INVESTIGATION ON ATTENTION BASED LSTM IMAGE CAPTIONING USING DEEP LEARNING AND CNN TECHNIQUES

[1]Y.Mounica, [2]U.Vijaya Bharathi, [3]B.Venkanna, [4]A.Devender Reddy
[1,2,3]Assistant Professor, [4]UG Student, [1,2,3,4]Department of CSD, Kasireddy Narayanareddy College of Engineering and Research, Hyderabad, Telangana

Abstract
Generating a description of an image is called image captioning. Image captioning is a challenging task, because it involves the understanding of the main objects, their attributes, and their relationships in an Image. It also involves the generation of syntactically and semantically meaningful descriptions of the images in natural language. A typical image captioning pipeline comprises an image encoder and a language decoder. Convolution Neural Networks (CNNs) are widely used as the encoder while long short-term memory (LSTM) networks are used as the decoder. A variety of LSTMs and CNNs including attention mechanisms are used to generate meaningful and accurate captions. Traditional image captioning techniques have limitations in generating semantically meaningful and superior captions. In this research, we focus on advanced image captioning techniques, which are able to generate semantically more meaningful and superior captions. As such we have made four contributions.

We investigate an attention-based LSTM on image features extracted by Dense Net, which is a newer type of CNN. We integrate Dense Net features with attention mechanism and we show that this combination can generate more relevant image captions than other CNNs.

Keywords: LSTM, CNN, Pooling, recurrent neural networks, Correlation

INTRODUCTION

Image captioning is the task of providing a natural language description of the content in an image and lies at the intersection of computer vision and Natural Language Processing (NLP). Automatic image captioning is useful to many applications, such as developing image search engines with complex natural language queries and helping the visually impaired people to understand their surroundings. Hence, image captioning has been an active research area. The advent of new convolutional neural networks and object detection architectures has contributed enormously to improving image captioning.

Moreover, sophisticated sequential models, such as attention-based recurrent neural networks, have also been presented for accurate image caption generation.

Inspired by neural machine translation, most modern deep learning-based image captioning methods use an encoder-decoder framework. In this framework, an encoder is used to encode an intermediate representation of the information contained within the image. A decoder is used to decode this information into a descriptive text sequence. Thus, this framework is composed of two principal modules.

We further extend the work by using an additional CNN layer to incorporate the structured local context together with the past and the future contexts attained by Bi-directional LSTM. A pooling scheme namely Attention Pooling is also used to enhance the information extraction capability of the pooling layer. Consequently, it is able to generate contextually superior captions.

Existing image captioning techniques use human-annotated real images for training and Testing, which involve an expensive and time-consuming process. Moreover, nowadays bulk of the Images are synthetic or generated by machines. There is also a need for generating captions for such Images. We investigate the use of synthetic images for training and testing image captioning. We show that such images can help improving the captions of real images and they can effectively be used in caption generation of synthetic images

We use bi-directional self-attention as a language decoder. Bi-directional decoder can capture the context in both forward and backward directions, i.e., past context as well as any future context, in caption generation. Consequently, the generated captions are more meaningful and superior to those generated by typical LSTMs and CNNs.

Convolutional Neural Network (CNN) as an encoder for image feature extraction and a Long Short-Term Memory (LSTM) model as a language decoder for caption generation. Different CNNs such as Alex Net, VGG Net, ResNet, and Dense Net have their own strengths and weaknesses. It is generally accepted that the deeper the network is, the more relevant are the learned features. However, if the depth of the network exceeds a threshold, one may obtain the opposite effect, i.e., a decline in performance. There are two main reasons behind this fact:

The vanishing-gradient problem: when the input or the gradient passes through many layers, it can vanish or gets "washed out" by the time it reaches the end of the network. The degradation problem. This problem has been addressed in the literature by using residual learning mechanisms such as ResNet. However, the element-wise addition used in the identity mapping in ResNet is computationally expensive during training. In contrast, with Dense Net, each layer has connections with every other layer in the network in a feed-forward manner. The network reuses the feature-maps and uses concatenation for various operations instead of addition. Therefore, it can reduce the number of parameters and it can be memory efficient. Moreover, since each layer of Dense Net receives feature maps from all previous layers, it gets diversified features and tends to have rich patterns. In this paper, we use Dense Net as an encoder to extract image features

However, encoder-decoder based methods focus only on the factual description of an image. They lose the information of the relevant objects in the scene. Visual attention mechanisms can selectively focus the relevant parts of the image for a period of time, similar to the human visual system. Simultaneously, they can discard irrelevant information. Several methods use attention-based techniques and can describe the relevant parts of the image successfully. All of these methods use the three most common datasets.

Microsoft COCO (MSCOCO), Flickr30k, and Flickr8k. The images of all these datasets are human-annotated. However, these deep learning-based methods require a large amount of labelled data in order for them to perform at their very best. Moreover, the manual generation of (additional) data is expensive and time-consuming.

