IMAGE CAPTIONING GENERATOR

K. SANDHYA, S. PRAVALIKA, D. ANKITHA

Mr. **Devavarapu Sreenivasa Rao (Guide)**

Sreenidhi Institute of Science and Technology

ABSTRACT

Our brain is able to annotate and label whenever we see an image in front of us. But what about computers How they are able to label the images. It's quite impossible for a computer to label them. But with the enhancement of computer vision and new algorithms, it becomes easier for generating a label for the image. Image Caption Generator is a popular research area of Deep Learning that deals with image understanding and a language description for that image. Generating well-formed sentences requires both syntactic and semantic understanding of the language. Image captioning becomes a vital role by automatically labelling all amount of data in a very short period of time. We will generate a image caption generator by using CNN and LSTM algorithms.

INTRODUCTION

Image Captioning is the process of describing the image in textual format. It tries to display the contents of an image in words. This task lies at the intersection of computer vision and natural language processing. Image captioning is one of the in-demand issue in the field of artificial intelligence, It has a wide range of application scenarios, It can be used in human-computer interaction, adding subtitles to video, video question answering , search important information according to image content and image search by keywords, etc.

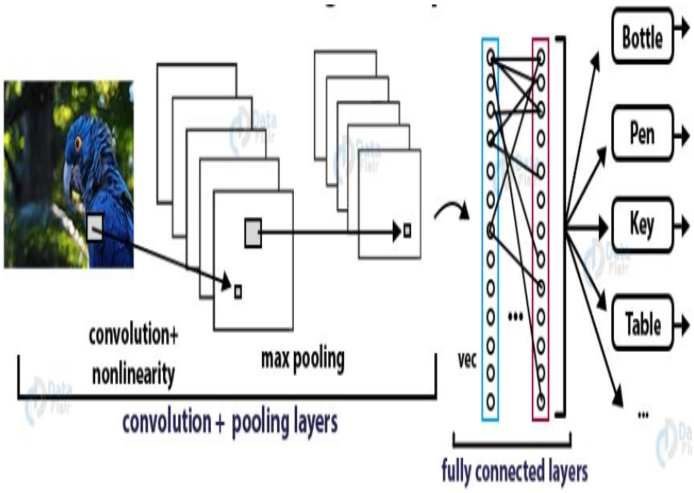
Automatically describing the content of an image is an interesting and challenging task in artificial intelligence. In this paper, an enhanced image captioning model—including image captioning—is proposed to automatically generate the textual descriptions of images. we have to pass the image to the model and the model does some processing and generating captions or descriptions as per its training. we will have implemented the caption generator using CNN (Convolutional Neural Networks) and LSTM a comprehensive human-like description makes a better first impression. Natural language descriptions will remain a difficult problem to address as long as machines do not think, talk, or behave like humans. Image captioning is used in a variety of sectors, including medical, commerce, web search, and the military.

CNN

Convolutional neural network is a class of deep learning methods which has become dominant in various computer vision tasks and is attracting interest across a variety of domains, including radiology.

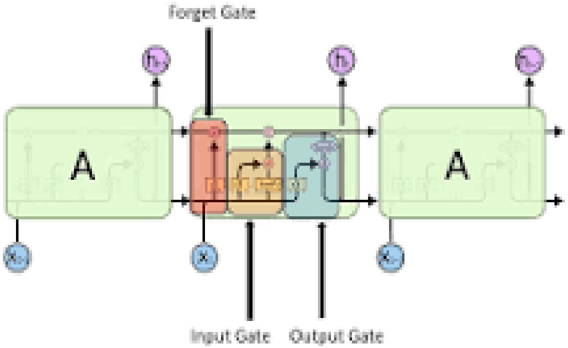
• Convolutional neural network is composed of multiple building blocks, such as convolution layers, pooling layers, and fully connected layers, and is designed to automatically and adaptively learn spatial hierarchies of features through a backpropagation algorithm.

