International Journal of Current Engineering and Technology
©2021 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Image Captioning Model using Visual Aligning Attention and Deep Matrix Factorization

Shweta Dhurmekar and Prof. Deipali Gore

Department of Computer Engineering PES Modern College of Engineering,Pune-05

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Image captioning technique is a complicated task that bridges both the visual and linguistic domains. Image captioning models are required to understand the content of input images to generate sentences with human languages. The attention technique, widely used for Image Captioning task provides more accurate information. Attention technique explicitly trains the deep sequential models. In this work, we have proposed a system using visual aligning attention model and deep matrix factorization; Visual aligning attention model focuses on the region of interest using CNN and LSTM as encoder- decoder. While DMF works on refinement and assignment of image tag. The dataset used is FLICKR8k for caption generation. The experimental results show that the proposed system gives more accurate results. Captions generated are more descriptive and accurate.

Keywords: *Encoder-decoder; Visual Aligning; Global Aligning; CNN; RNN; Semantic; Remote Sensing; LSTM; Language Model.*

Introduction

Image captioning technique is a complicated task that bridges both the visual and linguistic domains. Image captioning models are required to understand the content of input images to generate sentences with human languages. Attention techniques are known to explicitly train the attention layer which help in generating more accurate caption. Encoder-decoder based models are widely used models for image captioning task. In this model, a convolutional neural network (CNN) is selected as the encoder for extracting image features while the decoder is usually composed of LSTM, RNN or GRU for generating sentences[18]. To predict different words, different image information related to the predicted word is required. If all the image information is fed into the decoder without filtering out useless information, the decoder cannot capture purer information, which is useful for generating more accurate sentences[18]. This gap in image information limits the robustness of the existing models. To overcome such problem attention model was proposed[12s]. Attention layers are able to focus more on regions of interest in images. Images with attention techniques generate purer and useful image captions. Although the attention mechanism has certain effects in suppressing image useless information and improving image caption results, there are still some defects for processing remote sensing images. Compared with natural images, remote sensing images cover a wide area. The useless information in image features is more than that in natural images. Therefore, it is necessary to exclude the useless information by further improving the conventional attention mechanism. In detail, the training of attention layer is not directly constrained by the loss function. The propagation of function loss passes through the MLPs, RNN and embedded layer to reach the attention layer. When predicting visual words, this implicit constraint training process cannot ensure that the attention layer is accurately focused on interested region on the image. At the same time, they cannot guarantee that the useless information in the image features, which will be sent to the following RNN to predict words, can be filtered out at each time step.

Literature Survey

Describing the image and understanding the scene in an image has become an important aspect in industrial and many real life applications. The solution for this is image caption generation. To generate caption for any image machine needs to interpret the image information to generate semantically and syntactically correct caption; which will allow the users to understand the image better. Many techniques over the years have been proposed to solve this problem. Compared to Template based methods and Retrieval based method, Encoder-decoder based methods perform very well[1-3,5,6,11]. Encoder-decoder methods are deep learning based methodology. In this encoder is usually composed of convolutional neural network(CNN) which extracts the features of the image like image object, image features and then process it to decoder. Decoder is mainly of recurrent neural network(RNN), gated recurrent unit(GRU) or long short term memory(LSTM) which processes the image information and generates the appropriate caption. A. Karpathy and L. Fei-Fei et al proposed alignment model on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks (RNN) over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. This study then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. In this work the datasets used are Flickr8K, Flickr30K and MSCOCO datasets. However this work is subjected to few limitations where evaluating every region in isolation leads to computational inefficiency because one must forward every individual region of interest separately through the convolutional network[9]. ShiruQu, Yuling Xi, Songtao Ding et al, proposed a visual attention mechanism a neural and probabilistic framework which combines CNN with a special form of recurrent neural network (RNN) to produce an end-toend image captioning. This work uses a model that takes advantage of word to vector to encode the variable length input into a fixed dimensional vector. Considering the description of the object in an image is not specific enough, they introduce an attention mechanism through visualization to show how the model is able to fix its gaze on salient objects. The datasets used are Flickr8K, Flickr30K and MSCOCO. It has limitations where if the model is not able to identify unknown objects it generates the wrong caption[12]. cPGCON 2020 (Post Graduate Conference for Computer Engineering) of the best performing models. Among these Attention models are finest at generating captions also good at explicitly training the attention layers.

