Seminar: Generating Narrative Paragraph for Photo Stream via Bidirectional Attention Recurrent Neural Networks

Chen Zongqi
Matriculation Number 1564832
University of Mannheim, Germany
zchen@mail.uni-mannheim.de

Abstract

When machine learning techniques develop rapidly in image captioning, other extension relevant areas appeal to many researchers doing further study. Among them generating story from sequential photo stream becomes quite interesting and challenging task. Compared with single image captioning, this task needs to handle with two mainly problems facing to researchers, large visual variance in sequence and longterm language coherence among multiple sentences. Till now those two critical questions are solved by different approaches. In this paper, we mainly focus on one of them, generating story from sequential photos stream via Bidirectional Attention Recurrent Neural Network method, and discus several relevant approaches from first paper to state-of-the-art.

1 Introduction

2 Related Work

Due to interaction between computer vision and natural language processing is new topic, particularly in image description, there are several researches divided in three categories: single-frame to single-sentence, multi-frame to single-sentence and multi-frame to multi-sentence.

2.1 Single-frame to single-sentence

These researches focus on image captioning task, which can be classified into two sub-categories: semantic element based methods

2.2 Multi-frame to single-sentence

This family of approaches, mainly focus on video captioning to captures the temporal dynamics in variable-length of video frames sequence and to map them to a variable-length of words.

2.3 Multi-frame to multi-sentence

The work by is the first scheme to explore the task of image streams to sentence sequence.

3 Approach

In this part, we mainly introduce our topic model Bidirectional Attentional Recurrent Neural Networks, and then briefly explain two related but important models, Coherence Recurrent Convolutional Network (Park and Kim, 2015) and Generating Narrative Paragraph by Adversarial Training (Wang et al., 2018). Particularly the paper which proposes Coherence Recurrent Convolutional Network is the first paper in this field and the paper which proposes using Adversarial Training is the newest paper till now.

3.1 Bidirectional Attention Recurrent Neural Network

In order to generate narrative paragraph from photo stream, we need to not only learn from photos but also analyse from sentences. Based on this idea, the approach Bidirectional Attention Recurrent Neural Network is separated by two main parts, one is aimed to combine sentence embedding space with image embedding space called Semantic Space Embedding and another one is aimed to predict sentence embedding features using Bidirectional Attention Recurrent Neural Network. In general, photo streams and corresponding sentences as input would be manipulated by CNN and Word2Vecs. For image stream, each image would be transformed into 4096-dimension VGG features (Simonyan and Zisserman, 2014)and then mapped into 300-dimension image embedding space. For sentence, each sentence would be transformed by Word2Vecs

(Mikolov et al., 2013)into 300-dimension sentence semantic embedding space. There is a very important assumption that 300-dimension image embedding space and 300-dimension sentence semantic embedding space share one embedding space, which means if the contents of image and sentence are similar then their euclidean distance is close.

Joint Embedding for Semantic Space To create this semantic space embedding, we could use a loss function and then use learning algorithm to approach the minimum loss value. Here we use a contrastive loss function equation1:

$$\begin{aligned} C^{remb}(x,v) = & \sum_{x \in X, v \in V, v' \in V'} max(0, \alpha - xv + xv^{'}) \\ & + \sum_{x \in X, x' \in X', v \in V} max(0, \alpha - xv + x^{'}v), \end{aligned} \tag{1}$$

where X is the image embedding vector and V is the sentence embedding vector, $X^{'}$ and $V^{'}$ are negative paired image and sentence sample, α is entropy for judging positive image-sentence pair similarity. Contrastive loss function (Chopra et al., 2005) is aimed to make gradient of loss function successfully approach to minimum value. The principle of contrastive is that we want to introduce a negative pair to confirm that the result we get is what we expect but not stochastic error. For example we get a result from equation and I want to know this equation that is correctly. So we input a wrong number into the equation, if we get correct result then we can confirm this equation does not work.

