MANTECH Publications
Journal of Artificial Intelligence, Machine Learning and Soft Computing
Volume 4, Issue 1, January-June 2019

A Comprehensive Survey of Deep Learning for Image Captioning

G. Ramya, B. Chaitanya[1], J. Dinesh[2], D. Chandrasekhar[3], Dr. G. S. N. Murthy[4]
AITAM, AP, India
***Email id**: jdinesh949@gmail.com[2]*
***DOI:** http://doi.org/ 10.5281/zenodo.2728950*

Abstract

This Paper will involve developing a model that generates suitable captions for images. This will help in analyzing image and converting the textual content to other useful forms. Image descriptions provide textual information about non-text content that appears on your website, allowing it to be presented auditory, as visual text, or in any other form that is best for the user. Image descriptions are plain text descriptions of images, gifs, videos, and other media. Humans have been captioning images involuntary since decades and now in the age of social media where every image has a caption over various social platforms. Psychologically those things are affected by events and scenarios running in mind or influenced by nearby activities and emotion. Sometimes those are far-far away from real context. Describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing.

***Keywords:** Image Captioning, Artificial Intelligence, Python*

INTRODUCTION

Every day, we encounter a large number of images from various sources such as the internet, news articles, document diagrams and advertisements. These sources contain images that viewers would have to interpret themselves. Most images do not have a description, but the human can largely understand them without their detailed captions. However, machine needs to interpret some form of image captions if humans need automatic image captions from it. Image captioning is important for many reasons. For example, they can be used for automatic image indexing. Image indexing is important for

Content-Based Image Retrieval (CBIR) and therefore, it can be applied to many areas, including biomedicine, commerce, the military, education, digital libraries, and web searching. Social media platforms such as Facebook and Twitter can directly generate descriptions from images. The descriptions can include where we are (e.g., beach, cafe), what we wear and importantly what we are doing there.

IMAGE CAPTIONING

Image captioning is a popular research area of Artificial Intelligence (AI) that deals with image understanding and a language description for that image. Image understanding needs to detect and recognize objects. It also needs to understand scene type or location, object properties and their interactions. Generating well-formed sentences requires both syntactic and semantic understanding of the language.

Understanding an image largely depends on obtaining image features. The techniques used for this purpose can be broadly divided into two categories: (1) Traditional machine learning based techniques and (2) Deep machine learning based techniques.

In deep machine learning based techniques, features are learned automatically from training data and they can handle a large and diverse set of images and videos. For example, Convolutional Neural Networks (CNN) is widely used for feature learning, and a classifier such as Softmax is used for classification. CNN is generally followed by Recurrent Neural Networks (RNN) in order to generate captions.

CNN (Convolution Neural Network):

In neural networks, Convolutional neural network (Conv Nets or CNNs) is one of the main categories to do images recognition, images classifications. Objects detections, recognition faces etc., are some of the areas where CNNs are widely used.CNN image classifications takes an input image, process it and classify it under certain categories (Eg., Dog, Cat, Tiger, Lion).

A computer sees an input image as array of pixels and it depends on the image resolution. Based on the image resolution, it will see h x w x d(h = Height, w = Width, d = Dimension). Eg., An image of 6 x 6 x 3 array of matrix of RGB (3 refers to RGB values) and an image of 4 x 4 x 1 array of matrix of grayscale image.

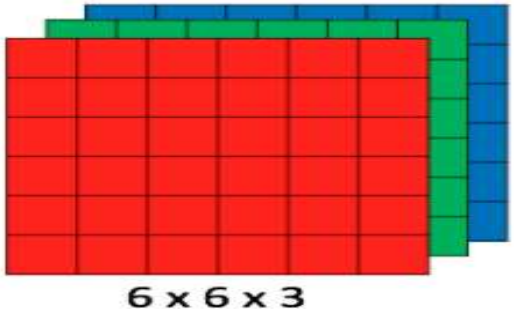

Fig 1: Array of RGB matrix

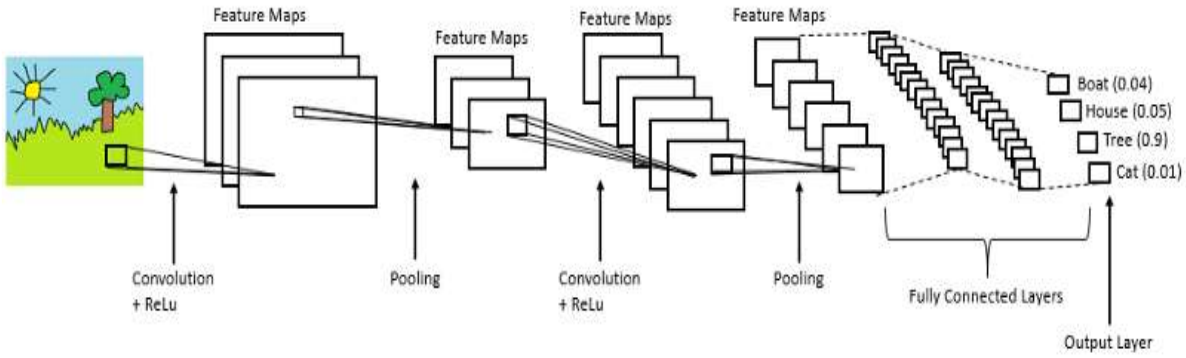

Fig 2: Complete CNN Architecture

Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1. The below figure is a complete flow of CNN to process an input image and classifies the objects based on values.

RNN(Recurrent Neural Network):

Recurrent Neural Networks (RNN) are a powerful and robust type of neural networks and belong to the most promising algorithms out there at the moment because they are the only ones with an internal memory. Because of their internal memory, RNN's are able to remember important things about the input they received, which enables them to be very precise in predicting what's coming next.

This is the reason why they are the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather and much more because they can form a much deeper understanding of a sequence and its context, compared to other algorithms.

TECHNOLOGY

A system configuration (SC) defines the computers, processes, and devices that compose the system and its boundary. More generally, the system configuration is the specific definition of the elements that define and/or prescribe what a system is composed of. Alternatively, the term "system configuration" can be used to relate to a model (declarative) for abstract generalized systems. In this sense, the usage of the configuration information is not tailored to any specific usage, but stands alone as a data set.

SOFTWARE

Software Name	Version
Ubuntu	18.04
Jupyter Notebook	1.8.0

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Data Collection:

There are many open source datasets available for this problem, like Flickr 8k (containing8k images), Flickr 30k (containing 30k images), MS COCO (containing 180k images), etc. But for the purpose of this case study, I have used the Flickr 8k dataset which you can download by filling this form provided by the University of Illinois at Urbana-Champaign. Also training a model with large number of images may not be feasible on a system which is not a very high end PC/Laptop.

MODEL ARCHITECTURE:

Since the input consists of two parts, an image vector and a partial caption, we cannot use the Sequential API provided by the Keras library. For this reason, we use the Functional API which allows us to create Merge Models.