Proposed Methodology

To overcome the existing system issue, in this project we proposed a system implementing an Image Captioning Model using Visual Aligning Attention (VAA) and Deep Matrix Factorization (DMF). There are two contributions of this project are as follows: 1) A novel attention model, VAA, is proposed to align the visual words and their corresponding image features for constraining the attention layers. 2) Deep matrix factorization framework for image tag refinement and assignment.

A. System Architecture

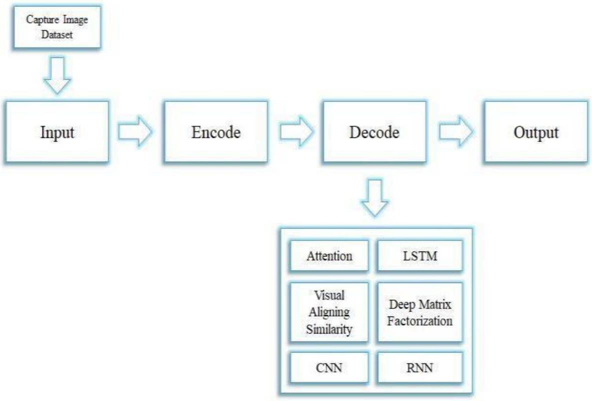


Figure 1. System architecture

CNN

Convolution neural network(CNN) acts as the encoder for image captioning based models. CNN extracts the image information, object detection, object identification to feed the decoder with this information. In encoder decoder based models the attention layers are trained for the model to understand the image.

LSTM

Long short term memory(LSTM) is the decoder which decodes the image information to generate caption. LSTM defines the image parameters; model those parameters. The parameters that are fed into decoder through the encoder are image parameters such image features, image information such people, animals, different objects and more. The model is trained to understand the relationship among these parameters so as to generate the appropriate caption.

Deep matrix factorization

Deep matrix factorization is used for image tag refinement and assignment. The images provided by the user have insufficient or incomplete tags for the images which makes it difficult to map correct image to appropriate tag. So deep matrix factorization will map the images to appropriate tags which will result to generate correct captions. This algorithm uses the reference tags as the input and maps the images containing the matching tag to generate the caption.

Mathematical Model

- S = {I, P, O}
- Where,
- I = Input images from dataset
- P = {P1, P2, P3}
- Where,
- P1 = Image detection
- P2 = Features extraction

- P3 = Generate caption
- O = The images containing captions with highest similarity score to the query are returned as output.
- F is the set of functions used for remote sensing image caption generation.
- F={F1; F2; F3}
- where,
- F1 is a function for CNN Encoder.
- F2 is a function for LSTM Decoder.
- F3 is a function for Deep matrix factorization.

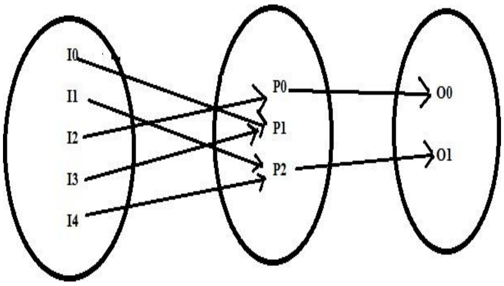
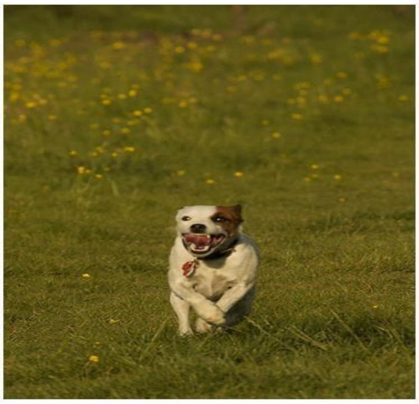


Figure 2. Mathematical model

D. System Requirement



A. Software Requirement

1. OS- Microsoft Windows 7 or Above
2. Programming Language- Python

B. Hardware Requirement

1. Processor- Core Intel I3 or Higher
2. RAM- 2GB or Higher
3. Hard Disk- 100GB (min)

E. Experimental Setup

Recommended System Requirements to train model.
A good CPU and a GPU with at least 8GB memory
At least 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

Result and Discussions

The system will take the image input from the FLICKR8k dataset and process the image using reference captions to generate the candidate caption. The model need to trained to generate captions. Dataset used has images with human generated captions as the reference for the machine to learn and generate candidate caption. The results show the caption using various input such as caption using word as input or using phrase as input.

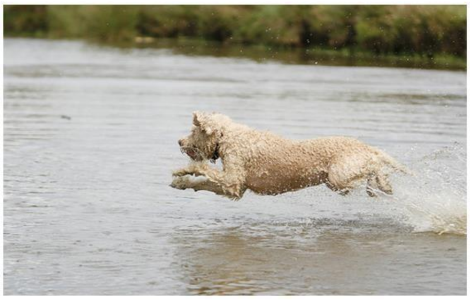


Figure 3. Results

Actual: A tan dog jumps into water.
Using word as input: A brown dog is walking in the water. **Using phrase as input**: A dog walking a shallow river looking into in the water gray shorts the horizon with in down water.

Figure 4. Results
Actual: A brown and white dog running in a field covered in yellow flowers.
Using word as input: A black dog running Frisbee in galloping.
Using phrase as input: A white and brown dog in his face a field.

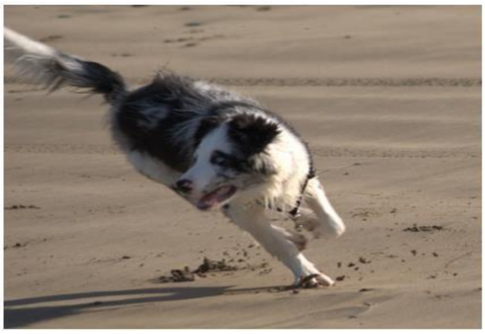


Figure 4. Results

Actual: A black and white dog is running on the beach.
Using word as input: Man running along the beach, running and a beach.

Using phrase as input: A black and white dog running UNK playing on a coffee shop a skate park.

Conclusions

In this work, a novel Visual Aligning Attention (VAA) and Deep Matrix Factorization (DMF) model is proposed for image captioning. This model is proposed to explicitly train the attention layer. CNN is the encoder to extract image features and LSTM is the decoder to generate sentences for describing the content of input images. More importantly, the trained attention layers are able to focus on the regions of interest. This provides purer and more useful image information for the decoder to generate sentences to describe the content of input images. While Deep Matrix Factorization tags the images and recommends to the system for relevant image caption.

References

[1]. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, ``Every picture tells a story: Generating sentences from images,'' in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15_29.

[2]. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, ``Baby talk: Understanding and generating simple image descriptions,'' in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1601_1608.

[3]. V. Ordonez, G. Kulkarni, and T. L. Berg, ``Im2Text: Describing images using 1 million captioned photographs,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143_1151.

[4]. Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”.

[5]. M. Hodosh, P. Young, and J. Hockenmaier, ``Framing image description as a ranking task: Data, models and evaluation metrics,'' *J. Artif. Intell. Res.*, vol. 47, no. 1, pp. 853_899, 2013.

[6]. Y. Gong, L.Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, ``Improving image-sentence embeddings using large weakly annotated photo collections,'' in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 529_545

[7]. C. Sun, C. Gan, and R. Nevatia, ``Automatic concept discovery from parallel text and visual corpora,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 2596_2604.

[8]. Karen Simonyan & Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, ICLR 2015.

[9]. Andrej Karpathy, Member, IEEE, and Li Fei-Fei, Member, “Deep Visual- Semantic Alignments for Generating Image Descriptions” Vol. 14 No. 8, August 2015.

[10]. Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola, “Stacked Attention Networks for Image Question Answering”, IEEE Conference on Computer Vision and Pattern Recognition, 2016.