the extensibility of this type of solution to conversational texts and unusual images.

In addition, we used MS-COCO [Lin et al., 2014] to train and test the proposed model. This dataset contains 123,287 images and 616,767 descriptions [Lin et al., 2014]. Each image contains 5 textual descriptions on average which collected by crowdsourcing on AMT. The average caption length is 8.7 words after rare word removal. We follow the protocol in [Karpathy et al., 2014] and use 5000 images for both validation and testing, and also report results on a subset of 1000 testing images in Table 4.1.

We collected 13751 tweets with 14415 images by a crawler based on the Twitter API. To make sure that the collection is diverse, we first created a list of seed users. Then, the followers of the seed accounts were added to the list. Next, the latest tweets of the users in our list were extracted and saved in the dataset. To make the data appropriate for our task, we removed retweets, the tweets with no images, non-English tweets and the ones that had less than three words. This led the dataset to have a relatively long description for each image and at least one image for every tweet. At the final step, the dataset was examined by two professionals and unrelated content was removed by them.

Appendix A shows samples of the extracted dataset. This collection is different from currently existing ones due to varied domains, informal texts and high level correlation between text and image. For instance, the tweets may contain abbreviations, initialisms, hashtags or URLs. On collected tweets, our model improves sentence retrieval by 14.3% relatively and image retrieval by 16.4% relatively based on $R@1$. The dataset has been available.[1]

[1]Dataset, source codes and model is publicly available at https://iasbs.ac.ir/~ansari/deeptwitter/index.html.

4.6 Conclusions

We propose a multi-modal image-text matching model using a convolutional neural network and a long short-term memory along with fully-connected layers. They are employed to map image and text inputs into a shared feature space, where their representations can be compared, to find the closest pairs. Additionally, a new dataset of images and tweets extracted from Twitter is introduced, with the aim of having a characteristically different collection from the benchmarks. Whereas the descriptions in the benchmarks are well-organized, our dataset has not been standardized and the image-text pairs can contain high semantic correlations. Also, because of a varied number of domains existent in the extracted dataset, the task of image-text matching becomes even more challenging. Therefore, it can be used to carry out new research and assess the robustness of the proposed frameworks. Our experiments on MS-COCO yield improved results over some previously proposed

Chapter 5

Conclusions

We have discussed the prospects and problems of automatic image captioning. Captioning is a good task to measure the progress in both visual feature extraction and language generation research. Despite recent rapid progress in visual recognition, it is clear that many challenges still remain before we can realize the Turings vision of machines that can sense the visual world and interact with us through natural language. In many cases, captions generated for images are descriptive and accurate. One can find a good amount of novel captions generated by the models, showing that they not only learn to just mimic the training data, but also to use the phrases seen in the training set in novel compositions. The generative language model is also able to use correct grammar while describing the visual content. Nevertheless, there are still many areas where the captioning models make mistakes.

First and foremost, the vocabulary the models learn to use is very limited because of the number of unique words that model seen at training time. A method to integrate new words into the vocabulary of the model, without re-training it entirely, would be a good extension to make such captioning models viable to use on images and videos in the wild.

Another major bottleneck hindering the progress of captioning systems is the lack of effective and efficient methods to evaluate them. These metrics perform better when they have access to a larger number of reference captions, but those are expensive to collect. Despite the impressive performance of the model, its performance is not interpretable. Specifically, it is not apparent how much the visual features are responsible for the generation of a specific word and how much it is caused solely by the bias in the model especially in the language model.

In addition, we showed this problem can be applied to other domains by changing the

language and vision model. Also, learning can be done on noisy datasets with blur or low-resolution images without any good captions or newly collected dataset with non-standard text-image pairs.