Bidirectional Attention Recurrent Neural Network for Textual Story Generation The goal of this part is to use Bidirectional RNN with Attention modelling to predict sentence embedding features. The input of BARNN model is image and sentence embedding vectors. The output h is sentence embedding features with image stream inputting.

The idea of Attention modelling (Itti et al., 1998) is that when human focus on one object, they would ignore something non-relevant,so that we can introduce new attention weight to balance what should be concentrated on or what should be ignored. Meanwhile, under the assumption that same content has close distance in embedding space, we introduce equation 2:

$$R_{pt} = x_p x_t, \tag{2}$$

where x_p and x_t are p and t time inputting images, and R_{pt} is the relation of image x_p and image x_t. We use R_{pt} as our attention weight to focus on objects that are more important, then we add them more weight.

Later we have a new designed Gated Recurrent Unit with skip gate we called Skip-GRU. The classic GRU(Chung et al., 2014) has update gate and reset gate. In order to introduce our attention modelling, we add a skip gate. The advantage of skip-gate is that we can add attention weight to current hidden state and furthermore it can influence the main object features in our semantic embedding space. See our Skip-GRU equations 3:

$$\begin{aligned} z_t &= \sigma(W_{zx}x_t + W_{zh}h_{t-1}) \\ r_t &= \sigma(W_{rx}x_t + W_{rh}h_{t-1}) \\ s_t &= \sigma(W_{sx}x_t + W_{sh}h_p) \\ \tilde{h} &= \tanh(W_{hx}x_t + W_{hh}rt \odot h_{t-1} \\ & \quad + \sum_{p<t} R_{pt} \cdot W_{hp}s_t \odot h_p) \\ h_t &= z_t\tilde{h} + (1 - z_t h_{t-1}), \end{aligned} \tag{3}$$

where t and p are times, x_t and x_p are t time and p time input image, z_t, r_t and s_t are t time update gate, reset gate and skip gate, $\tilde{h}$ and h_t are current hidden state and t time output, $\odot$ means element-wise multiplication and $\tanh$ means hyper tangent function.

For Bidirectional Framework, we apply our new skip-GRU into our framework in both forward and backward pass. See equation 4below:

$$\begin{aligned} (z_t^f, r_t^f, s_t^f, \tilde{h}^f, h_t^f) &= sGRU(x_t, h_{t-1}^f, R, h_p^f; W^f) \\ (z_t^b, r_t^b, s_t^b, \tilde{h}^b, h_t^b) &= sGRU(x_t, h_{t-1}^b, R^T, h_p^b; W^b) \\ h_t &= W_h^f h_t^f + W_h^b h_t^b, \end{aligned} \tag{4}$$

where f means forward pass and b means backward pass. We learn $W = (W^f, W^b, W_h^f, W_h^b)$ as our parameters to learn. And h is what we expect that predict sentence embedding features.

The contrastive loss function is quite same with equation 1 see equation 5 below:

$$\begin{aligned} C^{cpt}(h,v) = & \sum_{v' \in V'} max(0, \gamma - hv + hv^{'}) \\ & + \sum_{h' \in H'} max(0, \gamma - hv + h^{'}v), \end{aligned} \tag{5}$$

where $h^{'}$ and $v^{'}$ are negative image-sentence pairs, and γ is contrastive margin. By using equation 5 we can learn parameter W combined with corresponding image captioning.

Finally, we combine two contrastive loss functions, joint embedding semantic space and bidirectional attention RNN, as one equation 6 see below:

$$C = \sum_{X,V} C^{emb}(x, v) + \sum_{H,V} C^{cpt}(h, v), \qquad (6)$$

where X is sequential image embedding vectors, V is corresponding sentence embedding vectors and H is predict corresponding sentence vectors from BARNN model.

3.2 Coherence Recurrent Convolutional Network

3.3 Adversarial Training

4 Experiment

5 Evaluation

6 Discussion

References

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, June.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

L. Itti, C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 73–81. Curran Associates, Inc.

K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. 2018. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. February.