Most deep learning based image captioning methods fall into the category of novel caption generation. Therefore, we focus only on novel caption generation with deep learning. Second, we group

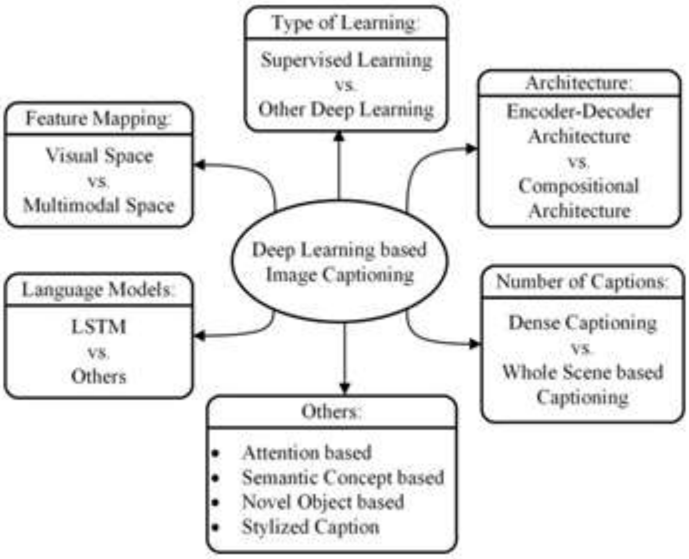


Fig.1 the deep learning-based image captioning methods into different categories namely

In CNN – CNN based Model where CNN is used for both Encoding and Decoding purpose, we observe that CNN – CNN model has High Loss which is not Acceptable as the Generated Captions won`t be Accurate and the Captions Generated here will be Irrelevant to the given test Image.

While in the case of CNN – RNN base Captions there might be less loss compared to the CNN – CNN based model but the Training Time is more. Training Time Effects the whole Efficiency of the model and here we also Encountered another problem i.e. Vanishing Gradient Problem.

In this paper we are Proposing this System so as to Increase the Efficiency of Generating the Captions for the Image and also to Increase the Accuracy of the Captions.

Here are Two Architecture for our proposed model:

- Architecture of ResNet – LSTM Model
- Model Implementation of Flow Chart
- Data Set Collection
- Image and Text Preprocessing
- Vocabulary Building, Defining and Fitting the Model

In deep learning, a **convolutional neural network** (**CNN/Conv Net**) is a class of deep neural networks, most commonly applied to analyse visual imagery. Now when we think of a neural network, we think about matrix multiplications but that is not the case with Conv Net. It uses a special technique called Convolution. Now in mathematics **convolution** is a mathematical operation on two functions that produces a third function that expresses how the shape of one is modified by the other.

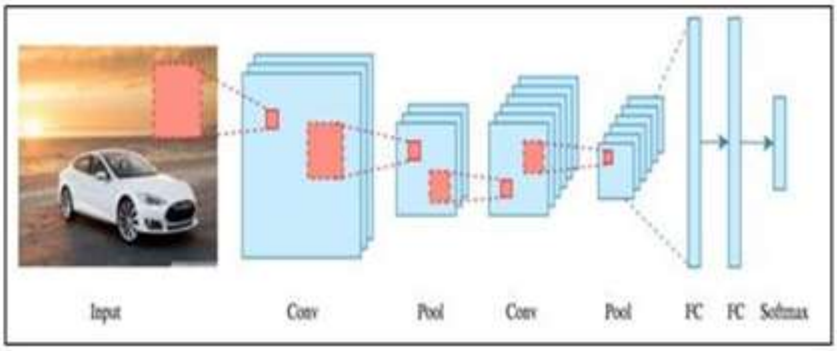


Fig.2

How does it work

Before we go to the working of CNN's let's cover the basics such as what is an image and how is it represented.

Fig.3

For simplicity, let's stick with grayscale images as we try to understand how CNNs work.

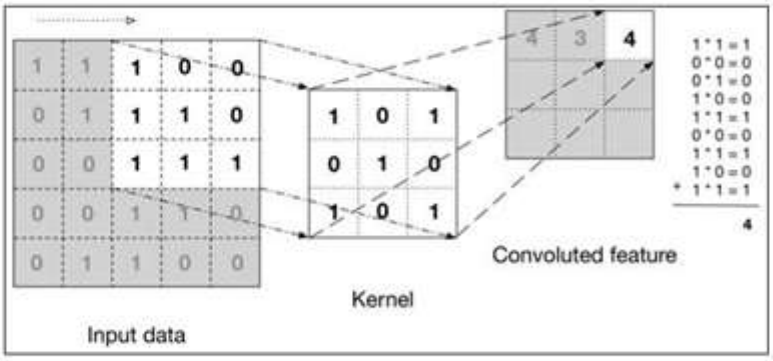


Fig.4

The above image shows what a convolution is. We take a filter/ kernel (3×3 matrix) and apply it to the input image to get the convolved feature. This convolved feature is passed on to the next layer.

