

Image Captioning: Transcribing Image into Words

Adesh Vaidya¹ Prof. A. Nachankar²

¹Student ²Professor

^{1,2}Department of Computer Technology

^{1,2}KDK College of Engineering, Nagpur, India

Abstract— In the past few years, generating descriptive sentences by the machine of any image data gained an immense curiosity in computer vision and natural language processing research. Captioning the image is a basic job which requires the understanding of image data and potential to generate descriptive sentence with the correct structure. Image captioning models follow the conventional encoder-decoder architecture which uses the features of the image data as input and generates transcription. Image captioning needs to identify the important entities, their attributes, and their connection in an image data. It is also primely important that the model can make the sentence semantically and systematically. The deep learning-based model is proficient to manage these complications and challenges of image captioning.

Keywords: Image Caption Generation, LSTM, Deep Learning, Natural Language Processing, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Text Generation, Transcript Generation

I. INTRODUCTION

Generating caption is a fascinating problem of artificial intelligence. Describing automatically what exactly happens in the image is a challenging task where image data is converted into descriptive sentences. It includes some techniques which involve computer vision for understanding what exactly happens in the graphical data and the natural language processing technique is used to generate the contextual data in a correct format. Image captioning has a large variety of applications such as providing recommendations in image editing software, implementing this model in software that can help people who have low eyesight or no eyesight, in various social media platforms, and more other natural language processing applications. Here we are going to use the model names InceptionV3 which is a pre-trained model. This model is pre-trained on the very famous 'Imagenet' dataset. This model is trained on over 1000 different classes and has 93.9% of accuracy.

The neural network restrictions are decided mainly by what amount of GPU memory is available and what amount of time duration we have to train the model. The result given by model may be improved by providing the faster and haviour GPU.

II. WORKING MECHANISM

There are several steps are required to build an image caption generator. Here I am going to use the deep learning approach so it requires a dataset. I will use the MS-COCO dataset to build my model. The MS-COCO dataset contains has about 85,000 image data where each image data has 5 different annotations. The size of this dataset is about 13 gigabytes. But we don't have a powerful processor to process this large amount of data. So, to speed up the training process I am

going to limit the size of the data for training our model to 30,000 annotations for 20,000 image data.

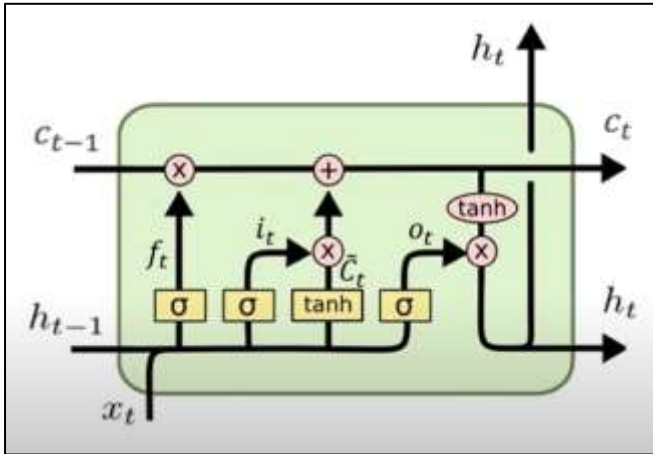
Next, we need to pre-process all the images for that I am going to use the InceptionV3 model which is a pre-trained model on a famous dataset called 'Imagenet'. The InceptionV3 model is used to classify every image. To use this pre-trained model, I need to implement another approach that is 'Transfer Learning'. In this approach, we need to replace its last layer with our own convolutional layer to extract the image features. Firstly, we need to convert the image data which is accepted by the InceptionV3 model. Here we going to resize the image data into 299 pixels by 299 pixels.

Now we need to work on annotation data. We will tokenize the data. Tokenization means spiting the sentences, phrases, or paragraphs into smaller entities. Every smaller entity is called a token. In tokenization of the data, we will limit the size of vocabulary to the first 5000 entities to save memory and time and replace all other words with the token name "Unknown". Now we need to create 'index_to_word' and 'word_to_index' mappings. These are two methods to give numbering to each token. Finally, we resize all the tokens to the same size by using 'pad_sequences'. All tokens should be the same size as the longest token.

The further step is splitting our dataset into training and validation section randomly. The training sections take 80% of the data where the remaining 20% taken by the validation section. Now we make a model using convolutional and dense layers and train our data on it.

The key factor of our model is we are not using the recurrent neural network (RNN) for text generation because RNN can predict the word when the gap between the prediction and the necessary context is very small e.g. "The grass is green". Here the gap between the predicted word and the necessary context which is 'grass' is too small so RNN can predict the word 'green' easefully. But if we take another example like "I live in Mumbai in west India, where the weather is generally humid". In this example, the gap between the prediction and the necessary context is more than the previous one, so here the correct prediction with RNN is very less.

To overcome this drawback, we are using the 'Long Short-term Memory (LSTM)' network. This network holds one memory line to carry previously gathered data. LSTM contains gates that can allow or block the data from the previous cell. These gates incorporate a sigmoid neural layer along with a dot multiplication operation. The sigmoid is an activation function and it ranges from 0 to 1. Where 0 doesn't allow data to flow and 1 allows everything to flow.



Long Short-term Memory (LSTM)

Colah, Christopher. *Understanding LSTM Networks* -- Colah's Blog. 27 Aug. 2015, colah.github.io/posts/2015-08-Understanding-LSTMs/.

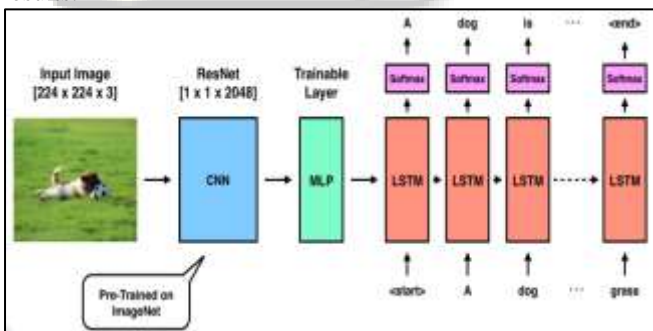
Equations are,

$$\begin{aligned}
 f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
 i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
 \tilde{c}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
 o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
 c_t &= f_t * c_{t-1} + i_t * \tilde{c}_t \\
 h_t &= o_t * \tanh(c_t)
 \end{aligned}$$

Where

- W = weights in the network
- b = bias
- σ = activation function
- tanh = activation

Following the training, we predict the accuracy as well as the loss of the model. We can also try to generate our own data but keep in mind that the model is trained in a relatively small amount of dataset so it can give some weird results.

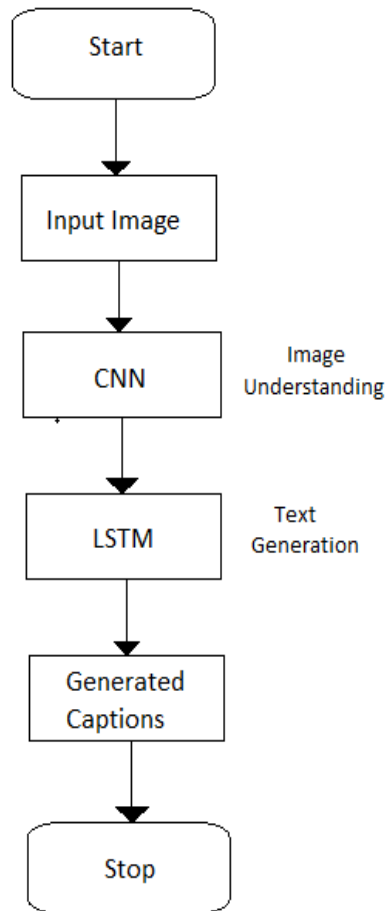


Neural Network For Image Captioning

Pesch, R., & Dogonadze, N. (2020, April 30). *End-to-End Image Captioning [Image]*.

<https://www.inovex.de/Blog/End-to-End-Image-Captioning/>. https://www.inovex.de/blog/wp-content/uploads/2020/04/encoder_decoder.png

III. FLOWCHART



IV. RELATED WORK

The problem of generating transcription from image data is studied from past in computer vision mainly for videos. The automatic image caption generation is the method where, a computer automatically generates the captions or keywords from the image data. There are a lot of approaches aimed towards generating transcription with the use of different templates in which way attributes, blank slots, and actions are detected first and then blank spaces in the templates are filled.

For example, Farhadi et al. use a trinity of arena components to infuse the template for generating caption data. Correspondingly Li et al. pull out the wordings related to predicted objects, features, and their bonding for this purpose. Kulkarni et al. accept a conditional random field (CRF) method to conclude the objects, features, and relationship before padding in the voids. In this work we integrate the deep convolutional networks for image data classification with recurrent neural networks for sequence modelling, to generate a single neural network that create transcriptions of image data. The recurrent neural network is trained in the surroundings of the one "end-to-end" neural network. The model is motivated by modern triumph of sequence generation in machine learning, with the dispute that rather of starting with a sentence, we provide an image data prepared by a convolutional neural network.

V. CONCLUSION

In the advanced project, an image transcription generator has been bloomed using a CNN-RNN architectural model. Some key features about our model that our model is based on the image data and some annotation, so, it is unable to predict and generate the transcriptions that are out of its vocabulary. A dataset consisting of 20,000 images and 30,000 annotations are used here. But for large-scale models i.e., higher accuracy and low loss models, we need to train the model on more than 100,000 image and annotation datasets so that better accuracy and lower the loss can be developed by the model. As lower the loss and higher the accuracy the model can generate transcription more effectively.

VI. FUTURE SCOPE

- 1) Classify and sort your photos into different categories
- 2) Recommending the required editing in photo editor software.
- 3) Can be used in virtual assistance for better guidance.
- 4) Visually impaired or completely blind people can use this application to get the guidance to walking on streets.
- 5) Social media platforms like instagram and facebook can conclude straightly from the image that what actually happened, what type of colors in the image and where the image is taken from.

REFERENCES

- [1] Xiujun Li, Xi Yin , Chunyuan Li , Pengchuan Zhang, Xiaowei Hu , Lei Zhang , Lijuan Wang , Houdong Hu , Li Dong , Furu Wei , Yejin Choi , and Jianfeng Gao, "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks", July 2020.
- [2] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara, "Meshed-Memory Transformer for Image Captioning" March 2020.
- [3] Luowei Zhou , Hamid Palangi , Lei Zhang , Houdong Hu , Jason J. Corso and Jianfeng Gao, "Unified Vision-Language Pre-Training for Image Captioning and VQA", December 2019.
- [4] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi and Jianfeng Gao, "VinVL: Making Visual Representations Matter in Vision-Language Models", January 2021.
- [5] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari, "Connecting Vision and Language with Localized Narratives", July 2020.
- [6] Grishma Sharma, Priyanka Kalena, Nishi Malde, Aromal Nair and Saurabh Parkar, "Visual Image Caption Generator Using Deep Learning", 2019.
- [7] Sailee P. Pawaskar and J. A. Laxminarayana, "Image Caption Generation A Comprehensive Survey", March 2018.