IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 7.301

Volume 3, Issue 2, March 2023

Multi Task Learning for Captioning Images with Novel Words

Sreekantha B[1], Saniya Sultana[2], Shabreen Taj[3], Shikhar[4], Tasmiya Khanum[5]

Associate Professor, Department of Information Science and Engineering[1]
Students, Department of Information Science and Engineering[2,3,4,5]
HKBK College of Engineering, Bangalore, Karnataka, India
shreekantha.is@hkbk.edu.in, saniya544505@gmail.com, shabreenshabu240@gmail.com,
shikhar0055@gmail.com, khntasmiya@gmail.com

Abstract: *In this article, we introduce a Multi-task Learning Approach for Image Captioning (MLAIC), which is inspired by the fact that people can readily finish this job given their proficiency in a variety of fields. There are three crucial components that make up MLAIC in particular:(i)A multi-object categorization model that uses a CNN image decoder to learn intricate category-aware picture representations (ii) A model for creating image captions that uses an LSTM-based decoder that is grammar conscious and shares its CNN encoder and LSTM decoder with an object categorization job to create text summaries of pictures. The additional object categorization and grammatical skills are particularly relevant to the job of creation. (ii) A syntactic generation model that enhances LSTM-based decoders that are syntax cognizant. An effective grammar creation model for the image labeling model is (iii).Our model beats other strong competitors in terms of efficiency, according to testing results on the MS-COCO dataset.*

Keywords: Multi-task Learning Approach for Image Captioning

I. INTRODUCTION

Humans are inherently multi-tasking cognitive creatures, which explains their exceptional ability toverbally describe a visual. Humans have acquired those talents since infancy by adapting to understandthe 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 astrong 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 reliablyand appropriately describes an image in a more satisfying manner. We believe that a multitasking learning framework can help research based on this discovery and Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18). Based on these results, we created a computerized image labeling assistant that can also perform some other related tasks. The idea that conscious AI is inherently multitasking inspires us. Image annotation—a key task in computer vision and natural language processing—is the creation of a sentence that captures the salient features of an image.

Picture [Bernardi et al.] in 2016 have frequently addressed the use of a supervised learning framework in recent years, which involves building models based on collecting human-generated examples and comparing the generated text with the collected feedback. These learning models, in our view, have only modest theoretical implications in both directions. First of all, a model built using data collected relative to the complexity of the given examples is the only way to understand the complexity of the problem. Since the information is essentially a finite set, the complexity in which it is displayed should be as low as possible. Second, many features of the structured output that are not prioritized in the traditional evaluation scale of image feedback, such as object categories and syntax, often produce sentences that are insensitive to the loss function used to numerically optimize the model and, in fact, provide more relevant information and targets for the learning structure, which is beneficial in both ways. The study shows. Although multitasking learning is not a new idea in the field of machine learning, it still proves to be a challenging step in building experimentally successful systems. We argue that this proves fundamental to establishing an efficient image

DOI: 10.48175/IJARSCT-8928

classification system. In the ablation analysis of our models (see Table 1), a model image that is not known about the syntax may lead to a sentence that only partially describes it. A form of grammatical annotation that performs another function may also result in a sentence that is unfamiliar with sentence structure. 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'tbe quantified by traditional evaluation measures.

By taking advantage of new developments in decoder architecture [Karpathy and Fei-Fei, 2015; Vinyals et al., 2015], our method for generating image captions. The basic idea behind this approach is to use a convolutional neural network (CNN) as an encoder to extract features from an input image associated with visual comprehension, and then pass that feature vector to a recurrent neural network (RNN) based decoder. To create translations for images. Inthis 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 asummary of our key contributions:

We introduce MLAIC, a multi-task learning system for simultaneously teaching an image labeling task and two additional related tasks, multi-object classification and rule generation. The CNN encoder and RNN decoder in the image labeling model are enhanced by supplement functions. In order to build an object-rich image decoder and improve the recognition accuracy of image environmental information, a method specifically seeks,

1. Co-trained multi-object classifier with image annotation.
2. Variations in description language and style with respect to different groups of objects are examined under closely supervised test conditions.
3. From a language modeling perspective, an RNN encoder can use word-level syntax to generate high-level translations.

It eliminates problems caused by redundant clauses and incomplete sentences. Online server evaluation and Karpati's offline split test results show that MLAIC performs excellently on the popular MSCOCO dataset.

II. 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 popular way to generate image captions is CNN and RNN [Karpathy and Fei-Fei, 2015; Viñales et al. .

Translation is done using vectors and has a significant positive effect on the underlying structure of the visual vision decoder. For example, Xu et al. [2015] proposed a vision-based approach that naturally learns where to listen when generating visual images. Focus refocuses the feature map of the final convolutional layer of the CNN according to geographic probability. In multilevel feature maps, Chen et al. [2017] change the context of sentence construction in terms of where (i.e., alerted spatial locations at different scales) and which (i.e., alerting channels) are suitable for recording visual attention. Increasing interest in combining encoder and decoder design with reinforcement learning models for image labeling has been demonstrated [Liu et al., 2016; Rennie et al., 2016; Zhao et al., 2017]. For example, Liu et al. [2016] directly optimized a linear combination of SPICE and CIDEr metrics using the political gradient (PG) method, where the SPICE score ensures that the translation is artificially flexible and the CIDEr score ensures that it is conceptually faithful to the image. The. Publish a Critical Self-Sequence Training (SCST) method that uses the well-known REINFORCE algorithm 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 withoutside language models, Venugopalan et al. [2016]investigated linguistic enhancements. By combining

Basunuru and Bansal [2017] found that classification of videos improved by experimenting with two similar prompted generation tasks—a temporally self-directed video prediction task and a logical direction language generation function. Our strategy is different from the above methods. We perform image annotations using knowledge-sharing multitask learning with three associated tasks, multilabel classification, image annotation, and rule generation, to improve the performance of both the CNN encoder and LSTM decoder.

III. EXISTING SYSTEM

In this study, we build a computerized image labeling assistant that is capable of performing a few additional connected tasks. The idea that a conscious AI is multitasking by nature is what motivates us. A important task in computer vision and natural languageprocessing is image captioning, which involves creating a sentence expressing the key elements of animage [Bernardi et al., 2016]. In recent years this has often been addressed using a supervised learning framework, which involves collecting human-made examples and developing models based on comparing the resulting text with collected comments. We believe that such educational models hold two areas of limited promise. First, the complexity of the cases determines how complex the problem is according to the model created using the collected data. The data set is basically a finite set, so the specific complexity should be kept to a minimum. Second, many features of the structured output are not highlighted in the traditional evaluation scale of image feedback, such as the classes of objects and the syntax of the sentences produced, which are often sensitive to the loss function used in the numerical optimization of the model. . In fact, our study shows that it can be beneficial in two ways to provide more relevant information and goals to the learning system.

IV. PROPOSED SYSTEM

For the object classification of a picture x, we havethe equation $y^o = y^o{}_1, y^o{}_2, \ldots, y^o{}_C$, 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 $y^w = y^w , y^w , \ldots, y^w$, 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 $y^s = y^s{}_1, y^s{}_2, \ldots, y^s T$ .The captions, annotated CCG supertags, and object categories vocabularies are each denoted by the letters W_W, W_S, and W_O respectively. The structureof 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 sharethe same encoder.To aid the LSTM decoder in focusing on various facets of the pictures with regardto the object labels, all object labels are encoded as low dimensional distributed embeddings and handledas 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.

V. SYSTEM ARCHITECTURE

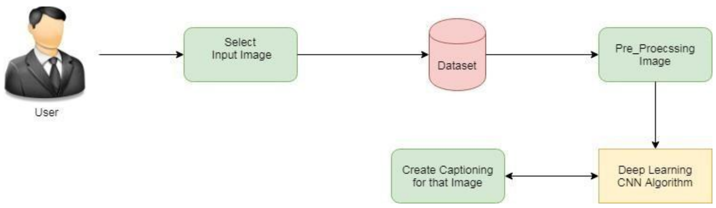

Image based model extracts the features of our image from dataset and it pre-processes image. For our image based model, we use deep learning and CNN algorithm, the image summarizes the approach of image caption generator, usually image based model rely on convolutional neural network . A pre-trained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

VI. 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, whilethe syntax generation assisted in reducing the issue ofproducing redundant and incomplete phrases. On thewidely known MSCOCO dataset, we carried outextensive tests to confirm the efficacy of our strategy.The experimental findings showed that, in comparison to other potent rivals, our approach produced excellent outcomes.

REFERENCES

[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