Appendix A

Examples from collected dataset

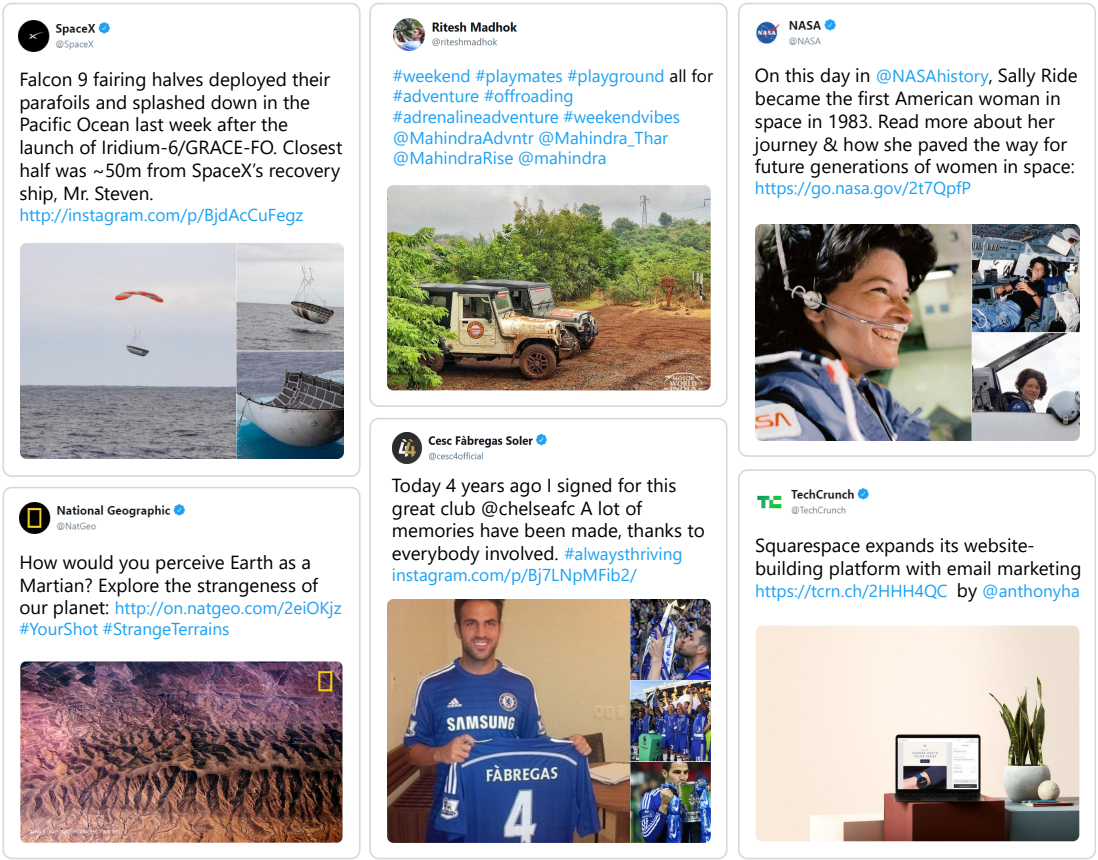


Figure A.1: Few example Tweets in our collected dataset.

Appendix B

Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

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.

Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. Déja image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 504–514, 2015.

Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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*, pages 376–380, 2014.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*, 2015.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

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.

Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4601–4611, 2017.

Desmond Elliott and Arjen de Vries. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 42–52, 2015.

Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, 2013.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

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 computer vision and pattern recognition*, pages 1473–1482, 2015.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.

Yansong Feng and Mirella Lapata. Automatic caption generation for news images. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):797–812, 2012.

Ralf Gerber and N-H Nagel. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 2, pages 805–808. IEEE, 1996.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *IET*, 1999.

Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision*, pages 529–545. Springer, 2014.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2, 2006.

Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018.

Ankush Gupta, Yashaswi Verma, and CV Jawahar. Choosing linguistics over vision to describe images. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

Micah Hodosh and Julia Hockenmaier. Sentence-based image description with scalable, explicit models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 294–300, 2013.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013.

Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia*, 20(4):927–938, 2017.

Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017.

Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

Instagram. Instagram press. `https://instagram-press.com/`.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision*, pages 2407–2415, 2015.

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*, pages 3128–3137, 2015.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015.

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

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (12):2891–2903, 2013.

Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 359–368. Association for Computational Linguistics, 2012.

Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2:351–362, 2014.

Yann Le Cun, Leon Bottou, and Yoshua Bengio. Reading checks with multilayer graph transformer networks. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 151–154. IEEE, 1997.

Rémi Lebret, Pedro O Pinheiro, and Ronan Collobert. Simple image description generator via a linear phrase-based approach. *arXiv preprint arXiv:1412.8419*, 2014.

Rémi Lebret, Pedro O Pinheiro, and Ronan Collobert. Phrase-based image captioning. *arXiv preprint arXiv:1502.03671*, 2015.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*, pages 833–850. Springer, 2016.

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*, pages 220–228. Association for Computational Linguistics, 2011.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of*

the North American Chapter of the Association for Computational Linguistics, pages 150–157, 2003.

Dahua Lin, Chen Kong, Sanja Fidler, and Raquel Urtasun. Generating multi-sentence lingual descriptions of indoor scenes. *arXiv preprint arXiv:1503.00064*, 2015.

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.

Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277. Springer, 2016.

Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4107–4116, 2017.

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.

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

Rebecca Mason and Eugene Charniak. Nonparametric method for data-driven image captioning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, 2014.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1881–1889, 2017.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151, 2011.

Luis Gilberto Mateos Ortiz, Clemens Wolff, and Mirella Lapata. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515, 2015.

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.

Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108 (1-2):59–81, 2014.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

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.

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.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

Chen Sun, Chuang Gan, and Ram Nevatia. Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE international conference on computer vision*, pages 2596–2604, 2015.

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*, pages 1–9, 2015.

Telegram. Telegram press. `https://telegram.org/press`.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Yoshitaka Ushiku, Masataka Yamaguchi, Yusuke Mukuta, and Tatsuya Harada. Common subspace for model and similarity: Phrase learning for caption generation from images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2668–2676, 2015.

