Image Captioning using Keras

Prof.Kaustubh Hiwarkar
BRACT VIT, Pune, India

Rohan Ranjan - 11910134
Dept.Of Computer Engineering ,
BRACT VIT, Pune, India

Varun Ringnekar - 11910590
Dept.Of Computer Engineering,
BRACT VIT, Pune, India

Himangi Rinwa - 11910386
Dept.Of Computer Engineering,
BRACT VIT, Pune, India

Rohit Singh - 11910771
Dept.Of Computer Engineering,
BRACT VIT, Pune, India

Abstract

Recent advancement in artificial intelligence (AI) has significantly progressed the execution of models. However, the results are still not adequately fulfilling. Machines cannot mimic human brains and the way they communicate, so it remains an ongoing task. Due to the increasing sum of data on this subject, it is exceptionally difficult to keep on track with the most current research and comes about accomplished in the image captioning field. We have developed an image captioning model that can generate suitable captions based on the image provided.

Keywords: Image Captioning

I. INTRODUCTION

Within the past few years, computer vision in the image processing zone has made significant progress, like classification of images [1] and detecting an image object [2]. The processing of images has grown much faster and more effective in cost and its uses. Also, Time consumed is quite less. Caption generation comes under AI standards where a printed description must be generated for a given photograph. It requires both strategies from computer vision to understand the substance of the picture and a language demonstrated from the field of normal dialect preparing to turn the understanding of the picture into words within the right arrangement. As of late, deep learning strategies have accomplished state-of-the-art results on examples of this problem.

Image Captioning is a well-known zone of Artificial Intelligence that deals with picture understanding and a language is described for that picture. Generating sentences related to the image gives us an understanding of the same. It's quite challenging depicting the insights of such an image, but it seems moreover to have a great effect, it has a great impact on visually challenged people who like this application would make them a lot easier to understand certain images.

This assignment is altogether harder in comparison to the picture classification or object recognition tasks that have been well researched. The greatest challenge is most certainly being able to make a portrayal that must capture not only the objects contained in a picture but also it will be explained what is the connection between them or how they are related to the surroundings of the same image.

A)Keras

Keras is an Open Source Neural Network library composed in Python that runs on Theano or the so-called TensorFlow. It is planned to be measured, quick, and simple to utilize. Theano is a library in python which is utilized for faster computing uses mostly numerically. TensorFlow is the foremost typical math library utilized for making neural networks and profound learning models. TensorFlow is exceptionally adaptable and the essential advantage is distributed computing. Keras is outlined to rapidly characterize deep learning models. Keras is an ideal choice for deep learning applications and has been proved effective in dealing with them.

B) TensorFlow

TensorFlow is a library created in python for quick numerical computing created and released by Google. It is an established library that can be used to make Deep Learning models straightforwardly or by utilizing wrapper libraries that rearrange the method built on the beat of TensorFlow.

II. Literature Review

Image captioning is defined as the process of creating captions or descriptions for an image based on the features of that image. This problem of image captioning and solutions to this have existed since the birth of machine learning. Researchers from different backgrounds have put forward various methods and techniques. Krizhevsky created a neural network model using GPU implementation of the convolution function, reducing overfitting using dropout. Their network consisted of a max-pooling layer and a 1000-way softmax. Deng introduced a new ImageNet database. This database consists of an extensive collection of images based on the core of the WordNet structure. Karpathy and FeiFei used the database of images and sentence descriptions to learn about the correlation between visual data and language. Their work described a Multimodal Recurrent Neural Network architecture that utilizes the inferred co-linear arrangement of features to learn how to generate novel descriptions of images. Yang came up with a system that automatically generated natural language which will bring the predictions as close to human predictions as possible. To address the problem of LSTM units being complex and inherently sequential across time, Aneja proposed a convolutional network model for machine translation and conditional image generation. Pan experimented with different network architectures, by implementing them on large datasets consisting of various styles of content. Doing so he was able to come up with a unique model with improvement on captioning accuracy over previous models. Vinyals used a deep recurrent network to create a generative model that uses machine translation and computer vision and was able to generate natural descriptions of an image while making sure that it was an accurate description of the image. Xu proposed an attention-based model that described the image regions automatically. The model was trained through standard backpropagation techniques. The model was able to generate descriptive sentences while being able to automatically identify object boundaries.

III.PROJECT METHODOLOGY

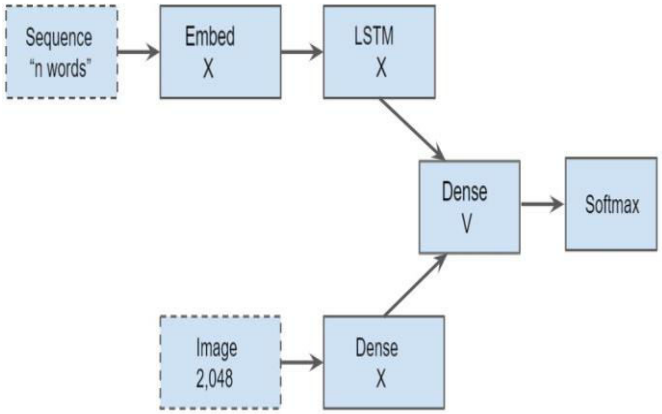
Data Description

Several datasets are utilized for preparing, testing, and assessment of the picture captioning methods. These datasets vary in different viewpoints such as the number of pictures, the number of captions per picture, organizing the captions, and picture measure. Three datasets: Flickr8k, Flickr30k, and MS COCO Dataset are prevalently used. In the Flickr8k dataset, each picture is associated with five diverse captions that portray the entities and occasions depicted within the picture that were collected. By associating each picture with different, independently delivered sentences, the dataset captures some of the phonetic variety that can be utilized to describe the same image. Flickr8k is a good dataset for the start as it is small in the estimate and can be trained effortlessly even on low-end laptops/desktops using a CPU. For training, features had to be extracted from both images and text, and then the model will be trained on them. For extracting the image features, the ImageNet model was considered the first choice but because of the deep nature of ImageNet and huge dataset extracting features consumed a lot of time and processing power. So we went on the image features that were provided in the dataset which is quite similar to what we would have received from ImageNet.

