ZKG INTERNATIONAL

ISSN: 2366-1313

LSTM-based Tensor-flow mechanism for generating captions for the images

[1]Anumalla Niharika, [2] Dr.T. Venugopal

[1]M. Tech Scholar, [2]Professor, Department of CSE and Vice Principal,

JNTUH UNIVERSITY COLLEGE OF ENGINEERING, JAGTIAL, T.S., INDIA

Abstract: -

Image captioning is the process of generating descriptions about what is going on in the image. By the help of Image Captioning descriptions are built which explain about the images. Image Captioning is basically very much useful in many applications like analyzing large amounts of unlabeled images and finding hidden patterns for Machine Learning Applications for guiding Self-driving cars and for building software that guides blind people. This Image Captioning can be done by using Deep Learning Models. With the advancement of deep learning and Natural Language Processing now it has become easy to generate captions for the given images. In this paper we will be using Neural Networks for the image captioning. Convolution Neural Network (CNN) is used as encoder which access the image features and Recurrent Neural Network (Long Short-Term Memory) is used as decoder which generates the captions for the images with the help of image features and vocabulary that is built.

Keywords: - Deep Learning; Image Captioning; Convolutional Neural Networks; Recurrent Neural Networks; CNN; Long Short-Term Memory; insert (key words).

1. INTRODUCTION:

Image captioning models typically follow an encoder-decoder architecture which uses abstract image feature vectors as input to the encoder and generates a caption. Generating a natural language description from images is an important problem at the section of computer vision, natural language processing, artificial intelligence and image processing. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. The application of image caption is extensive and significant, for example, the realization of human-computer interaction. This summarizes the related methods and focuses on the attention mechanism, which plays an important role in computer vision and is recently widely used in image caption generation tasks. Furthermore, this project model highlights some open challenges in the image caption task.

Photo captions aim to describe objects, actions, and details found in an image using natural language. Most image caption research focuses on single-sentence captions, but the descriptive capabilities of this form are limited; one sentence can only describe in detail a small part of an image. Recent work has been challenged instead of captions for the role of the image for the purpose of reproduction (usually sentence 5-8) describing the image. Compared to single-sentence captions, section captions are a relatively new task. The caption data set for the main role is Visual Genome corpus, presented by Krause et al. (2016). When solid single-sentence caption models are trained in this database, they produce repetitive sections that can explain various aspects of the images.

The generated sections repeat the slightest variation of the same sentence many times, even when beam search is used. Similarly, the different methods used for classification, namely: Long Dial Network Repetition: Input can be an image or image sequence in a video frame. captions [4]. Visual Paragraph Generation: This method is meant to provide a coherent and detailed category. Few semantic regions are acquired in the image using the attention model and sentences are generated sequentially and phase is generated [14]. RNN: Continuous neural network is a special neural network for processing data sequences with a timestamp. index t from 1 to t. In activities that include sequential inputs, such as speech and language, it is usually best to use RNNs. before it. GRU: Repetitive unit with the latest development

ISSN: 2366-1313

gateway proposed by Cho et al. Similar to the LSTM unit, GRU has gate-gate units that model the flow of information within a unit, however, without having separate memory cells.

The Gated Recurrent Unit (GRU) lists two gates called renewal and reset gates that control the flow of information for each hidden unit. Each hidden state during step t is calculated using the following calculations: Update gate formula, Reset gate formula, new memory formula, and final memory formula. This gate controls how many parts of the new memory and old memory should be integrated into the final memory. Similarly, the reset gate is calculated but with a different set of weights. Controls the balance between previous memory and new input information for new memory.

2 RELATED WORKS:

Image captioning means automatically generating a caption for an image. As a recently emerged research area, it is attracting more and more attention. To achieve the goal of image captioning, semantic information of images needs to be captured and expressed in natural languages. Connecting both research communities of computer vision and natural language processing, image captioning is a quite challenging task. Various approaches have been proposed to solve this problem. -e number of digital images increases rapidly; hence, categorizing these images and retrieving the relevant web images are a difficult process. For people to use numerous images effectively on the web, technologies must be able to explain image contents and must be capable of searching for data that users need. Moreover, images must be described with natural sentences based not only on the names of objects contained in an image but also on their mutual relations.