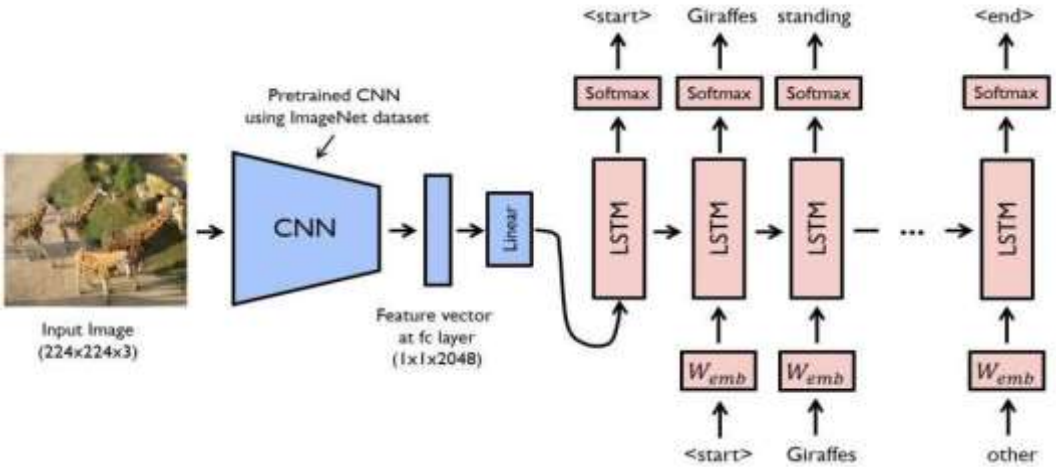

Fig 3: Model Architecture

Image Feature Extraction

In Image Captioning, a CNN is used to extract the features from an image which is then along with the captions is fed into an RNN. To extract the features, we use a model trained on Image Net. I tried out VGG-16, Resnet-50 and InceptionV3. Vgg16 has almost 134 million parameters and its top-5 error on Image Net is 7.3%. InceptionV3 has 21 million parameters and its top-5 error on Image Net is 3.46%. Human top-5 error on Image Net is 5.1%.

I used VGG-16 as my first model for extracting the features. I took an hour to extract features from 6000 training images. This is very slow. Imagine how much time it will take to extract features in the MS-COCO dataset which has 80,000 training images. Resnet-50 was the second model I tried for extracting features. But I didn't train the model for long time because InceptionV3 has a better accuracy than Resnet-50 and almost the same number of parameters. Finally, it was the time of InceptionV3. Since it has very less parameters as compared to VGG-16, it took 20 mins for InceptionV3 to extract features from 6000 images. I also ran this on MS-COCO dataset which contains 80,000 training examples and it took 2 hours and 45 minutes to extract the features.

CONCLUSION

We have presented a deep learning model that automatically generates image captions with the goal of helping visually impaired people better understand their environments. Our described model is based on a CNN that encodes an image into a compact representation, followed by a RNN that generates corresponding sentences based on the learned image features. We showed that this model achieves comparable to state-of-the-art

performance, and that the generated captions are highly descriptive of the objects and scenes depicted on the images. Because of the high quality of the generated image descriptions, visually impaired people can greatly benefit and get a better sense of their surroundings using text-to-speech technology.

REFERENCES

1. Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan Show and Tell: ANeural Image Caption Generator.

2. M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899 http://www.jair.org/papers/paper3994.html

3. CS231n Winter 2016 Lesson 10 Recurrent Neural Networks, Image Captioning and LSTM https://youtu.be/cO0a0QYmFm8?t=32m25s.

4. https://github.com/yashk2810/Image-Captioning

5. https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-todescribe- pictures-c88a46a311b8.

6. https://machinelearningmastery.com/develop-a-deep-learning-caption-generationmodel-in-python/

Cite this Article

G. Ramya, B. Chaitanya, J. Dinesh, D. Chandrasekhar, Dr. G. S. N. Murthy, (2019). **A Comprehensive Survey of Deep Learning for Image Captioning** "Journal of Artificial Intelligence, Machine Learning and Soft Computing", 4(1), 26- 32

AUTHORS' PROFILE

[1] G.Ramya, *Student*
Department: Computer Science Engineering
College: AITAM, AP, India

[2] B.Chaitanya, *Student*
Department: Computer Science Engineering
College: AITAM, AP, India

[3] J.Dinesh, *Student*
Department: Computer Science Engineering
College: AITAM, AP, India

[4] D.Chandrasekhar, *Student*
Department: Computer Science Engineering
College: AITAM, AP, India

[5] Dr.G.S.N.Murthy, *Professor*
Department: Computer Science Engineering
College: AITAM, AP, India