Convolutional neural networks are composed of multiple layers of artificial neurons. Artificial neurons, a rough imitation of their biological counterparts, are mathematical functions that calculate the weighted sum of multiple inputs and output an activation value. When you input an image in a Conv Net, each layer generates several activation functions that are passed on to the next layer.

The first layer usually extracts basic features such as horizontal or diagonal edges. This output is passed on to the next layer which detects more complex features such as corners or combinational edges. As we move deeper into the network it can identify even more

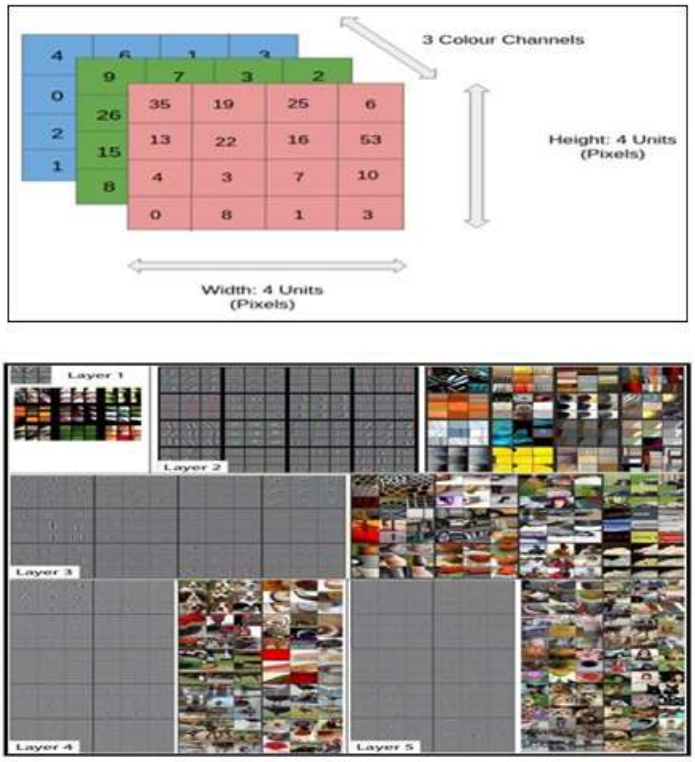


Fig.5

Features such as objects, faces, etc.

Based on the activation map of the final convolution layer, the classification layer outputs a set of confidence scores (values between 0 and 1) that specify how likely the image is to belong to a "class." For instance, if you have a Conv Net that detects cats, dogs, and horses, the output of the final layer is the possibility that the input image contains any of those animals.

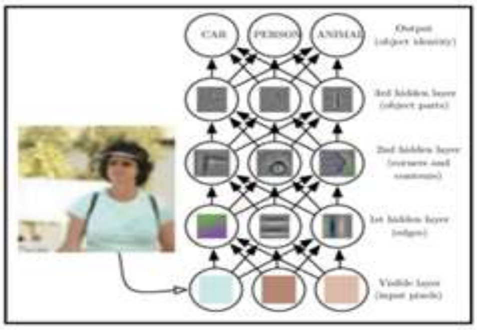
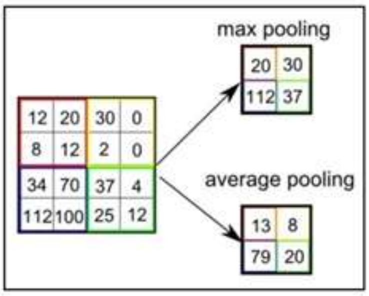

CAR
PERSON
ANIMAL
Output (object identity)
3rd hidden layer (object parts)
2nd hidden layer (corners and contours)
1st hidden layer (edges)
Visible layer (input pixels)

Fig.6

Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to **decrease the computational power required to process the data** by reducing the dimensions. There are two types of pooling average pooling and max pooling. I’ve only had experience with Max Pooling so far I haven’t faced any difficulties.

So, what we do in Max Pooling is we find the maximum value of a pixel from a portion of the image covered by the kernel. Max Pooling also performs as a **Noise Suppressant**. It discards the noisy activations altogether and also performs de-noising along with dimensionality reduction. On the other hand, **Average Pooling** returns the **average of all the values** from the portion of the image covered by the Kernel. Average Pooling simply performs dimensionality reduction as a noise suppressing mechanism. Hence, we can say that Max Pooling performs a lot better than Average Pooling.



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

max pooling

20	30
112	37

average pooling

13	8
79	20

REFERENCES

1. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Available: https://papers.nips.cc/paper/4824-imagenetclassificationwith-deep-convolutionalneural-networks.pdf
2. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database
3. Andrej Karpathy, Li Fei-Fei, Deep VisualSemantic Alignments for Generating Image Descriptions, Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf
4. Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.pdf

5. Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Convolutional Image Captioning, Available: https://arxiv.org/pdf/1711.09151.pdf
6. Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, Automatic Image Captioning, Conference: Multimedia and Expo, 2004. ICME ’04. 2004 IEEE International Conference on, Volume: