International Journal of Advanced Science and Technology
Vol. 29, No. 3s, (2020), pp. 975-980

IMAGE CAPTION GENERATOR USING DEEP LEARNING

[1]B.Krishnakumar , K.Kousalya[2], S.Gokul[3], R.Karthikeyan[4] and D.Kaviyarasu[5]
1Assistant Professor Department of CSE, Kongu Engineering College, Anna University, India
[2]Professor Department of CSE,Kongu Engineering College, Anna University, India
[345]UG student Department of CSE, Kongu Engineering College,Anna University, India

Abstract

Computer Vision and Natural Language Processing in artificial intelligence is used for automatically describing the content of an image. In order to describe the image a well-formed English phrases is needed. Automatically describing image content is very much helpful to the visually impaired people to understand the problem better. The paper is intended to identify objects and inform people through audio and text messages. It recognizes image and converts to audio using GTTS and converts to text using LSTM. Initially, the input image is converted to a grayscale image that is processed through the Convolution Neural Network (CNN) to correctly identify the objects. Objects in the image are correctly identified using OpenCV, which is then converted to audio and text messages. The proposed method for blind people is designed to expand to people with vision loss in order to achieve their full potential.

Keywords: *Neural Networks, Imag, Caption, RNN, LSTM, GTTS , Deep Learning.*

1. INTRODUCTION

Machine translation is evolving at an alarming rate, leading to the technological developments in the field of machine learning. The rapid developments in machine learning influence and strengthen other branches of the technical industry.The application of Artificial Intelligence and Neural Networks to complicated natural language processing problems such as speech recognition and machine translation is leading to surprisingly rapid developments. Among the many advances, one such example is progress in the area of "Describing Images." The job of presenting a summary of the picture is difficult. First it requires understanding the visual information and using natural language processing software to translate the knowledge into sentences. This includes the creation of a model capable of capturing the association present on the related image in the visual and natural language. The problem is multimodal, which generates the need to construct a hybrid model that can leverage the problem's multidimensionality. Approaches such as prototype based and retrieval based approaches have historically been used to solve the issue. The major drawback of these approaches, however, is that the results don't translate into new images and thus these methods struggle to generalizeSecond, to turn the details into sentences, it requires understanding the visual information and using natural language processing tools. It involves the development in visual and natural language of a model capable of capturing the connection present on the relevant image. The problem is multimodal, which generates the need to construct a hybrid model that can take advantage of the multiplication of problems capacity of deep neural networks has also been successfully tested to construct image captions and their ability to generalize is much better than the traditional methods. Such models often use specific deep neural networks such as convolutional neural network (CNN), long-term memory (LSTM) networks, recurrent neural network (RNN) to

ISSN: 2005-4238 IJAST

learn the common embedding implicitly by encoding and decoding the direct modalities. These methods give better results compared with earlier methods on all can subtitle generation datasets.A variety of models have been produced during the last few years. A modified "merge model" paradigm built in by integrating CNN and LSTM, is the common theme in these approaches. The most popular and easily available datasets used to test new methods are:

- Flickr8k(14) - It has 8000 images and 5 captions for each image.
- Flickr30k(14) - This dataset has 31783 images with 5 full sentence level caption for each image.

2. LITERATURESURVEY

The problem using natural language processing in the automatic description of visual data has been studied in [5,22]. It uses visual primitive recognizer with formal language for converting visual to text, e.g. And-Or Graphs or logic systems are converted to natural language through rule-based systems. Such systems are weak and used only on very limited areas, e.g. Sports, traffic scenes etc.. The issue of picture depiction using natural language processing has picked up importance in current trends. In recent days natural language processing methodologies were used to identify objects using attributes and locations.

Farhadi et al. [4] states the process involved in converting image to text. Li et al. [15] describes the object as sentence by dividing the objects and matches the corresponding phrases, finally all phrases are joined and form a sentence. Kulkani et al. [12], used a numerous detection graphs beyond triplets wih high complexity and text generation is done by using tempates. Several language models used in [17,18,13,14,3] are using language parsing. The mentioned methodologies are improper in terms of text generation.