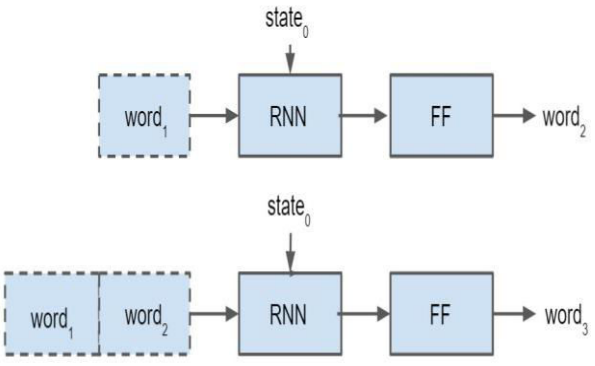
For text features, we created unique tokens for each word and let the embedding layer create the word embeddings on its own.
Since we already have the image features there wasn't any preprocessing required for images. For text preprocessing, we removed all the punctuations as they would make up extra unwanted space in vocabulary. We converted all the descriptions to lowercase and removed single-lettered words. Start and end tokens were encoded to each description to mark the beginning and

end of sentences during the prediction

For training data was converted in sequences where at first timestamp only first token is passed and based on that next token is predicted, then this predicted token is appended with the first token and both are passed in the second timestamp and the third token is predicted and so on.

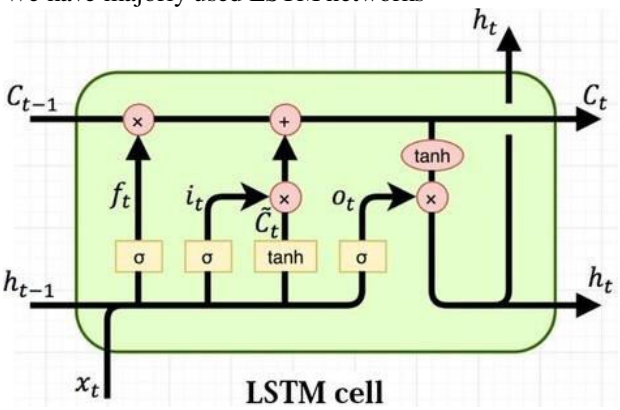


The model has been created in two parts and then merged for a final prediction.
The first part deals with the images which process the image features and passes the weights of the Dense layer.
The second part deals with the text descriptions which process the descriptions create the embeddings and passes weights to the further layers.
The common layer combines the weights from the image part and text part and provides them further to the Dense layer that predicts the probability of the words in the vocabulary.



While predicting captions, image features are extracted, and the model is given a start token and it predicts the first word and then that word is passed along with start topredict the second word and so on.

We have majorly used LSTM networks



Forget gate : af = Wf . Zt + bf ft = sigmoid(af)
Input gate : ai = Wi . Zt + bi it = sigmoid(ai)
: ac = Wc . Zt + bc ct = tanh(ac)
Output gate : ao = Wo . Zt + bo ot = sigmoid(ao)

IV. CONCLUSION

In this advanced Python project, an image caption generator has been developed using the RNN model. The developed model generates a reasonable description in plain English.
The project is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence for describing the image. As the size of the available datasets for image description increases, the performance of the model also increases.
The training model depends on the data, so it cannot predict the words that are out of its vocabulary. And so, for this reason, the flickr8k dataset is used which consists of 8000 images.
But for production-level models i.e. higher accuracy models, we need to train the model on larger than 100,000 images datasets so that better accuracy models can be developed.

V. REFERENCES

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and Top-down which is attention for image captioning and visual question answering. In Proceedings of the IEEE Conference about Computer Vision and also Pattern Recognition, 2018.

[2] M. Denkowski and A. Lavie. Meteor universal: Language-specific translation evaluation for any kind of target language. In Proceedings of the ninth workshop on statistical machine translation, pages 376–380, 2014.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. 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.

[5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and also Pattern Recognition, pages 1473–1482, 2015.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 580–587, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference about Computer Vision and Pattern Recognition, pages 770–778, 2016.

[8] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3588–3597, 2018.

[9] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4565–4574, 2016.

[10] A. Karpathy and L. 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.

[11] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In International Conference on Machine Learning, pages 595–603, 2014.

[12] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out, 2004.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[14] D. Liu, Z.-J. Zha, H. Zhang, Y. Zhang, and F. Wu. Context-aware visual policy network for sequence-level image captioning. In Proceedings of the 26th ACM International Conference on Multimedia, MM ’18, pages 1416–1424. ACM, 2018.

[15] R. Luo. An image captioning codebase in PyTorch.https://github.com/ruotianluo/ ImageCaptioning.pytorch, 2017.

[16] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint arXiv:1412.6632, 2014.