International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 20th July 2020. Link

:http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=ISSUE-07

Title: IMAGE CAPTION GENERATION USING CNN-RNN MODEL AND AUTOMATING WEB ACCESSIBILITY

Volume 09, Issue 07, Pages: 144 - 148

Paper Authors

K. V. Kiran, Sanmitra Dharmavarapu, A.Teja Ratna Kumari

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IMAGE CAPTION GENERATION USING CNN-RNN MODEL AND AUTOMATING WEB ACCESSIBILITY

K. V. Kiran[1 *], Sanmitra Dharmavarapu[2 **], A.Teja Ratna Kumari[3*]**
Dept. of CSE, Wellfare Institute of Science Technology & Management, Visakhapatnam, India.
*goodfriend2899@gmail.com, **kasi.kvk@gmail.com, ***tejakumari555@gmail.com

Abstract: Image captioning is one of the important tasks in the field of Artificial Intelligence that leverages the advantages of both Natural Language Processing and Computer Vision techniques to generate an accurate caption to the given image. It is one of the important fields of research and a lot of research is being carried out in this field. There are a lot of advantages to this task and one of them is, it can help the people with visual impairment to understand the content of the image which is one of the primary goals of Web Accessibility. Till now this is achieved on the web by providing alternate text to an image and is manually done by a web developer. Here we automated this task by reproducing a CNN-RNN model on following the best practices from previous research in the field and created a web interface, an API and a chrome extension to make them available to end-users..

1. Introduction

Artificial Intelligence is revolutionising every field and is completely transforming the lifestyle of us. Many researchers across the globe are making things possible which were once thought to be impossible. Image Captioning is one such example. There is a lot of research taking place in this field and many state-of-the-art models were already proposed. We thought this research can help us to show the world to a visually impaired person in a better way. Some systems like Google Lens, Vivo Jovi which are widely in use and are capable of detecting different objects in the images. But there are no widely used systems which are readily available and can explain the contents of an image in a human spoken language as a caption. To understand the contents of the image and to generate captions to it in a human spoken language we developed a CNN-RNN model. This model contains two neural networks. They are Convolutional Neural Network (CNN) which is used to extract the features from an image and Recurrent Neural Network (RNN) to generate caption to the image taking the extracted features as input. On a whole, it takes images as an input and provides a caption as an output.

Web Accessibility

Accessibility is one of the design standards that are proposed by the World Wide Web Consortium (W3C). According to W3C, a website is termed to be web-accessible when it is accessible to all classes irrespective of the characteristics of a person and his possessions. It says on its website that the web is fundamentally designed for all the people irrespective of

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

the hardware, software, language, ability. The goal of accessibility is to make the web available to people with a diverse range of hearing, sight, movement and cognitive abilities.

One of the major recommendations to achieve web accessibility is every image on the website should have an equivalent alt(alternate) text. It must explain the whole scene as best as possible. Once, it has equivalent alt text then a blind person who uses the website with a screen reader can have the same experience of the normal person as the text reader reads out the alternate text in place of the image. Due to the surge of image data that is available every day, it is definitely a hard task for a web developer to add alt text to every image. It is much more difficult in case of websites like social networking sites where the user directly uploads an image.

So to automate this process we used the CNN-RNN model to generate captions to images automatically and create an API (Application Programmer Interface) to make it available to the web developers to automate the task. We also created a chrome extension to make this feature available to every website who uses it.

II The CNN - RNN model

We started working on this model by following the best practices from the previous research [1][2] which are referred to as the state-of-art. The model consists of two neural networks: Encoder CNN and Decoder RNN.

Encoder CNN:

We used the transfer learning technique to create the Encoder CNN, it is a Convolutional Neural Network created using pre-trained resnet50 architecture. We removed the last fully connected layer which was of output size 1000 and introduced a new fully-connected layer with output size equal to the embedding dimension of the Decoder RNN model. We freeze all the layers of this pretrained network except the last one we added because we want to use it only as a feature extractor. We took this step because this pretrained on showed the state of results on ImageNet dataset. We thought this works best for us as the feature extractor.
In this way, we created Encoder CNN.

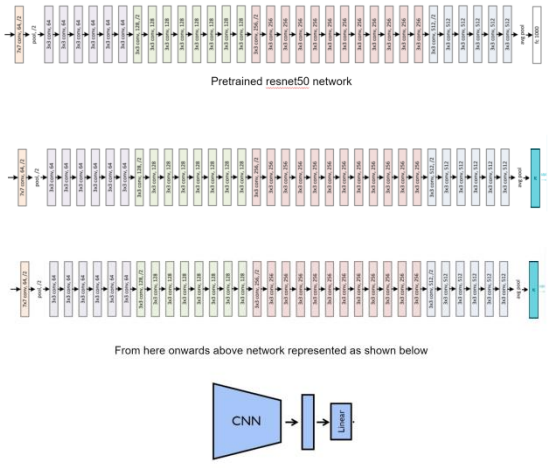

Fig 1: Encoder CNN

Decoder RNN:

We created a Recurrent Neural Network with LSTM units to use as Decoder RNN. This takes a feature vector from Encoder CNN which is of size equal to embedding dimensions. We used 512 as the embedding dimension size. This is one

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

which generates captions for the given image feature vector.

We used MSCOCO [4] dataset for training the model and used hyperparameters suggested from previous research [1][2]. Used embedding size as 512, batch size as 64, Adam as an optimizer and Cross-Entropy Loss as loss function and trained for 70 hours on a GTX 1060 GPU for 23 epochs of training. And the loss decreased as shown in the fig. 2.

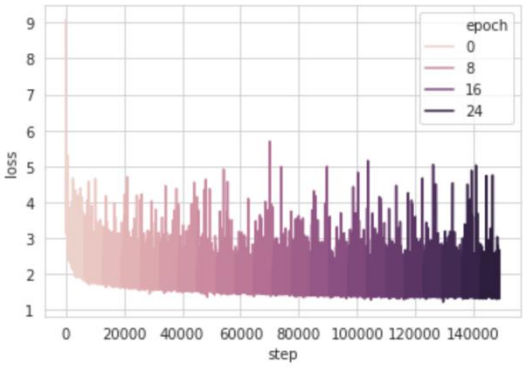

Fig 2: a graph showing losses at different steps of training.

We also calculated the perplexity at each epoch which is a measurement of how well a model predicts. The lower its value the better is its prediction. The movement of its value is shown in fig 3. Here we are mainly interested in reproducing the model from the state of art research done so far. As our primary goal is to use this research to automate the image captioning task and to increase web accessibility we didn't train for long hours to compete with the previous research.

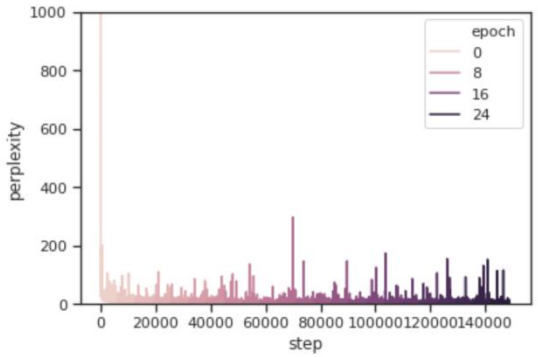

Fig 3: perplexity at different steps of training.

Image Captioning API:

To make the image captioning feature available to a large set of web developers across the globe we came up with an API which accepts the requests from different clients and sends the image caption as the response to the requested image. Fig 4 explains how this works.

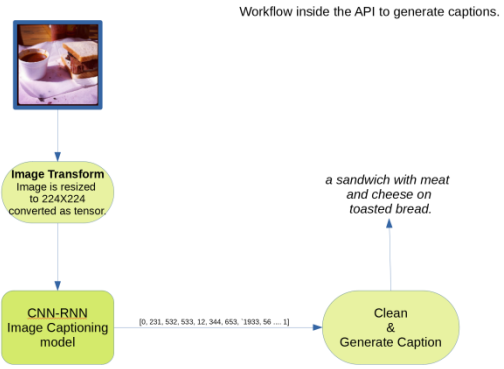

Fig 4: Workflow diagram of Image Captioning API

It accepts two types of requests. One is if the client sends the image as a request, the API processes the image and sends the caption as the response. As shown in fig 5.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

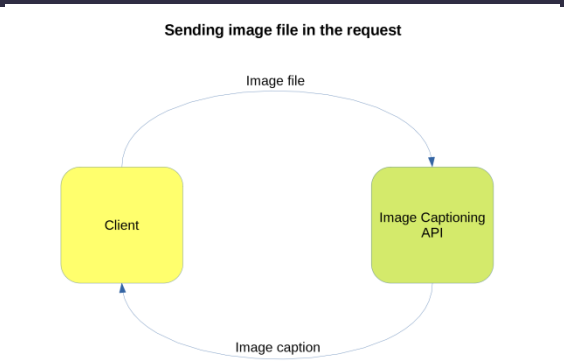

Fig 5: API accepting image as request.

On the other hand, it will also accept the image URL as the request. The main reason for handling this type of request is the surge of image data on the web. This is shown in fig 6.

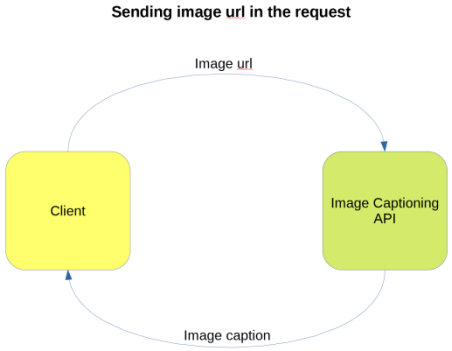

Fig 6: API accepting image URL as request

Web Interface:

Using this API we created a web interface which has an image upload option. On uploading an image this will show captions generated by the API. We showed five captions here by simply looping the process 5 times. We did this to show how it generates the captions. The screenshot of the web interface is shown in fig 7.

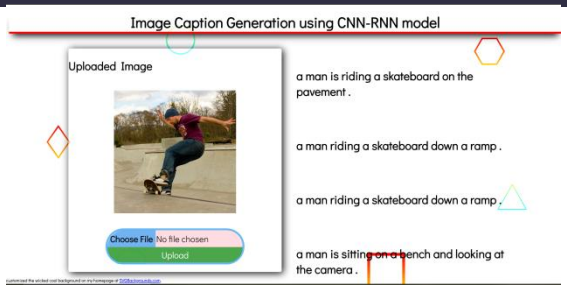

Fig 7: Web interface screenshot.

"aat: add alt text"

Chrome extension:

To make our Image Captioning API available to all the visually impaired people we created a chrome extension called "add alt text". This can be installed on chrome browser and it automates the task of adding alternate text to the image using our Image captioning API. There are many instances where a website can contain a lot of images which may not contain alternate text. Then any visually impaired person will have a bad experience of using that website. To avoid that we did this. The workflow of the extension is explained in fig 8.

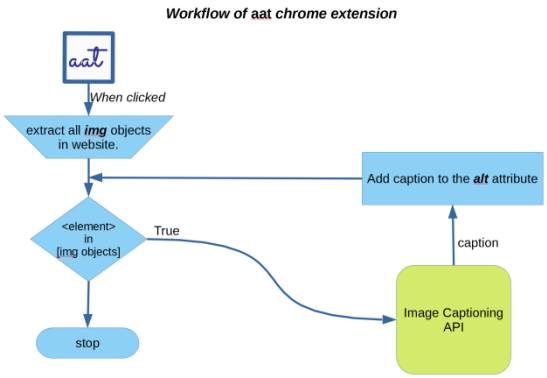

Fig 8: Workflow of "aat" chrome extension.

It extracts all the images objects in the website and searches for the images which do not have alt text for them. Or if they have an empty string as alt text. Then it selects all such images and sends their URLs as requests to our Image Captioning API. After receiving the captions from the API they will be added to their respective images. So whenever the user uses the screen reader they can hear the alt text which is added by us with our chrome extension and the experience of the user will be far better than the previous scenario.

Conclusion:

We developed a CNN-RNN model to generate caption in human spoken language to the given image. We proposed an automated way of increasing web accessibility to address the large amount of image data that is being created every day. We proposed different methods to achieve this like our Image Captioning API, chrome extension and can easily be implemented in other platforms. With all these things, we proposed a way to increase the accessibility on the web and to make the web accessible to more people on the globe.

References:

1. Oriol Vinyals and Alexander Toshev and Samy Bengio and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. *arXiv: 1411.4555*, 2014.

2. Kelvin Xu and Jimmy Ba and Ryan Kiros and Kyunghyun Cho and Aaron Courville and Ruslan Salakhutdinov and Richard Zemel and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv: 1502.03044*, 2015.

3. Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385*, 2015.
4. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Com- ´ mon objects in context. arXiv:1405.0312, 2014.
5. WEB ACCESSIBILITY reference: https://www.w3.org/standards/webdesign/accessibility