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.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

Yashaswi Verma and CV Jawahar. Im2text and text2im: Associating images and texts for cross-modal retrieval. In *BMVC*, volume 1, page 2. Citeseer, 2014.

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*, pages 3156–3164, 2015.

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

Qingzhong Wang and Antoni B Chan. Cnn+ cnn: Convolutional decoders for image captioning. *arXiv preprint arXiv:1805.09019*, 2018.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

Semih Yagcioglu, Erkut Erdem, Aykut Erdem, and Ruket Cakici. A distributed representation based query expansion approach for image captioning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 106–111, 2015.

Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3441–3450, 2015.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454. Association for Computational Linguistics, 2011.

Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, pages 110–120, 2014.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

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.

Youtube. Statistics. `https://www.youtube.com/yt/press/en-GB/statistics.html`.

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*, 2017.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016, 2013.

Appendix C

The summary in Persian

خزنده Crawler

دنبال‌کننده Follower

رابط برنامه‌نویسی کاربردی Application Programming Interface

ریشه‌یاب Stemmer

زبان طبیعی Natural Language

شبکهٔ عصبی بازگشتی Recurrent Neural Network

شبکهٔ عصبی پیچشی Convolutional Neural Network

قابل یادگیری Learnable

قابلیت تعمیم Generalizability

گذر پیشرو Forward Pass

لایهٔ تماما متصل Fully-connected Layer

لغت‌نامه Dictionary

متراکم Dense

مجموعه‌داده Dataset

نشانی وب Uniform Resource Locator

نقطه به نقطه End-to-end

نویسه Character

نهُفته Embedding

همبستگی Correlation

هم‌ترازی Alignment

هوش مصنوعی Artificial intelligence

یادگیری عمیق Deep Learning

واژه‌نامه فارسی به انگلیسی

اختصارات ..... Initialisms

استخراج ویژگی ..... Feature Extraction

الحاق کردن ..... Concatenate

اولیه ..... Seed

بازنمایی ..... Representation

بازیابی ..... Retrieval

بینایی رایانه‌ای ..... Computer Vision

تنظیم دقیق ..... Fine-tune

توالی ..... Sequence

توسعه‌پذیری ..... Extensibility

توصیف ..... Description

توصیف متنی تصویر ..... Image Captioning

توصیف‌گر بصری ..... Visual Descriptor

جاسازی کلمه ..... Word Embedding

جملات کوتاه‌شده ..... Abbreviation

چندحالته ..... Multimodal

حاشیه ..... Margin

حاشیه‌نویسی ..... Annotate

حافظهٔ کوتاه‌مدت ماندگار ..... Long Short-term Memory

آن‌هایی که کمتر از سه کلمه داشتند را حذف کردیم. این‌کار باعث شد تا مجموعه‌دادهٔ جدید شامل توصیفات نسبتا طولانی برای هر تصویر و حداقل یک تصویر به‌ازای هر توئیت باشد. در مرحلهٔ پایانی مجموعه‌داده توسط دو فرد متخصص بررسی شد و محتوای غیرمرتبط توسط آن‌ها حذف شد. مجموعهٔ فعلی به علت گوناگونی ساختار، متن‌های غیررسمی و همبستگی بالای معنایی تصاویر و متن توئیت‌ها با مجموعه‌داده‌های موجود متفاوت است. برای مثال، توئیت‌ها ممکن است حاوی اختصارات، جملات کوتاه‌شده، هشتگ[1] یا نشانی وب باشند یا ارتباط تصاویر یا متن ضمنی باشد. این مجموعه‌دادهٔ جدید در دسترس عموم قرار داده شده است.

نتیجه‌گیری

ما یک مدل چندحالته با استفاده از شبکهٔ عصبی پیچشی، حافظهٔ کوتاه‌مدت ماندگار و لایه‌های تماما متصل برای انطباق تصویر و متن پیشنهاد کردیم. این شبکه‌ها با ترسیم داده‌ها در یک فضای مشترک جدید که بازنمایی قابل‌مقایسه‌ای داشته باشند، جفت‌های مشابه را می‌یابند. علاوه‌براین، با هدف داشتن مجموعهٔ کاملا متفاوت از ویژگی‌ها، مجموعه‌دادهٔ جدیدی از توئیت‌ها که از توییتر دریافت شده‌است، معرفی شده‌است. در حالی که توصیف‌ها در داده‌های معیار به خوبی سازمان‌یافته هستند، مجموعه‌دادهٔ ما استاندارد نشده است و داده‌ها می‌توانند حاوی همبستگی معنایی بالا باشند. بنابراین می‌توان از آن برای انجام تحقیقات جدید و ارزیابی میزان خطای مدل های فعلی استفاده کرد. همچنین آزمایش ها بر روی مدل پیشنهادی نتایج بهبودیافته در مقایسه با برخی مدل های قبلی را نشان داده‌اند.