More amount of work has discussed for converting a given image [8, 6, 20]. Those approached are dealt with the text and images in same vector space. The neural network [19] are used to embed images and sentence together. Image cropping and sub sentences mapped [10] are discussed. The above techniques are retrieving words from each object in image, but generated descriptions are not good.

Sequence modelling is done by combining deep convolutional neural network which is used for image classification in [9] along with the recurrent neural networks to design a separate networks that generates image descriptions. The recurrent neural network is trained with this end to end architecture as a frame of reference[7].Instead of starting with the sentence, the convolutional net concept was used for sequence generation in machine [2, 1, 21]. Similar work Kiros et al. [11] was done by usin neural net as feed forward technique, i.e. predict the next word by knowing previous images.

Mao et al. [16] had done the task of prediction by using RNN as his recent work. The work done in this paper is very much identical with Mao's work but with some necessary changes. Here the visual images is given as input to the model, the architecture of the model is built in such a way that it act as more powerful. Using this powerful model and the image, the objects present in the image can be explained clearly as text. As a result the proposed work of this paper provides a better result when compared to other works for the well known benchmark datasets. Kiros et al.[14] uses a model based on computer vision which is more powerful and LSTM model which is used to generate text to build a multimodal embedding model. Here two different model one for image and the other for text is combined to create joint multimodel embedding model. The proposed model produces text which are

tuned towards ranking.

3. PROPOSED WORK

In this proposed work Convolutional Neural Network is used for extracting features from the image. We used the pre-trained model VGG16. Long Short Term Memory (LSTM) use the features from Convolutional Neural Network to generate a description of the image. gTTS api used to generate the image caption to audio.

3.1 MODEL

We use A. Karpathys pretrained model as our baseline model. This model is trained only on MSCOCO dataset. The model uses a 16-layer VGG Net for embedding image features which is fed only to the first time step of the single layer RNN which is constituted of longshort term memory units (LSTM). The RNN size in this case is 512. Since words are one hot encoded, the word embedding size and the vocabulary size is also 512. The two parts, CNN and RNN, are joined together by an intermediate feature expander, that feeds the output from the CNN into the RNN. Recall, that there are 5 labeled captions for each image. The feature expander allows the extracted image features to be fed in as an input to multiple captions for that image, without having to recompute the CNN output for a particular image.

In VGG-Net, the convolutional layers are interspersed with maxpool layers and finally there are three fully connected layers and softmax. The softmax layer is required so that the VGGNet can eventually perform an image classification.But for the purpose of image captioning, we are interested in a vector representation of the image and not its classification. And so, the last two layers are eliminated and the output from the fully connected layer can be extracted and expanded to feed into the RNNpart of the architecture.

3.2 VGG-16

VGG-16 is a convolutional neural network architecture, it's name VGG-16 comes from the fact that it has 16 layers. It's layers consists of Convolutional layers, Max Pooling layers, Activation layers, Fully connected layers. There are 13 convolutional layers, 5 Max Pooling layers and 3 Dense layers which sums up to 21 layers but only 16 weight layers.Conv 1 has number of filters as 64 while Conv 2 has 128 filters, Conv 3 has 256 filters while Conv 4 and Conv 5 has 512 filters.VGG-16 network is trained on ImageNet dataset which has over 14 million images and 1000 classes, and acheives 92.7% top-5 accuracy. It surpasses AlexNet network by replacing large filters of size 11 and 5 in the first and second convolution layers with small size 3x3 filters.