Paper [20] uses an annotation mechanism to overcome the problem of images. Here, two mechanisms are followed; they are manually annotating the images by using the human interface, and the annotated images are stored in the repository. Automatic annotation: they are obtained by performing feature extraction and clustering algorithm. For feature extraction, SIFT algorithm is used. Our method for feature extraction is different; we used pretrained CNN models. -e last mechanism is learning annotations by clustering.

In the paper [21], the authors give a comprehensive overview of the automatic caption generation for medical images covering existing models, the benchmark medical image caption datasets, and evaluation metrics that have been used to measure the quality of the generated captions

Also, paper [22] is concerned with the task of automatically generating captions for images, with concrete implementations for many image-related applications. Apart from images, they also used video retrieval as well as the development of tools that aid visually impaired individuals to access pictorial information. -is approach leverages the vast resource of pictures available on the web and the fact that many of them are captioned and collocated with thematically related documents. Authors approximate content selection with a probabilistic image annotation model that suggests keywords for an image. -e model postulates that images and their textual descriptions are generated by a shared set of latent variables (topics) and are trained on a labeled dataset (which treats the captions and associated news articles as image labels). Experimental results show that it is viable to generate captions that are pertinent to the specific content of an image and its associated article while permitting creativity in the

description.

In the paper [23], Ding et al. introduced the theory of attention in psychology to image caption generation. -ey used two approaches: stimulus-driven, where an object detection model is used to identify objects belonging to certain classes and localize them with bounding boxes, and concept-driven, where the visual question answering (VQA) model implements a joint embedding of the input questions and images and

ISSN: 2366-1313

then projects them into a common semantic space. -ey used a different approach from the one proposed in our paper, but since we were working on the same dataset,

3 PROPOSED SYSTEM

Dataset

The data set is a collection of 10000 images with five captions each, collected in one place, and available to be used for the benchmarking of image captioning and im- age querying approaches [11]. The authors show that better results can be achieved when multiple captions are used with each image, to train the model. A manual data set of 2000imageswas created with relevant 50000 captions is was used to provide final results of a model

Figure 1 is a sample image file in dataset. The image is paired with following five human-generated training captions: • Sunsets and oceans. It's what I do. • A mind-boggling, awe-inspiring, spine-tingling sunset. • A sunset that good doesn't need a filter. • Watch the sunset. Not Netflix. • Here comes the sunset. • Data Preprocessing

We divide the training data (10000) and the captions into three different data sets - the training set (8000), the validation set (1000) and the test set (1000). For each of the captions in the three data sets, we create a set of training input and target captions by shifting the training input caption by one word to get the training target caption.

Image Preprocessing

To generate image features we use pretrained weights of CNNs trained on ImageNet image classification dataset (VGG16, VGG19, and CNN) and remove the final dense layers from the model. We preprocess images and generate image features using the by performing a forward pass on the image on using these weights and save these features to a file. Caption Preprocessing.

To preprocess the image captions in the training data, we first identify all the words that are there in the data set. We then generate a histogram of the distribution of these words and drop the words that occur less than five times. We end up with a vocabulary of size 2531 words.

The model that was used for the project consists of two different input streams, one for the image features, and the other for the preprocessed input captions. The image features are passed through a fully connected (dense) layer to get a representation in a different dimension. The input captions are passed through an embedding layer. These two input streams are then merged and passed as inputs to an LSTM layer. The image is passed as the initial state to the LSTM while the caption embeddings are passed as the input to the LSTM. The architecture is shown in figure 2. Training

The model was trained first on Nvidia GeForce MX 250. We faced memory problems using different batch sizes and hence moved to a better GPU. We then used the Alienware R17 which includes intel i9 processor with 6 core, 32GBRAMandNvidia GTX1080 8GB. Training the model takes about 6 hours

To train the model, for each image and each of the input captions that were generated during preprocessing, we pass the image features through the dense layer, and the pre-processed input captions to the embedding layer. We then use the image the initial state to the LSTM, along with the caption which is passed as the input to the LSTM.

