MULTI TASK LEARNING FOR CAPTIONINGIMAGES WITH NOVEL WORDS

Sreekantha B
Associate Professor Department of Information Science and Engineering, HKBK College of Engineering, Bangalore, India
Email: shreekantha.is@hkbk.edu.in

Saniya Sultana
Department of Information Scienceand Engineering, HKBK College of Engineering,Bangalore, India
Email: saniya544505@gmail.com

Shabreen Taj
Department of Information Scienceand Engineering, HKBK College of Engineering, Bangalore, India Email:shabreenshabu240@gmail.com

Shikhar
Department of Information Scienceand Engineering, HKBK College of Engineering,Bangalore, India
Email: shikhar0055@gmail.com

Tasmiya Khanum
Department of Information Scienceand Engineering, HKBK College of Engineering,Bangalore, India
Email: khntasmiya@gmail.com

Abstract

In this article, we present a Multi-task Learning Approach for Image Captioning (MLAIC), which is motivated by the observation that humans can easily complete this task since they are skilled in a variety of disciplines. MLAIC is made up of three essential parts in particular:(i)A multi-object classification model that learns detailed category-aware picture representations by employing a CNN image encoder (ii) An image captioning model that generates text descriptions of images by sharing its CNN encoder and LSTM decoder with the object classification task and the syntax that learns better syntax aware LSTM based decoder.The added object classification and grammar knowledge is especially generation task, respectively. (ii) A syntax generation model that improves syntax aware LSTM based decoder. (iii) A syntax generation model advantageous for the picture captioning model.The experimental outcomes on the MS-COCO dataset show that our model outperforms other formidable rivals in termsof performance.

1. INTRODUCTION

Humans are inherently multi-tasking cognitive creatures, which explains their exceptional ability to verbally describe a visual. Humans have acquired those talents since infancy by adapting to understand the complicated outside environment through several channels of observation and communication, rather than just learning to accomplish a single activity. They receive training by executing a variety of pertinent activities simultaneously in order to build a strong foundation of knowledge and abilities for comprehending and describing scenarios. Studying all pertinent activities that lead to a machine intelligence's growth is a crucial first step if one hopes to build one that mimics the vast array of human skills In this might provide a phrase that reliably and appropriately describes an image in a more satisfying manner. Based on this finding and the Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), we think a multi-task learning framework can assist study, we construct a computerised picture captioning agent that can also handle a few other related tasks. We are inspired by the notion that a cognitive AI is by nature multi-tasking. A crucial task in computer vision and natural language processing is image captioning, which involves creating a phrase that describes the key elements of a picture. [Bernardi et al.,

2016] In recent years, it has frequently been tackled using a supervised learning framework, which involves gathering human- produced samples and building models based on matching the generated text to the gathered annotations. We assume that such learning frameworks have limited theoretical success in two dimensions. First off, the difficulty of the problem can only be understood by the model built from the data collected in relation to the degree of complexity of the instances given. The dataset is essentially a finite collection, hence the complexity that is shown should be kept to a minimum.Second, many characteristics of the structured output that have not been prioritised in the traditional evaluation metric for picture captioning, such as object categories and syntax of produced sentences, are frequently not sensitive to the loss function employed in numerically optimising the model.In fact, our research demonstrates how providing the learning framework with more pertinent data and objectives may be advantageous in both respects. Although the concept of multi-task learning is not new in the field of machine learning, it frequently remains a difficult step in the development of empirically effective systems. We contend that it basically supports the argument for developing a successful picture captioning system. In an ablation study of our models (see Table 1), a model that is not aware of the sentence syntax may produce a sentence that is incompletely describing a picture; a model that is

to carry out another duty of syntax annotation. The goal of co-training is to make up for the standard framework's unable to recognise all objects presented may produce a sentence that is incompletely describing a salient object in the image. However, a better system with components that have been trained on several related tasks simultaneously a captioning system perform better in ways that can't be quantified by traditional evaluation measures.

Our method generates picture captions by taking use of the encoder decoder framework's recent breakthroughs [Karpathy and Fei-Fei, 2015; Vinyals et al., 2015]. This framework's basic principle is to employ a convolutional neural network (CNN) as an encoder to extract features that correspond to the visual understandings of an input picture, and then feed the feature vector to a recurrent neural network (RNN)-based decoder to produce image captions. In this research, we suggest further regular isationsutilising multi-task learning, sharing this common framework with other comparable techniques. First, co-training to execute a second job of multi-object classification regularises our CNN encoder.Second, [Nadejde et al., 2017] our RNN decoder is additionally regularised using the co- training lack of an image caption regularisation requirement rather than to obtain the highest performance on these auxiliary tasks. Following is a summary of our key contributions:

- To jointly train the job of picture captioning and two additional related tasks, multi-object categorization and syntax generation, we present MLAIC, a multi-task learning system.The auxiliary tasks improve the CNN encoder and RNN decoder inthe image captioning model. In particular,

1. Multi-object classification co-trained with image captioning aims to build an object-rich image encoder and enhances the accuracy of identifying contextual information of an image.
2. Under carefully monitored experimental conditions, the differences in caption language and style in relation to several object categories are investigated.
3. From a language modelling standpoint, the RNN decoder is capable of utilising word-level grammar to produce high-quality captions. It eliminates the problems caused by repeated words and unfinished phrases.

