Attention Image Caption with DenseNet

To cite this article: Fucheng You and Yangze Zhao 2019 J. Phys.: Conf. Ser. 1302 032048

View the article online for updates and enhancements.
Attention Image Caption with DenseNet

Fucheng You and Yangze Zhao
Beijing Institute of Graphic Communication. No. 1, Xinghua Street (two section), Daxing District, Beijing, China
Email: 1263108136@qq.com

Abstract. Image caption aims to automatically generate a descriptive text from the image. In this paper, we get a more effective model than the infrastructure by replacing the pre-trained feature extraction model VGG and LSTM with DenseNet and GRU. Experimental results show that the training time of the modified model is shorter than that of the original model and the performance can be guaranteed.

1. Introduction
Vision plays an important part in getting information from the outside. But for visually impaired people, they can only feel the world by touch. In order to help this kind of people, it’s important to translate the surrounding as image, and then extract key information behind it. It’s easy to take the computer technology in mind and hope that we can make use of computer to automatically complete the process of image recognition and then generate the corresponding description. Image caption takes an image as input and then generate a suitable sentence of the content of the image [1,2,3]. This task is similar to translate a picture into a description sentence[7,9,10,]. This task has a variety of applications, so it is significant to explore this field.

This research is also full of challenges. In the perspective of experts, image caption contains not only image classification but also text generation. However, that does not means that this task only involves them, and it also include how to achieve the connections of image features and text vector. Majoring in this field is helpful to deepen the understanding of each field and expand research ideas.

Traditional methods of image caption have many disadvantages, such as low efficient and expensive computation. As the rise of deep learning, deep learning method has already become the dominant method by now. Inspired by machine translation, more and more models of image caption adopt encoder-decoder architecture, which regard image caption as a particular case of machine translation. This kind of architecture extracts image high-level semantic features as feature vector, which contains the image content [11,12,13]. And then image feature vector is fed to RNN or LSTM with text vector. It’s common to initialize the LSTM with image content vector and the training process will not take it again. We replace the pre-trained architecture from ResNet with DenseNet, because we think ResNet doesn’t contain full features, such as combination of high-level features with low-level features. DenseNet connect the output of pre-layer with all layers after current layer, it fuses all features together. In addition, we adopt GRU instead, because of its simplicity. The main model is shown as below.
2. Model Architecture

As is shown above, the model contains two core components, Encoder and Decoder. First, given a image, the pre-trained DenseNet extracts high-level map and map is separated into fixed length vector. The vector will be sent to GRU at each step, together with the word vector. And during training time, the model will assign different attention to each features vector. The main principle is that convolutional neural network has already been proved to be efficient to classify image. The extracted features inform the decoder the image contents. The source paper applies ResNet to extract image content features. However, we think ResNet can be more replaced with more suitable architecture of DenseNet, which combines low-level features with high-level features. As for the decoder, we replace the original LSTM with GRU because GRU is more simple and has only two gates, update gate and reset gate. The input word as a one-hot vector will be sent to GRU combined with the weights.

Principally, the model aims to maximize the probability of the correct description provided the picture. The formula is shown as (1)

$$\theta^* = \arg\max_{\theta} \sum (I, S) \log(p(S | I, \theta))$$

(1)

In the formula, \( \theta \) represent the parameters which are going to be trained. It is the input pictures. \( S \) represent the correct translation sentence. We compute the joint probability with chain rule. After we get the extracted features and weighted word vector, GRU takes both of them as input and then outputs the probability distribution across all words. We add correct word probabilities at each step and then output the final result of its log function

$$\log p(S | I) = \sum_{t=0}^{N} \log p(S_t | I, S_0 \ldots S_{t-1})$$

(2)

2.1. Training

In the training time image and word vector pairs are sent to the GRU, and GRU shares the same parameters at each step. GRU outputs the probability distribution of the next word, which has been defined in the word dictionary as fixed vector. The goal of training is to make the maximum likelihood obtain the maximum value. So our loss function can be the negative sum of all the predicted descriptions of training images.
2.2. Inference
When it comes to inference time, there are two common methods LPCN and Beam Search. In this paper we adopt the second method, which has been used by NIC. Traditional method always tries to find the best path, but this method also result a very large search space. And it’s easy to cause memory overflow. So Beam Search prefers breadth-first and then executes optimize search space in order to reduce memory consume. As for the precision, we adopt BLEU, which is based on N-gram matching rule. In other words, it compares the identity words between the output sentence and the source description and then compute the ration of these words in source description.

3. Dataset
Our experiment is based on MS COCO dataset. COCO dataset contains more than 120 thousand images. And each image is described of five different captions. The dataset is labeled with 80 categories, and MS COCO dataset has already been widely used in image caption task. In our experiment, we just train about 20 thousand images because we are limited by reality.

4. Result
The image effects are shown below.

![Figure 2. State of the art](image1.png)

![Figure 3. Ours](image2.png)
5. Conclusion
The project was partly based on Joint Funding Project of Beijing Municipal Commission of Education and Beijing Natural Science Fund Committee (KZ201710015010), Project of National Scientific Found (No.61370188), Project of Beijing Municipal College Improvement Plan (PXM2017_ 014223_000063) and New Project of Green Printing and Publishing Technology by Cooperative Creating Center (PXM_014223_000025) and BIGC Project (Ec201803 Ed201802 Ea201806).

6. References
[1] Vinyals O, Toshev A, Bengio S, et al. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(4): 652-663.
[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017: 5998-6008.
[3] Cho K, Courville A, Bengio Y. Describing multimedia content using attention-based encoder-decoder networks[J]. IEEE Transactions on Multimedia, 2015, 17(11): 1875-1886.
[4] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.
[5] Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 933-941.

[6] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1243-1252.

[7] Lu J, Xiong C, Parikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 375-383.

[8] Chen L, Zhang H, Xiao J, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5659-5667.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012: 1097-1105.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[11] Ordonez V, Kulkarni G, Berg T L. Im2text: Describing images using 1 million captioned photographs[J]. Advances in neural information processing systems, 2011: 1143-1151.

[12] Yang Y, Teo C L, Daumé III H, et al. Corpus-guided sentence generation of natural images[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 444-454.

[13] Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating image descriptions[J]. Proceedings of the 24th CVPR, 2011.