[1] Hashtag

$$E(x,y) = ||max(0, y-x)||^2$$

که در آن مقادیر بردارهای تصویر و متن خروجی (v_{img}, v_{txt}) در زمان آموزش، در بازهٔ $\mathbb{R}^+$ هستند. در ادامه برای آموزش تابع کدگذاری، تابع $S(T,I)$ را برای یافتن شباهت، بر اساس اتلاف هینج[1] سه‌گانه به شکل زیر درنظر می‌گیریم:

$$\sum_{(T,I)}(\sum_{T'}(max\{0, \alpha - S(T,I) + S(T',I)\} - \sigma^2(T'))$$
$$+ \sum_{I'}(max\{0, \alpha - S(T,I) + S(T,I')\})) - \sigma^2(I'))$$

که T' و I' از مقادیر درست با تطبیق دادن تصاویر غیرمرتبط با هر متن و برعکس استنباط می‌شود. این تابع تلاش می‌کند تا نمونه‌های مثبت هزینهٔ صفر و نمونه‌های منفی فاصله‌ای بیشتر از یک حاشیه داشته باشند. در این‌جا $\sigma^2(x)$، مقدار واریانس گسسته کل مجموعه‌داده‌هاست که از رابطه $\sum_n \frac{x-\mu_c}{|n|}$ محاسبه می‌شود.

جمع‌آوری داده‌ها

چندین مجموعه‌داده که در فصل ۳ معرفی شده‌اند، برای بازیابی تصویر–متن منتشر شده‌اند. ما یک مجموعه‌داده را به عنوان شاهدی بر ادعای قابلیت تعمیم‌پذیری و توسعهٔ این نوع راه‌حل ها به متون محاوره‌ای و تصاویر غیرمعمول، جمع‌آوری کرده‌ایم.

۱۳۷۵۱ توئیت با استفاده از یک خزنده، براساس رابط برنامه‌نویسی کاربردی توییتر جمع‌آوری شده است. برای اطمینان از این‌که مجموعه متنوع است، ابتدا فهرستی از کاربران اولیه ایجاد کردیم. سپس دنبال‌کنندگان فهرست اولیه به فهرست اضافه شدند و این‌کار تا رسیدن به یک فهرست بزرگ‌تر ادامه یافت. سپس، آخرین توئیت‌های کاربران در فهرست ما در مجموعه‌داده‌ها استخراج و ذخیره شد. برای این‌که داده‌ها برای کار ما مناسب باشند، ریتوئیت[2]ها، توئیت های بدون تصویر، غیرانگلیسی، و

[1] Hinge
[2] Retweet

برآورده کردن این معیار، زمانی که چندین جمله برای یک تصویر وجود دارد، ما همهٔ کلمات را الحاق می‌کنیم و یک توصیف طولانی برای آن تصویر می‌سازیم. هنگامی که طول جملهٔ نهایی بزرگ‌تر از طول پیش‌بینی‌شده باشد، کلمات اضافی حذف می‌شوند و وقتی کلمات کمتری وجود دارند، پدینگ صفر[۱] اعمال می‌شود. در آخر، یک فضای ۷۰ بُعدی برای نمایش جملات داریم و ورودی مدل بازنمایی متن، توالی اعداد صحیح است.

در مرحلهٔ بعد، از یک جاسازی کلمه برای کاهش تعداد کلمات مشابه یا حذف کلمات با تکرار پایین استفاده می‌شود، که در فرهنگ لغت وجود ندارد. این‌کار منجر به یک فضای نهُفتهٔ جدید می‌شود. چون اندازهٔ واژگان خیلی بزرگ است، کاهش، در افزایش قابلیت تعمیم این شبکه مفید است. سپس داده‌ها در این فضای نهُفته به یک حافظهٔ کوتاه‌مدت ماندگار داده می‌شوند تا یک توزیع احتمال بر روی توالی ورودی را یاد بگیرد و کلمهٔ بعدی را پیش‌بینی کند. بنابراین برای توالی ورودی T، توصیف‌گر متن f_{txt} با تابع $F_{txt}(.)$ نمایش داده می‌شود و به صورت $f_{txt} = F_{txt}(T)$ محاسبه می‌شود. بردار نهایی ویژگی کلمات v_{txt} نیز با $f_{txt}(.)$ تعریف می شود.

تابع هدف هم‌ترازی

با داشتن زوج‌های تصویر_متن، هدف یاد گرفتن امتیاز شباهت تصویر_متن $S(T, I)$ است که به صورت زیر تعریف می‌شود:

$$S(T, I) = -E(v_{txt}, v_{img})$$

که در آن v_{img} و v_{txt} بردارهای بازنمایی هم اندازه از تصویر و متن هستند که مطابق آنچه گفته شد، به یک فضای جزئی نهُفته بصری_معنایی نگاشت شده اند. تابع هدف نیز بر اساس هزینهٔ زیر بهینه خواهد شد:

[۱] Zero-padding

می‌شود.

اغلب مدل تصویر به عنوان بخشی از شبکهٔ عصبی درنظر گرفته می‌شود و با وزن‌های قبل از آموزش، به همراه دیگر بخش‌های شبکه تنظیم دقیق می‌شود. با این حال، با توجه به تعداد زیاد پارامترهای قابل یادگیری که فرآیند آموزش را بسیار وقت گیر می‌کند، ویژگی‌های از پیش محاسبه شده مورد استفاده قرار می‌گیرند.

در آخر، دو لایه ۱۰۲۴ تماما متصل برای تبدیل $f_{img}(.)$ به بردار ویژگی v_{img} برای بازنمایی تصویر، که توسط $b_{img} + f_{img}(.)W_{img} = v_{img}$ محاسبه می شود، به انتهای مدل اضافه میکنیم.

بازنمایی متن

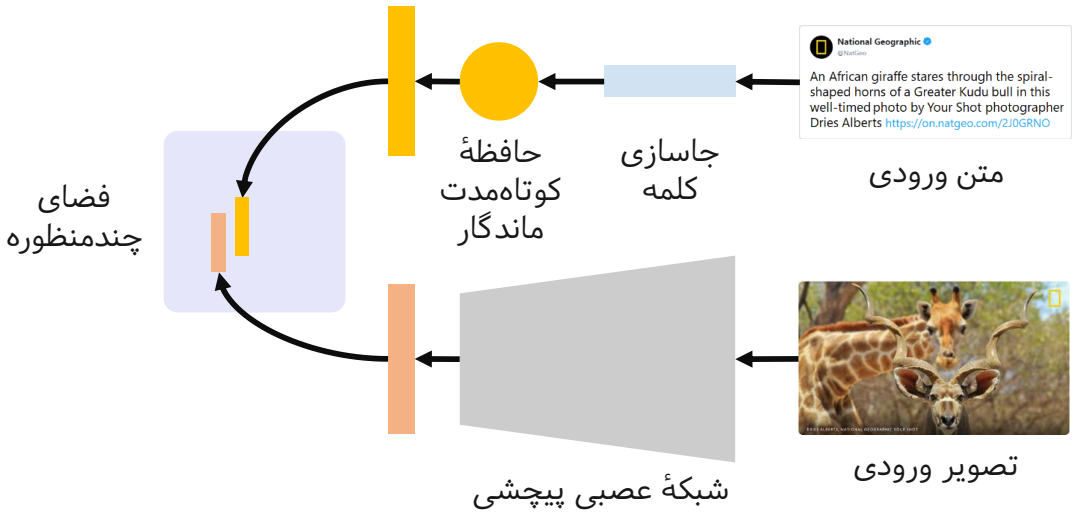
هر متن ورودی T ابتدا با یک ماتریس $d \times n$ بازنمایی می‌شود که n طول جمله و d اندازهٔ لغت‌نامه است. برای ساختن لغت‌نامه، کلمات و علائم نقطه‌گذاری حذف می‌شوند و همهٔ کلمات با استفاده از ریشه‌یاب پورتر[۱] به ریشهٔ اصلی بازگردانده می‌شوند. علاوه‌براین، حذف نویسه‌های خاص انجام می‌شود و بقیه کلمات در قالب حروف کوچک درنظر گرفته می‌شوند. هر کلمه در لغت‌نامهٔ نهایی با یک بردار وان‌ـ‌هات[۲] با d بُعد بازنمایی می‌شود که هر کلمه در شاخص l در دیکشنری قرار دارد. درنتیجه برای یک جملهٔ ورودی T با m کلمه، یک ماتریس $m \times d$ به صورت زیر وجود دارد:

$$T(i,j) = \begin{cases} ۱ & j = l_i \\ ۰ & otherwise \end{cases}$$

که در آن $۱ \leq i \leq m$ و $۱ \leq j \leq d$.

براساس این تعریف هر متن باید دارای طول ثابت باشد. در این مطالعه، از آن‌جا که دو مجموعه‌داده با توزیع‌های مختلف به کار گرفته می‌شوند، طول هر جمله یک عدد ثابت محسوب می‌شود. به منظور

[۱] Porter
[۲] One-hot



شکل ۱: ساختار مدل شبکهٔ عصبی نقطه به نقطه برای یادگیری بازنمایی تصویر و متن. ویژگی‌های تصویر با استفاده از شبکهٔ پیچشی و ویژگی‌های متن با شبکهٔ بازگشتی استخراج می‌شوند. سپس لایه‌های تماما متصل با تبدیل ویژگی‌ها، این دو را به یک فضای جدید می‌برند.

بازنمایی تصویر

برای مقداردهی اولیهٔ مدل، پس از حذف لایه‌های تماما متصل از رزنت‌ـ۵۰[۱] که با ایمج‌نت[۲] آموزش دیده شده است، لایه‌های باقیمانده را به عنوان استخراج‌کننده ویژگی درنظر می‌گیریم. سپس شبکه را با داده‌های جدید، تنظیم دقیق می کنیم. ورودی‌های شبکه تصاویر ۲۲۴×۲۲۴ پیکسل[۳] هستند و خروجی یک بردار ۷×۷×۲۰۴۸ است. به منظور مقایسهٔ بازنمایی‌های تصویر و متن، آن ها باید هم بُعد باشند. بنابراین یک لایهٔ متراکم با اندازه دامنهٔ متنی به انتهای شبکه اضافه می‌کنیم. علاوه بر بقیهٔ شبکه، این لایه که اکنون بخشی از مدل است، برای تولید بازنمایی‌های تصویر آموزش داده می‌شود. اگر بازنمایی تصویر ورودی را I بنامیم، f_{img} یک توصیف‌گر بصری است که نتیجهٔ گذر پیشرو در شبکه است. گذر پیشرو با $F_{img}(.)$ مشخص می‌شود، که یک تابع غیرخطی است و به صورت $f_{img} = F_{img}(I)$ تعریف

[۱] ResNet-۵۰
[۲] ImageNet
[۳] Pixel

مدل پیشنهادی

در مدل پیشنهادی با یادگیری یک فضای تعبیه‌شده تصویر–متن، بازیابی متقابل توصیفات تصاویر یا یافتن یک تصویر مرتبط با توصیف متنی ورودی، انجام می‌گیرد. پیدا کردن چنین فضایی یک کار چالش برانگیز است چون ویژگی‌ها و بازنمایی متن و تصویر قابل‌مقایسه نیستند. مدل یک شبکهٔ عصبی پیچشی–بازگشتی چندحالتهٔ نقطه به نقطه برای یادگیری همزمان بازنمایی زبان و بینایی برای یافتن شباهت تصویر–متن است. این مدل با استفاده از یک تابع هزینه بر پایه اتلاف هینج[۱] یاد می‌گیرد که کدامیک از جفت تصویر–متن منطبق (مثبت) یا کدامیک نامنطبق (منفی) هستند. همچنین برای یادگیری بازنمایی مشترک، از مجموعهٔ جدید استخراج‌شده‌ای از توییتر[۲] استفاده کرده‌ایم. ویژگی اصلی مجموعه‌دادهٔ مذکور این است که تصاویر و متن در توئیت[۳]ها همانند مجموعه معیار استاندارد نشده‌اند و همبستگی معنایی بالاتری با یکدیگر دارند. نتایج تجربی در مقایسه با مجموعه‌داده معیار نشان می‌دهد که مدل ما از برخی روش‌هایی که قبلا ارائه شده‌است بهتر عمل می‌کند و عملکرد رقابتی در مقایسه با آنها دارد.

معماری مدل که در شکل ۱ نشان داده شده است شامل دو زیرشبکهٔ اصلی است. یک شبکهٔ پیچشی برای بازنمایی تصویر و یک حافظهٔ کوتاه‌مدت ماندگار با یک جاسازی کلمه برای نگاشت کلمات به فضای جدید است. هدف یافتن یک نگاشت واحد در فضای مشترک تعبیه‌شدهٔ جدید برای هر جفت داده است. در این فضا، یک تصویر (یا متن) نمایش مشابه‌ای را به متن (یا تصویر) مرتبط با خود اما متفاوت از متون (یا تصاویر) دیگر خواهد داشت. هنگامی که مدل آموزش داده شده باشد، با ارائهٔ یک تصویر (یا متن) به مدل، مشابه‌ترین متن (یا تصویر) پیدا خواهد شد. توضیحات این بخش به تفضیل در فصل ۴ پایان‌نامه آمده است.

---

۱ Hinge

۲ Twitter

۳ Tweet

مقدمه

در مسئلهٔ توصیف متنی خودکار تصاویر، با داشتن یک تصویر ورودی، هدف تولید یک متن است که تمامی یا قسمت‌هایی از تصویر را توصیف کند. حل این مساله نیازمند این است که ماشین بتواند اشیا برجستهٔ تصویر و ویژگی‌های آن‌ها را تشخیص دهد، روابط بین این اشیا و اتفاقات پیرامون آن‌ها را درک کند و همچنین به درستی صحنه را تشخیص دهد. تولید یک توصیف متنی، نیازمند استخراج ویژگی‌های تصویر و همزمان شناخت زبان طبیعی است. در واقع، یک جمله باید نه تنها اشیا موجود در یک تصویر را توصیف کند، بلکه باید بیان کند که چگونه این اشیا به یکدیگر و نیز رابطه‌شان و فعالیت‌هایی که انجام می‌دهند مرتبط هستند.

تا چند سال پیش، تشخیص قابل‌اطمینان حتی یک شی در تصویر در یک مجموعه‌داده بزرگ و متنوع سخت بود. این امر به طور قابل‌توجهی با در دسترس بودن داده‌های حاشیه‌نویسی شده در مقیاس بزرگ و کاربرد تکنیک‌های یادگیری عمیق، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی تغییر کرده است که منجر به کاربرد موفق این شبکه‌های عمیق در وظایف مختلف از جمله توصیف متنی تصویر و بازیابی تصویر–متن شده‌است. شبکه های عمیق یاد داده شده بر روی این داده ها به مسائل دیگر به خوبی تعمیم داده می‌شوند و می‌توانند با دانش یک مجموعه‌داده یا مسئله، مسئله‌ای دیگر را حل کنند.

در این پایان‌نامه یک روش جدید پیشنهاد می‌کنیم که وظیفهٔ تولید خودکار یک توصیف متنی برای تصویر ورودی را بر عهده دارد. در این پژوهش، با ارائه یک شبکهٔ عصبی چندحالتهٔ نقطه به نقطه و کمک یادگیری ماشین این مسئله را حل کرده‌ایم. مدل پیشنهادی قادر به بازیابی یک تصویر براساس توصیف (جستجو توسط تصویر) و بازیابی توصیف بر پایهٔ یک تصویر (حاشیه‌نویسی تصویری) است.

پیشگفتار

«زلیخا گفتن و یوسف شنیدن شنیدن کی بُوَد مانند دیدن»

دیدن یک تصویر می‌تواند مفاهیم بیشتری را از شنیدن یا خواندن یک جمله انتقال دهد یا به بیان دیگر اطلاعات کدگذاری شده در تصویر ارزش چند هزار کلمه را دارند. این امر در استفادهٔ گسترده از تصاویر در رسانه‌های ارتباطی، از مجلات گرفته تا تلگرام و اینستاگرام مشهود است. انسان‌ها در پردازش تصاویر و ویدیو بسیار خوب هستند و همهٔ این اطلاعات کدگذاری شده را جمع‌آوری می‌کنند، اما کامپیوترها هنوز برای درک ساده‌ترین چیزها تلاش می‌کنند. در حال حاضر می توان گفت برای کامپیوترها، دیدن، تجزیه و تحلیل، جستجو و حتی درک هزاران کلمه از مقالهٔ بزرگ راحت‌تر از یک تصویر واحد ساده است.

استفاده از رسانه‌های تصویری در اینترنت در سال‌های اخیر به دلیل دسترسی آسان به دوربین در تلفن‌های هوشمند، به شدت افزایش یافته است. به عنوان مثال، بیش از ۷۰۰ میلیون عکس هر روز در تلگرام به اشتراک گذاشته می‌شود، حدود ۹۵ میلیون عکس و ویدیو هر روز در اینستاگرام پخش می‌شوند و حدود ۵۰۰ ساعت ویدیو در هر دقیقه در یوتیوب به اشتراک گذاشته می‌شود. افزایش سریع رشد داده‌های تصویری ناشی از این پدیده، چالش بزرگی برای درک تصاویر ایجاد می‌کند و فرصتی برای ایجاد الگوریتم‌های رایانه‌ای باهوش‌تر برای درک و خلاصه‌کردن داده‌ها است. از این‌رو، درک خودکار رسانه‌های دیداری یک مشکل قابل‌توجه و مهم در بسیاری از جنبه‌های بینایی رایانه‌ای و هوش مصنوعی است.

چکیده

یک موضوع تحقیقاتی مهم در شناخت بصری، توصیف متنی خودکار تصاویر است. این کار شامل الگوریتمی است که تصویر را به عنوان ورودی می‌گیرد و یک متن تولید می کند که تمامی یا قسمت‌هایی از تصویر را توصیف می‌کند. چنین سیستمی کاربردهای گسترده‌ای از قبیل حاشیه‌نویسی تصاویر و استفاده از توصیفات طبیعی برای جستجوی تصاویر یا متون دارد. این امر به طور قابل‌توجهی با در دسترس بودن داده‌های حاشیه‌نویسی شده در مقیاس بزرگ و کاربرد تکنیک‌های یادگیری عمیق، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی تغییر کرده است که منجر به کاربرد موفق این شبکه‌های عمیق در وظایف مختلف از جمله توصیف متنی تصویر و بازیابی تصویر_ متن شده‌است.

در این پایان‌نامه، یک شبکهٔ پیچشی_ بازگشتی چندحالتهٔ نقطه به نقطه را برای یادگیری همزمان بازنمایی زبان و بینایی برای یافتن شباهت تصویر_ متن پیشنهاد شده است. این مدل قادر به بازیابی یک تصویر براساس توصیف (جستجو توسط تصویر) و بازیابی توصیف بر پایهٔ یک تصویر (حاشیه‌نویسی تصویری) است. در این‌جا برای یادگیری بازنمایی مشترک، از مجموعهٔ جدید استخراج‌شده‌ای از توییتر استفاده می‌کنیم. ویژگی اصلی مجموعه، این است که داده ها بدون تغییر در متن و تصویر، همبستگی معنایی بالاتری با یکدیگر در مقایسه با مجموعه‌داده‌های معیار که در آن توصیفات به خوبی سازماندهی شده‌اند، دارد. هنگامی که مدل آموزش داده می‌شود، با ارائه یک تصویر (یا متن) به مدل، مشابه‌ترین متن (یا تصویر) را پیدا می‌کنیم. در این پژوهش نشان داده شده است که مدل پیشنهادی و داده‌های جدید، می‌توانند از نظر معیارهای ارزیابی روی مجموعه‌داده معیار نسبت به مدل‌های قبلی بهتر عمل کنند. مجموعه‌داده جمع‌آوری‌شده و پیاده سازی مدل در دسترس عموم قرار داده شده‌است.