• Familiarity with the concepts and advantages, as well as limitations, of convolutional neural network is essential to leverage its potential to improve radiologist performance and, eventually, patient care.

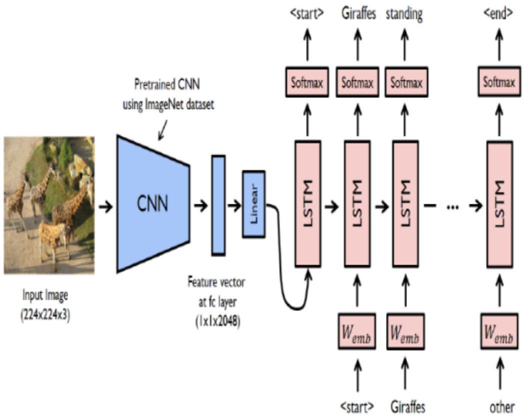


LSTM

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies, LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behaviour, not something they struggle to learn.



It is used for time-series data processing, prediction, and classification. LSTM has feedback connections, unlike conventional feed-forward neural networks. It can handle not only single data points (like photos) but also complete data streams (such as speech or video).



Our model consists of 3 phases:

1.Image Feature Extraction

2.Sequence processor

3.Decoder

Image Feature Extraction

- The features of images are extracted from the Xception model. This model has a great experience in object identification.
- Xception is **a convolutional neural network that is 71 layers deep**. You can load a pretrained version of the network trained on more than a million images from the ImageNet database . The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals.
- These are processed by a Dense layer to produce a 2048 vector element representation of the photo and passed on to the LSTM layer.
- This model configuration learns very fast.

Sequence processor

- The function of a sequence processor is for handling the text input It acts as a word embedding layer. The embedded layer consists of two rules:
  Extract the required features of text
  Ignore padded values.
- It contains a mask to ignore the padded values.
- The network is then connected to a LSTM for the final phase of the image captioning.

Decoder

- The final phase combines the input from Image feature extraction and Sequence processor using an additional operation.
- It fed into a 256 neural layer and then to a final output.
- Dense layer that produces a prediction of the next word in the caption over the entire vocabulary which was formed from the text data that was processed in the sequence processor phase.

The core concept of LSTM's are the cell state, and it' various gates. The cell state act as a transport highway that transfers relative information all the way down the sequence chain. You can think of it as the "memory" of the network. The cell state, in theory, can carry relevant information throughout the processing of the sequence. So even information from the earlier time steps can make it's way to later time steps, reducing the effects of short-term memory. As the cell state goes on its journey, information get's added or removed to the cell state via gates. The gates are different neural networks that decide which information is allowed on the cell state. The gates can learn what information is relevant to keep or forget during training. Sigmoid Gates contains sigmoid activations. A sigmoid activation is similar to the tanh activation. Instead of squishing values between -1 and 1, it squishes values between 0 and 1. That is helpful to update or forget data because any number getting multiplied by 0 is 0, causing values to disappears or be "forgotten." Any number multipliedby 1 is the same value therefore that value stay's the same or is "kept." The network can learn which data is not important therefore can be forgotten or which data is important to keep. Sigmoid squishes values to be between 0 and 1

FLOW OF THE IMAGE CAPTIONING

The following diagram shows us the exact flow of process of captioning image.

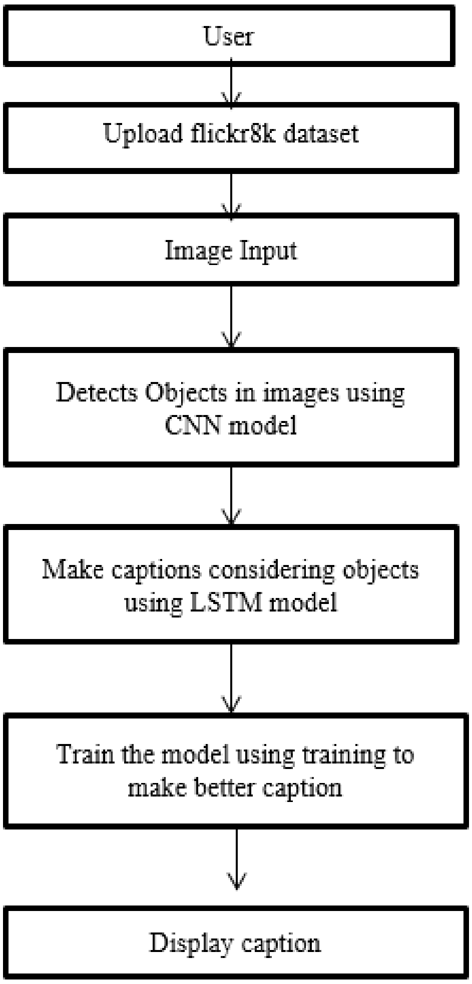
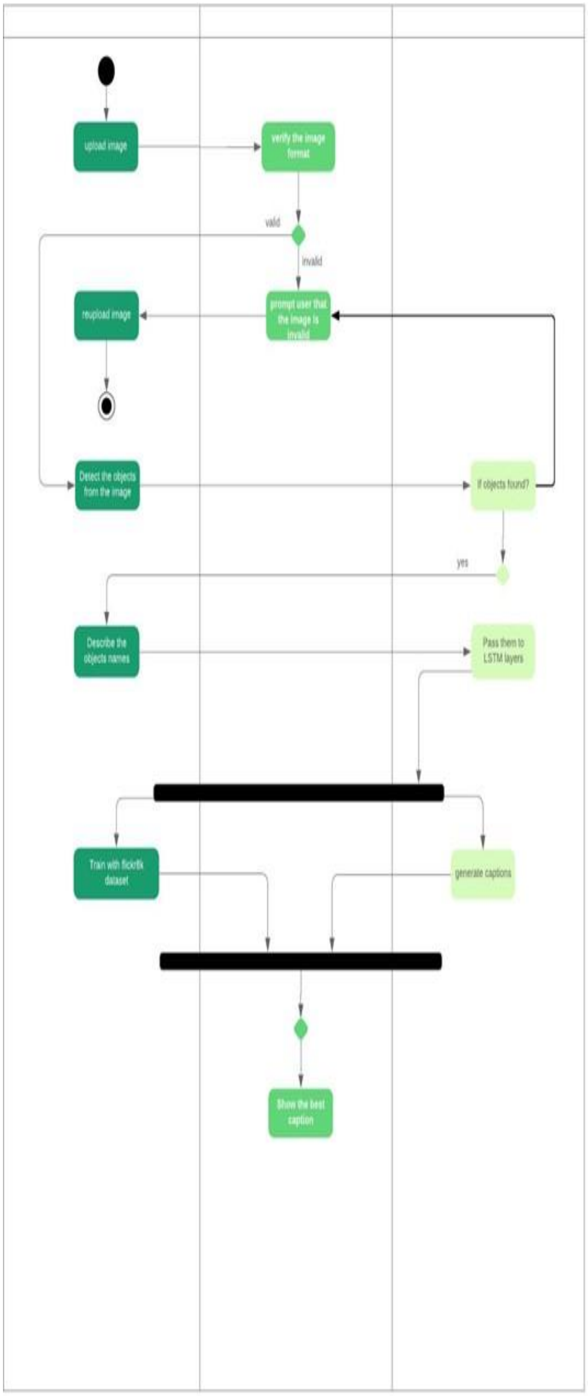

User
Upload flickr8k dataset
Image Input
Detects Objects in images using CNN model
Make captions considering objects using LSTM model
Train the model using training to make better caption
Display caption

Figure: Flow Chart

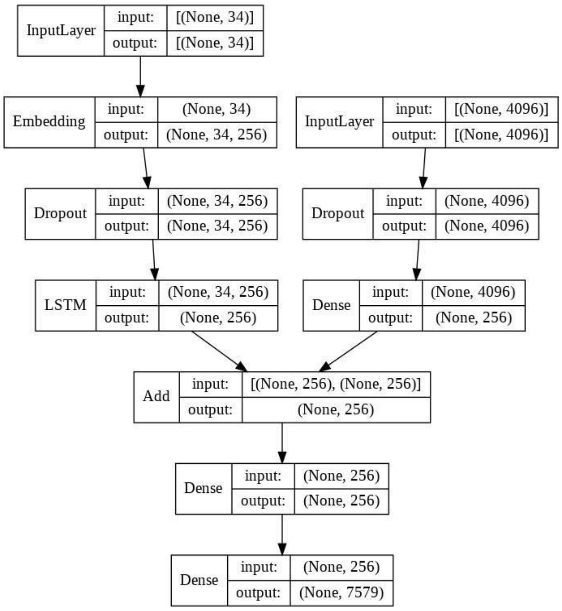


The steps involved in implementation are:

- Get Dataset
- Prepare Photo Data
- Prepare Text Data
- Load data
- Encode text data
- Define model
- Fit model
- Evaluate model

- Generate captions

Trained model architecture



DATASETS

We decide to use the Flickr8k dataset. It has 8092 images and 5 captions for each image. Each image has 5 captions because there are different ways to caption an image. This dataset has predefined training, testing and evaluation subsets of 6000, 1000 and 1000 images respectively.

HARDWARE AND SOFTWARE REQUIREMENTS

Recommended System Requirements to train model.

- A good CPU and a GPU with atleast 8GB memory
- Atleast 8GB of RAM
- Active internet connection so that keras can download inceptionv3/vgg16 model weights

Required libraries for Python along with their version numbers used while making & testing of this project

- Python - 3.6.7
- Numpy - 1.16.4
- Tensorflow - 1.13.1
- Keras - 2.2.4
- nltk - 3.2.5
- PIL - 4.3.0
- Matplotlib - 3.0.3
- tqdm - 4.28.1

ACKNOWLEDGEMENTS

We would like to express our special gratitude to our mentor Mr.Devarapu Sreenivasa rao who gave us a golden opportunity to do a wonderful project on this topic. It makes us to do a lot of research and learnt new things. We are really thankful to that.

In addition to that , we would also thank my friends who helped us a lot in finalizing this project within the limited time frame.

CONCLUSION

After training our model using VGG16 and LSTM validation loss is less for this model when compared to all other models. The BLUE score for the predicted captions are at an average of 0.72 which is better among any other models .we can improve our model's accuracy by increasing number of epochs and using better pretrained .

The text description of the image can improve the content-based image retrieval efficiency, the expanding application scope of visual understanding in the fields of medicine, security, military and other fields, which has a broad application prospect. Theoretical framework and research methods of image captioning can promote the development of the theory and application of image annotation.

REFERENCES

[1] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings us- ing large weakly annotated photo collections. European Conference on Computer Vision. Springer, pages 529–545, 2014.

[2] Peter Young Micah Hodosh and Julia Hockenmaie, Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47:853–899, 2013.

[3] Ryan Kiros, RuslanSalakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. Workshop on Neural Information Processing Systems (NIPS), 2014.

[4] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. International Conference on Machine Learning, 2048- 2057, 2015.

[5] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. IEEE International Conference on Computer Vision (ICCV), pages 4904–4912, 2017.

[6] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4651–4659, 2016.

[7] He, Kaiming, et al. "Deep Residual Learning for
Image Recognition." IEEE Conference on Computer
Vision and Pattern Recognition IEEE Computer Society, 770-778. (2016)

[8] Mao, Junhua, et al. "Explain Images with Multimodal
Recurrent Neural Networks." Computer Science (2014)

[9] Vinyals, Oriol, et al. "Show and tell: A neural image
caption generator." IEEE Conference on Computer
Vision and Pattern Recognition IEEE Computer Society, 3156-3164. (2015)

[10] Xu, Kelvin, et al. "Show, Attend and Tell: Neural
Image Caption Generation with Visual Attention."
Computer Science ,2048-2057. (2015)