3.3 EVALUATE MODEL

Once the model we can evaluate the skill of its predictions on the test dataset. We evaluate a model by generating descriptions for all photos in the test dataset and evaluating those predictions with a standard cost function. First, we need to be able to generate a description for a photo using a trained model.
This involves passing in the start description token startseq, generating one word, then calling the model recursively with generated words as input until the end of sequence token is reached endseq or the maximum description length is reached. Then the behavior and generates a textual description given a

trained model, and a given prepared photo as input. It calls the function word for id in order to map an integer prediction back to a word.We will generate predictions for all photos in the test dataset and evaluate a trained model against a given dataset of photo descriptions and photo features. The actual and predicted descriptions are collected and evaluated collectively using the corpus BLEU score that summarizes how close the generated text is to the expected text.

BLEU scores are used in text translation for evaluating translated text against one or more reference translations. Here, we compare each generated description against all of the reference descriptions for the photograph. We then calculate BLEU scores for 1, 2, 3 and 4 cumulative n-grams. A higher score close to 1.0 is better, a score closer to zero is worse.

3.4 GENERATE NEW CAPTION

We need to generate captions for entirely new photographs is in the model. We used the Tokenizer for encoding generated words for the model while generating a sequence, and the maximum length of input sequences, used when we defined the model.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed model is implemented in keras framework used in jupyter notebook environment that runs entirely on the system. Neural Network layers, cost functions, optimizers, initialization schemes, activation functions, and regularization schemes are all standalone modules that are combine to create new models. Keras offers broad adoption, support for a wide range of production deployment options, integration with at least five back-end engines and strong support for multiple GPUs and distributed training.

4.1. DATASET

The Image captioning is a challenging dataset composed of 8,091 images of (500×333pixels).The dataset has a pre-defined training dataset (6,000 images), development dataset (1,000 images), and test dataset (1,000 images).The All the samples selected were resized to a 224 ×224 RGB image. Then samples were uniformly distributed into three sets: the training, testing and validation sets.

5. CONCLUSION

The model has been successfully trained to generate the captions as expected for the images. The caption generation has constantly been improved by fine tuning the model with different hyper parameter. Higher BLEU score indicates that the generated captions are very similar to those of the actual caption present on the images.

We have implemented a CNN-RNN model by building an image caption generator. Some key points to note are that our model depends on the data, so, it cannot predict the words that are out of its vocabulary. We used a small dataset consisting of 8091 images. For production-level models, we need to train on datasets larger than 100,000 images which can produce better accuracy models.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and

translate.arXiv:1409.0473, 2014.

[2] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares,H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation.In EMNLP, 2014.

[3] D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP, 2013.

[4] 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 ECCV, 2010.

[5] R. Gerber and H.-H. Nagel. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In ICIP. IEEE, 1996.

[6] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In ECCV, 2014.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory.Neural Computation, 9(8), 1997.

[8] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR, 47, 2013.

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In arXiv:1502.03167, 2015.

[10] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. NIPS,2014.

[11] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In arXiv:1411.2539, 2014.

[12] R. Kiros and R. Z. R. Salakhutdinov. Multimodal neural language models. In NIPS Deep Learning Workshop, 2013.

[13] 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 CVPR, 2011.

[14] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In ACL, 2012.

[15] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi. Treetalk:Composition and compression of trees for image descriptions.ACL, 2(10), 2014.

[16] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In Conference on Computational Natural Language Learning,2011

[17] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Explain images with multimodal recurrent neural networks. In arXiv:1410.1090, 2014.

[18] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. C.Berg, K. Yamaguchi, T. L. Berg, K. Stratos, and H. D. III.Midge: Generating image descriptions from computer vision detections. In EACL, 2012.

[19] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In NIPS,2011.

[20] R. Socher, A. Karpathy, Q. V. Le, C. Manning, and A. Y. Ng.Grounded compositional semantics for finding and describing images with sentences. In ACL, 2014.

[21] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In NIPS, 2014.

[22] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2t: Image parsing to text description. Proceedings of the IEEE, 98(8), 2010.

[23] Marc Tanti, Albert Gatt, Kenneth P. Camilleri :Where to put the Image in an Image Caption Generator .arXiv:1703.09137