واژه‌های کلیدی: *یادگیری ماشین، یادگیری عمیق، شبکهٔ عصبی پیچشی، شبکهٔ عصبی بازگشتی، مدل چندحالته، بازیابی تصویر_ متن، توصیف متنی تصویر*

وزارت علوم، تحقیقات و فناوری
دانشگاه تحصیلات تکمیلی علوم پایه
گاوازنگ، زنجان

توصیف متنی خودکار تصاویر
با استفاده از شبکهٔ عمیق نهُفته چندمنظوره

پایان‌نامهٔ کارشناسی ارشد علوم کامپیوتر
دانشکدهٔ علوم رایانه و فن‌آوری اطلاعات
دانشگاه تحصیلات تکمیلی در علوم پایه زنجان

هادی عبدی خجسته

اساتید راهنما: ابراهیم انصاری
پروین رزاقی

شهریور ۱۳۹۸

اعضای کمیته پایان نامه

۱. جناب آقای دکتر ابراهیم انصاری
(استاد راهنما)

۲. سرکار خانم دکتر پروین رزاقی
(استاد راهنما)

۳. جناب آقای دکتر مهدی وثیقی
(داور داخلی)

٤. جناب آقای دکتر محسن افشارچی
(داور خارجی)

۵. جناب آقای دکتر محمدرضا فرجی
(ناظر)

نحوه ارزیابی پایان‌نامه‌های کارشناسی ارشد علوم کامپیوتر (پیشنهاد دانشکده به داوران)

موارد مورد بررسی در ارزیابی پایان‌نامه	نمره
سهم علمی	**7**
آیا مسئله مورد بررسی با مسئله ادعا شده در پروپوزال متناسب است؟ آیا مسئله در تاریخ انجام پایان‌نامه به عنوان یک مسئله باز مورد بررسی بوده است؟ آیا سهم علمی پایان‌نامه متناسب با مدت زمان انجام آن بوده است؟ آیا یک ایده جدید بوده است یا از بسط ایده‌های قبلی و ترکیب آنها شکل گرفته است؟	
بیان دقیق صورت مسئله	**1**
آیا صورت مسئله، فرض‌ها، محدودیت‌های و کاربردهای آن به درستی بیان شده است؟	
تشریح و ارائه دقیق روش یا سهم علمی پایان‌نامه (نگارش علمی)	**3**
آیا الگوریتم یا روش ارائه شده به صورت دقیق و مشخصی توضیح داده شده است؟ آیا درستی روش- اثبات‌ها و قضایا- نشان داده شده است؟	
آنالیز نتایج	**3**
آیا ارزیابی و مقایسه روش انجام شده با روش های دیگر و همچنین نقاط قوت و ضعف در تئوری و کاربرد انجام شده است؟ آیا پیاده‌سازی و شبیه‌سازی‌ها (در صورت نیاز) به درستی انجام شده است؟	
تاریخچه	**2**
آیا ادبیات و کارهای مرتبط، معایب، مزایا به طور کامل بررسی و آنالیز آنها انجام شده است؟	
ساختار و نگارش پایان‌نامه	**2**
آیا فصل‌های پایان‌نامه و ارتباط آنها به درستی ساختاربندی شده است؟ آیا اصول نگارشی- املایی و گرامری - و سهولت خوانایی آن به درستی انجام شده است؟ آیا کیفیت جداول، اشکال و دیاگرام‌ها مناسب است؟ آیا قالب مراجع به درستی رعایت شده است؟	
ارائه شفاهی	**2**
آیا پاسخ‌های دانشجو به سوالات مطرح شده از سوی داوران قانع کننده بوده است؟ آیا ارائه شفاهی پایان‌نامه در جلسه دفاع به صورت مناسب- کوتاه و مختصر همراه با تاکید دقیق بر سهم علمی پایان‌نامه و رعایت زمان‌بندی انجام شده است؟	
جمع نمرات	**20**
توجهات: موارد زیر برای افزایش یا کاهش نمرات پیشنهاد می شود که در نظر گرفته شوند	**2**
آیا مقاله‌ای در کنفرانس یا ژورنال‌های معتبر به چاپ رسیده است.؟ کنفرانس یا مجله چه مقدار معتبر است؟ آیا تولید علمی دیگری مانند نرم‌افزار و ... داشته است؟ **آیا دستاوردهای مساله متناسب با طول مدت تحصیل است**؟ چه مقدار تاخیر (نسبت به تحصیل دوساله کارناسی ارشد) در دفاع از پایان نامه دارد؟	