The Summary of ”Deep Visual-Semantic Alignments for Generating Image Descriptions”[1] Paper

Furkan Gül

I. Motivation

A semantic description about a visual scene in an image can be inferred in a detailed manner by a quick and short looking of a human at the image. However, in the case of image recognition models, it is a bit hard problem to get close to the visual recognition level of humans. The main focus of almost all prior studies in the visual recognition area has been placed in categorizing images into different fixed classes of labels. In the last decade, there have been tremendous achievements in this specific visual recognition task. However, modeling different image classes in a given dataset are majorly limited when the human ability to deduct rich semantic description from only one image is considered.

II. Approach

The main aim of this article is to create novel descriptions of image regions. There are two different models to achieve this ultimate goal. The first model discussed in the paper is responsible for aligning some parts of sentences to the image regions. In the training phase of this model, the input consists of images and sentence descriptions for each specific image. Captured hidden correspondences between sentence snippets and image regions in the first model are used as a training set for the second model. The second model, which is a Multimodal Recurrent Neural Network, aims to extract the novel sentence snippets from given image regions.

In the first model, for the representation of images, Region Convolution Neural Network(RCNN) by Girshick et al.[2], is pretrained over ImageNet dataset[3]. Besides the entire image, the top 19 detected regions are also used. By training about 60 million parameters in RCNN, each image is transformed into *h*-dimensional vectors, where *h* stands for the size of the multimodal embedding space. In their studies, *h* varies between 1000 and 1600. For the representation of the words in sentences, Bidirectional Recurrent Neural Network (BRNN)[4] is used. Since visual regions are represented by *h*-dimensional embedding space, the BRNN should represent each word as an *h*-dimensional vector for compatibility between the sizes of the embedding vector of images and words. In the BRNN, 300 dimensional word2vec[5] weights initiate a word embedding matrix. It remains the same because of the overfitting issue. Moreover, the rectified linear unit (ReLU) is preferred as activation functions. The dot product of $v_i^T s_t$ between the *i*-th region and *t*-th word is represented as a similarity metric to identify the score between image *k* and sentence *l* as the following:

$$S_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} max(0, v_i^T s_t) \tag{1}$$

In the equation above, g_k and g_l represent the set of image regions in the image *k* and the set of sentence snippets in the sentence *l* respectively where *k* and *l* indices iterate over the whole images and sentences in the training set. The cost function is a type of SVM loss, which can be seen in the below equation.

$$C(\theta) = \sum_k [\sum_l max(0, S_{kl} - S_{kk} + 1) + \sum_l max(0, S_{lk} - S_{kk} + 1)] \tag{2}$$

It is not sufficient to define the score between the *i*-th image region and the *t*-th word since the main aim of this study is to create novel sentence snippets instead of just a single word. That’s why they try to correlate a sequence of words with a single bounding region in an image. To solve this issue, they mention a Markov Random Field (MRF). The outcome of this process is a set of visual regions with the best segments of text.

In the second model, the input to the model consists of not only the set of full-sized images and their sentence descriptions but also the set of image regions and text snippets as obtained in the prior model. To predict variable-sized segments of text with a given input image, they use a Recurrent Neural Network (RNN) model as a language model with one modification, which is conditioning the RNN on the embedding vector of an input image via bias interventions in the first place. This conditioned RNN model is called the Multimodal RNN. In the Multimodal RNN model, the hidden layer size is generally 512 neurons. The first starting state is initiated as a zero vector. The softmax classifier is the cost function to maximize the log probability of the predicted sequence of words. For the optimization step in the first alignment model, 100 mini batches of image and sentence couples via SGD with 0.9 momentum are used. Crucial hyperparameters which are the weight decay and the learning rate are tuned by the cross-validation method. Gradients are clipped element-wise at 5. The droup-out method is applied in all layers except the recurrent layers to avoid the overfitting issue.

III. Dataset & Results

Flickr8K[6], Flickr30k[7], and MSCOCO[8] datasets are used in their experiments. All three datasets contain 8000, 31000, 123000 images respectively. There are at least 5 sentence descriptions for every image in these three datasets. For both the validation and test phases with MSCOCO dataset, 5000 images are used. In training phase of both Flickr datasets, for both the validation and test step, 1000 images are used.

For the first model, ranking experiments are conducted to assess the prediction quality of the image-sentence alignment. There are two metrics reported in the articles for the evaluation, the median rank and Recall@K. Recall@K is a measure of the fraction of times that a true object was in the top K results. The BRNN model has 22.2, 48.2, 61.4 for R@1, R@5, R@10 metric respectively. In terms of Med rank, the BRNN results in 4.8. At the time of the publication year, around the end of 2014, these results were the best among others.

For the evaluation of the second model in terms of full images and sequences of words, BLUE-n scores up to BLUE-4 are measured. By looking at the result table, it can be said that the results do not outperform the other models such as Google NIC[9].

IV. A Couple of Words

The next paper to be summarized will be from 2015 [10]. But it will be the last paper that I will be reading in terms of the late publication year. The reason for choosing these two papers from 2014 and 2015, which seems a bit older, is that in my opinion, they will be a favorable starting point to build a concrete base in terms of understanding NLP usage in image captioning problems with neural networks point of view especially BRNN. As you mentioned earlier in the proposal paper review, I am right into changing the articles that I will examine to the new released NLP based ones.

References

[1] A. Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017.

[2] Ross B. Girshick, J. Donahue, Trevor Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[3] Jia Deng, W. Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, 2009.

[4] Mike Schuster and K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45:2673–2681, 1997.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *ArXiv*, abs/1310.4546, 2013.

[6] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). *J. Artif. Intell. Res.*, 47:853–899, 2013.

[7] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[8] Tsung-Yi Lin, M. Maire, Serge J. Belongie, J. Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[9] Oriol Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, R. Salakhutdinov, R. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *ArXiv*, abs/1502.03044, 2015.