- According to both the findings of the offline Karpathy test split and the online server assessment, MLAIC performs remarkably well on the widelyused MSCOCO dataset.

2. RELATED WORK

A difficult task is to produce written descriptions from photographs. The majority of current methods, benchmark datasets, and assessment metrics for picture captioning were thoroughly reviewed by Bernardi et al. in their 2016 paper.

Deep neural network technology has recently made significant improvements that have significantly enhanced the process of picture captioning. A common method for creating captions for images is to combine CNN and RNN [Karpathy and Fei-Fei, 2015; Vinyals et al., 2015], where CNN is used to extract the compact representational vector of the entire image and RNN is used to create the language model that will be applied to the representational vectors to create captions.The fundamental encoder- decoder structure has been shown to benefit greatly from visual attention. When creating visual descriptions, for instance, Xu et al. [2015] proposed an attention-based approach that automatically learns where to attend. The CNN's final convolutional layer's feature map is reweighted by the attention as a function of spatial probabilities.In order to encode where (i.e., attentive spatial locations at various layers) and what (i.e., attentive channels) the visual attention was, Chen et al. [2017] dynamically varied the sentence production context in multi-layer feature maps.The integration of the encoder-decoder architecture with reinforcement

learning paradigms for picture captioning has attracted growing interest [Liu et al., 2016; Rennie et al., 2016; Zhao et al., 2017]. For instance, Liu et al. [2016] used the policy gradient (PG) approach to directly optimise a linear combination of the SPICE and CIDEr metrics, where the SPICE score made sure the captions were syntactically fluid and the CIDEr score made sure they were semantically loyal to the picture.A self- critical sequence training (SCST) technique using the well-known REINFORCE algorithm was put out by Rennie et al. in 2016.By training a task concurrently with related tasks, multi-task learning is a valuable learning paradigm to increase a task's ability to be supervised and generalised [Caruana, 1998].By combining the video caption decoder with outside language models, Venugopalan et al. [2016] investigated linguistic enhancements. By combining
their expertise with two related directed generation task in computer vision and natural language processing is image captioning, which involves creating a sentence expressing the key elements of an image [Bernardi et al., 2016].In recent years, it has frequently been tackled using a supervised learning framework, which involves gathering samples created by humans and building models based on matching the generated text to the annotations gathered. Such learning frameworks, in our opinion, have limited potential in two aspects. First off, the model built using the data gathered can only understand how complicated an issue is based on how complex the instances are.The stated complexity should be kept to a minimum because the dataset is essentially a finite collection. Second, many characteristics of the structured output that have not been highlighted in the traditional evaluation metric for picture captioning, such as object categories and syntax of produced sentences, are frequently not sensitive to the loss function employed in numerically optimising the model. Actually, our research demonstrates that providing the learning framework with more pertinent data and objectives might be beneficial in both aspects.

4. PROPOSED SYSTEM

For the object classification of a picture x, we have the equation $y^o = y^o_1, y^o_2,..., y^o_C$, where y o I = 1 if item I is annotated in this image; otherwise, $y^o_i = 0$, and C
is the number of object categories.When captioning apicture, we use the formula $y^w = y^w, y^w,..., y^w$ tasks a temporally-directed unsupervised video prediction task and a logically-directed language entailment generation job Pasunuru and Bansal [2017] enhanced video captioning.
Our approach is different from the methods mentioned above. In order to enhance the performance of both the CNN encoder and LSTM decoder, we execute image captioning using multi- task learning, which exchanges knowledge with three related tasks: multi-label classification, image captioning, and syntax creation.

3. EXISTING SYSTEM

In this research, we construct a computerised picture captioning agent that can also carry out a few other related tasks. We are inspired by the notion that a cognitive AI is by nature multi-tasking. A important where T is the length of the series, to represent the image description. For syntax generation, the combinatory category gramma (CCG) super tag sequence with regard to the associated caption of picture x is denoted by the formula $y^s = y^s_1, y^s_2,..., y^s_T$ .The captions, annotated CCG supertags, and object categories vocabularies are each denoted by the letters W_W, W_S, and W_O respectively. The structure of our model MLAIC shows how it trains the object classification and syntax generation tasks alongside the picture captioning job. The CNN encoder for the picture captioning job and the object classifier share the same encoder.To aid the LSTM decoder in focusing on various facets of the pictures with regard to the object labels, all object labels are encoded as low dimensional distributed embeddings and handled as additional input. The decoder for image captioningand the syntax generation job share an LSTM decoder. By training the shared LSTM decoder
similarly, [Nadejde et al., 2017] predicts words and syntax.