Word Embedding

Word Embeddings provides a vector representation of words that can capture something about the context of the word. There are many pretrained word embeddings available; however, our model learns the word embedding as part of the model itself.

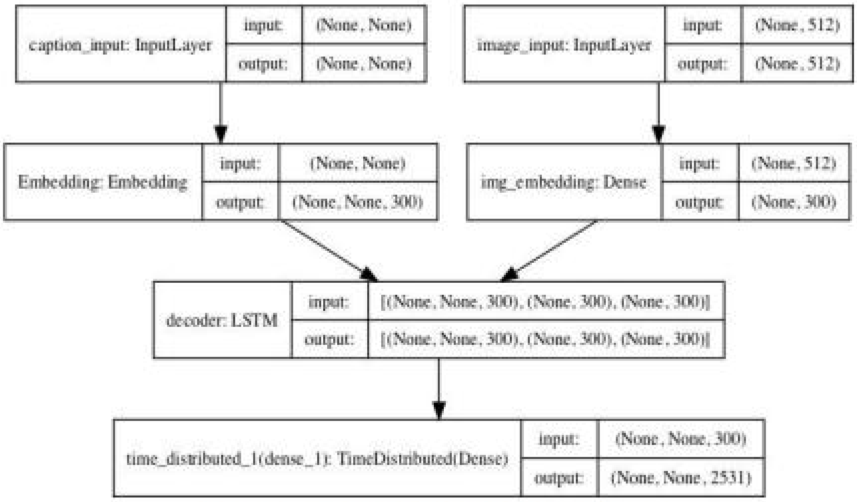


Figure 2: Model Architecture

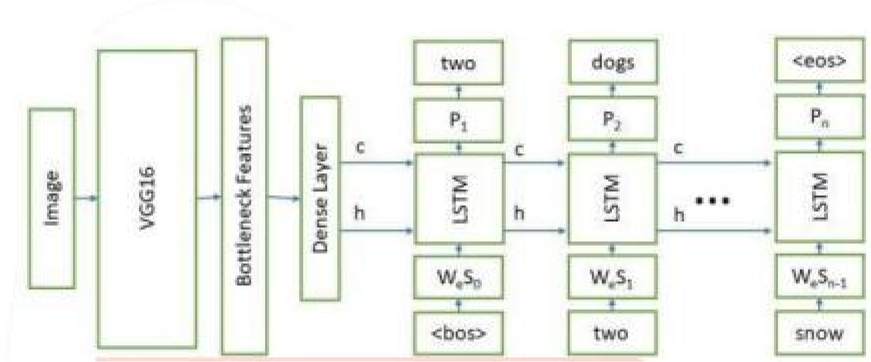


Figure 1: Training the model using the LSTM model with CNN image features

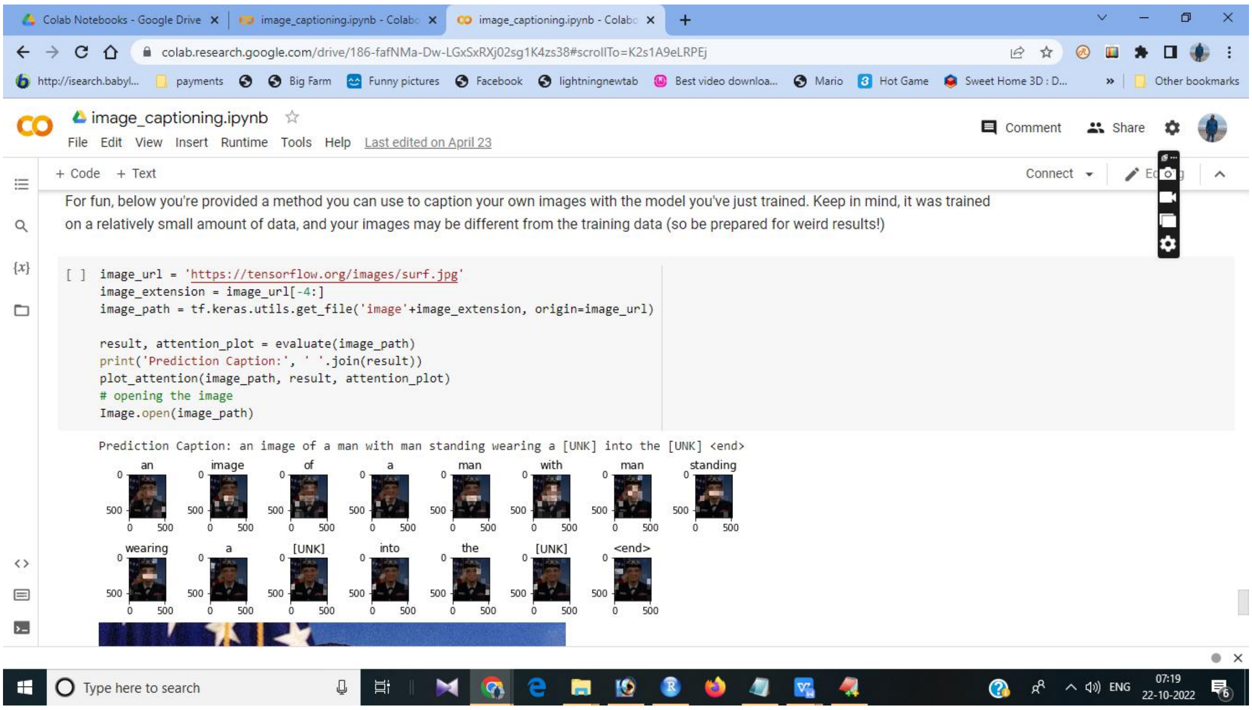
4. EXPERIMENTS

The model was trained 3 times for each of the CNNs models that we used. First, we trained the model using the learning rate of 0.0001 for 10 epochs and used a greedy approach to generate captions. The LSTM scores for this set of hyperparameters are shown in table 1. Next, we decreased the learning rate to 0.000051, trained the model to 33 epochs and used beam search to generate sentences. The LSTM scores for this set of hyper-parameters are given in table 2. We can see that there is a significant improvement in the LSTM scores.

Finally, we increased the size of the embedding layer and the dense layer from 300 to 512, increased the LSTMsizefrom300 to 512 and trained the model again for 33 epochs. The LSTM scores for these hyper parameters. We can see that increasing the LSTM size and the size of the embedding layer lead to even better results, even though it took a significantly longer time to train the models with these hyper parameters. We cleaned up our code and created two files that use argument parser to specify parameters for training and the model and generating captions using trained weights. We also created a simple web application in python with the help of flask that allows users to upload an image and uses the trained weights to generate image captions.

5. RESULTS