5. SYSTEM ARCHITECTURE

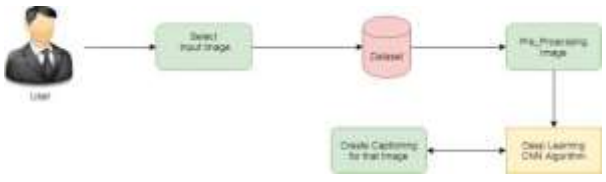

Image based model extracts the features of our image from dataset and it pre-processes image. For our image

based model, we use deep learning and CNN algorithm, the image summarizes the approach of image caption generator, usually image based model rely on convolutional neural network . A pre-trained CNN extracts the features from our input image, creates the vocabulary for the image.

6. INFERENCE

In the inference process, we use Beam Search: iteratively consider the set of the k best sentences up to time t as candidates to generate sentences of size, and keep only the resulting best k of them. We improve the Beam Search approach in the following

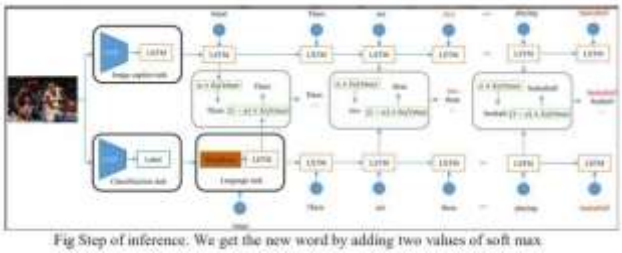

Fig Step of inference. We get the new word by adding two values of soft max

experiments with a beam of size 3. For each image I outside of image–caption datasets, the IM can recognise the label L. Then the word2vector of L would be the zero state of LM. The CM may generate a wrong sentence, L_{CM} and the LM can create a sentence L_{LM} conditioned on L. We produce a new sentence by combining L_{CM} and L_{LM}

$$h_{\mathrm{CM}_{-1}} = \mathrm{CNN}(I), \quad (17)$$

$$k = \mathrm{argmax}(\mathrm{softmax}(W_{\mathrm{fc}} h_{\mathrm{CM}_{-1}})), \quad (18)$$

$$h_{\mathrm{LM}_{-1}} = W_v k, \quad (19)$$

$$h_{\mathrm{CM}_t} = W_e w_{\mathrm{CM}_t}, \quad t \in \{0, \ldots, N-1\}, \quad (20)$$

$$h_{\mathrm{LM}_t} = W_e w_{\mathrm{LM}_t}, \quad t \in \{0, \ldots, N-1\}, \quad (21)$$

$$p_{\mathrm{LM}_{t+1}} = \mathrm{softmax}(\mathrm{fc}(\mathrm{LSTM}(h_{\mathrm{LM}_t}))), \quad (22)$$

7. CONCLUSION

By concurrently training object categorization and syntax generation with image captioning, we suggested a unique multi-task learning technique to enhance image captioning. The object categorization assisted in developing more accurate picture representations and enhanced visual attention, while the syntax generation assisted in reducing the issue of producing redundant and incomplete phrases.On the widely known MSCOCO dataset, we carried out extensive tests to confirm the efficacy of our strategy. The experimental findings showed that, in comparison to other potent rivals, our approach produced excellent outcomes.

8. REFRENCES

[1] [Anderson et al., 2017] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. vqa and image captioning require both bottom-up and top-down focus. ArXiv preprint 1707.07998, 2017

[2][Bernardi et al., 2016] Raffaella Bernardi, RuketCakici,Desmond Elliott, AykutErdem, ErkutErdem, NazliIkizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. A review of models, datasets, and assessment metrics for automatic description creation from photos. JAIR,\s55:409–442, 2016.

[3] [Caruana, 1998] Multitask learning, by Rich Caruana, is discussed on pages 95 to 133 in Learning to Learn. Springer, 1998.

[4][Chen et al., 2017] Long Chen, Hanwang Zhang, Jun Xiao, LiqiangNie, Jian Shao, Wei Liu, and Tat-Seng Chua. Scacnn: Spatial and channel-wise attention in convolutional networks for picture captioning. In CVPR, 2017.

[5][Gu et al., 2017a] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Coarse-to-fine learning for captioning images is called stack captioning. ArXiv preprint 1709.03376, 2017.

[6][Gu et al., 2017b] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. a linguistic empirical research for captioning images on CNN. In ICCV, 2017.

[7][He et al., 2016] Kaiming He, Jian Sun, Xiangyu Zhang, and Shaoqing Ren. Image identification using deep residual learning. Pages 770–778 of CVPR, 2016.

[8] [Karpathy and Fei-Fei, 2015] Li FeiFei with Andrej Karpathy. Deep visual-semantic alignments for producing picture descriptions. 2015 CVPR, pages 3128–3137

[9][Li et al., 2017] Yale Song, Jiebo Luo, and Yuncheng Li. improving pairwise ranking for multi- label image classification. ArXiv preprint 1704.03135, 2017.

[10] [Lin et al., 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft Coco: "Common items in context." Pages 740–755 of ECCV,Springer, 2014