The table 3 shows the bleau score is highest with the model that uses CNN to generate image features. The results shown here are for the CNN model using a LSTM size of 512 and a embedding layer and a dense layer size of 512.



For fun, below you're provided a method you can use to caption your own images with the model you've just trained. Keep in mind, it was trained on a relatively small amount of data, and your images may be different from the training data (so be prepared for weird results!)

```
image_url = 'https://tensorflow.org/images/surf.jpg'
image_extension = image_url[-4:]
image_path = tf.keras.utils.get_file('image'+image_extension, origin=image_url)

result, attention_plot = evaluate(image_path)
print('Prediction Caption:', ' '.join(result))
plot_attention(image_path, result, attention_plot)
# opening the image
Image.open(image_path)
```

```
Prediction Caption: an image of a man with man standing wearing a [UNK] into the [UNK] <end>
```

Figure 2: Model displaying the text about the image.

The images in the results show captions, generated. The sentence corresponds to the best caption according to the beam search.

6 CONCLUSIONS

Through this project, we learned about the deep learning techniques used for image captioning problem. We experimented with three distinct CNN models and compared the results of different models using LSTM scores, by comparing we came to a conclusion that CNN was the most suitable and efficient model for us. It captured more than enough information about the image to generate captions. We learned that the result of generated captions is influenced by the training dataset. The Flickr8k and manual data set contains many outdoor images of Nature, Beaches, and sunsets and our model gives better results on outdoor images without people and is capable of differentiating between various natural objects. We implemented beam search and found that the LSTM scores for sentences generated using beam search are significantly better Future Work.

7 FUTURE WORK

In this project, we implemented a model and experimented with different parameters to see the results. Future exploration can be done to compare the current results to those obtained with using different CNN models. Further comparisons can be made to approaches that include visual and temporal attention. We also intend to create an Android app for the users.

7 REFERENCES:

[1] Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra and Nand Kumar Bansode (2017) International Conference on Computational Intelligence in Data Science (ICCIDS) : Camera2Caption: A Real-Time Image Caption Generator

[2] Chaudhuri, D., Samal, A., An automatic bridge detection technique for multispectral images. IEEE Trans. Geosci. Remote Sens. 2008

[3] Hu, J., Razdan, A., Femiani, J.C., Cui, M., Wonka, P., Road network extraction and intersection detection from aerial images by tracking road footprints. IEEE Trans. Geosci. Remote Sens., 2007

[4] Goodin, D.G., Anibas, K.L., Bezymennyi, M., 2015. Mapping land cover and land use from object-based classification: an example from a complex agricultural landscape. Int. J. Remote Sens. 36, 4702-4723.

[5] Han, J., Zhang, D., Cheng, G., Guo, L., Ren, J., Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Trans. Geosci. Remote Sens. 2015.

[6] Wang, J., Song, J., Chen, M., Yang, Z., Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. Int. J. Remote Sens.2015

[7] Duarte, D., et al. "Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach." ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 4.2 (2018).

[8] Zhong, Y.; Fei, F.; Liu, Y.; Zhao, B.; Jiao, H.; Zhang, L. SatCNN: Satellite image dataset classification using agile convolutional neural networks. Remote Sens. Lett. 2016, 2, 136–145.

[9] Arshitha Femin and Biju K.S., “Accurate Detection of Buildings from Satellite Images using CNN” IEEE, 2020.

[10] G. Scott, M. England, W. Starms, R. Marcum, and C. Davis, “Training deep convolutional neural networks for land–cover classification of high-resolution imagery”, IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 4, 2017.

[11] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, “Transferring deep convolutional neural networks for the scene classification of highresolution remote sensing imagery,” Remote Sens., vol. 7, no. 11, pp. 14 680–14 707, 2015.

[12] D. A. Neu, J. Lahann, and P. Fettke, “A systematic literature review on state-of-the-art deep learning methods for process prediction,” Artificial Intelligence Review, vol. 55, pp. 1–27, 2021.

[13] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, “Deep learning approaches applied to remote sensing datasets for road extraction: a state-of-the-art review,” Remote Sensing, vol. 12, no. 9, p. 1444, 2020.

[14] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, “Deep learning in multi-object detection and tracking: state of the art,” Applied Intelligence, vol. 51, no. 9, pp. 6400–6429, 2021, https://doi.org/10.1007/s10489-021-02293-7.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, “Explain Images with Multimodal Recurrent Neural Networks,” 2014, https://arxiv.org/abs/1410.1090.

[16] Y. You, C. Lu, W. Wang, and C. K. Tang, “Relative CNNRNN: learning relative atmospheric visibility from images,” IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 45–55, 2018.

[17] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, “Cnn-rnn: a unified framework for multi-label image classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294, San Juan, PR, USA, June 2016.

[18] Y. Guo, Y. Liu, E. M. Bakker, Y. Guo, and M. S. Lew, “CNNRNN: a large-scale hierarchical image classification framework,” Multimedia Tools and Applications, vol. 77, no. 8, pp. 10251–10271, 2018.

[19] S. M. Xi and Y. Im Cho, “Image caption automatic generation method based on weighted feature,” in Proceedings of the 2013 13th International Conference on Control, Automation and Systems (ICCAS 2013), pp. 548–551, IEEE, Gwangju, Korea (South), October 2013.

[20] A. S. Reddy, N. Monolisa, M. Nathiya, and D. Anjugam, “Automatic caption generation for annotated images by using clustering algorithm,” in Proceedings of the International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1–5, IEEE, India, March 2015.

[21] I. Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz, “Automatic caption generation for medical images,” in Proceedings of the 3rd International Conference on Smart City Applications (SCA’18), 3rd International Conference on Smart