Seminar: Generating Narrative Paragraph for Photo Stream

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 long-term 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

With computing ability rapidly developing and new theory proposing, there are more and more intersection researches between computer vision area and natural language processing area occurring. This paper will focus on one specified topic from them, which is generating narrative paragraph for photo stream.

The basic task for this topic is that given sequential photos as input and return a story to describe as output. The core problems which need to be solved are how to generate corresponding sentence and how to organize sentences as human-level story. The previous question is based on the related field Image Captioning, such as generator with deep recurrent structure (Vinyals et al., 2014), deep visual-semantic alignments (Karpathy and Fei-Fei, 2017). However the different between this topic with image captioning is that image captioning mainly focus on individual image and this topic needs to handle with sequential photos. It would be more challenged that computer not only needs to understand the content of images but also need to find their inner relation.

2 Related Work

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 Zisser-

man, 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:

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

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:

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

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:

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

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

For the first paper in this field, the model Coherence Recurrent Convolutional Network has a simple idea to implement the task that generates corresponding sentences from image stream. They divide CRCN model in three parts, first is Bidirectional Recurrent Neural Network (BRNN) which learns image captioning crawling from user blog, second is Convolutional Neural Network (CNN) which learns image representation then transforms into 4,096 image VGG vectors and the third is local coherence model (Barzilay and Lapata, 2008) which learns sentence patterns and structures in order to combine several sentences as story or paragraph. For CNN part they directly use VGGNet (Simonyan and Zisserman, 2014), so in this approach, we mainly focus on Bidirectional Recurrent Neural Network (BRNN) and local coherence model.

Bidirectional Neural Network model The goal of BRNN is to learn corresponding sentences content and also consider previous and next sentence content as we already mention in BARNN approach. In BRNN model, there are five layers, input layer, forward layer, backward layer, output layer and ReLU activation layer. See the equation 7 below:

$$\begin{aligned} x_t^f &= f(W_i^f p_t + b_i^f) \\ x_t^b &= f(W_i^b p_t + b_i^b) \\ h_t^f &= f(x_t^f + W_f h_{t-1}^f + b_f) \\ h_t^b &= f(x_t^b + W_b h_{t-1}^b + b_b) \\ o_t &= W_o(h_t^f + h_t^b) + b_o, \end{aligned} \quad (7)$$

where p_t is 300-dimension embedding vectors from sentence as our input, x_f and x_b are forward and backward units activated input, h_f and h_b are forward and backward units output, o_t is in time the output of ReLU layer as the content of this sentence, f is ReLU activation function $f(x) = max(0, x)$ and W and b are weights and bias term.

The local coherence model In order to understand the sentence structure and content, they introduce local coherence model. The basic idea is that they make a table which contains discourse entity as row and sentence as column. The discourse entity grid is from sequential parse trees where extract from sequential text by using Stanford core NLP library (Manning et al., 2014). And then in grammatical aspect, they separate into four categories: S denotes subject, O denotes object, X denotes others except for subject or object, and $-$ denotes absent. Finally we computer their ratio by occurrence frequency and transform them into 300-dimension vector by zero-padding.

Combination of CNN, BRNN, and Coherence Model To combine BRNN and coherence model, they use two fully connected layers (FC) (Barzilay and Lapata, 2008). The two FC layers can combine both sentence content and sentence structure then map them into 4,096-dimension vector which is same dimension as image vectors. To compute sentence vector and image vector they use this score function 8 (Karpathy and Fei-Fei, 2015) as follow:

$$S_{kl} = \sum_{t=1...N} s_t^k \cdot v_t^l + g^k \cdot v_t^l, \quad (8)$$

where S_{kl} means score between image k and sentence l, v_t^l is CNN output of stream l in time t, s_t^k means the set of words s for image k, and g^k is the set of image fragments in image k (Karpathy et al., 2014). Finally, they create a loss function 9 to learn their Coherence Recurrent Convolutional Network model as follow:

$$\begin{aligned} C(\theta) = \sum_k \langle \sum_l \max(0, 1 + S_{kl} - S_{kk}) \\ + \sum_l \max(0, 1 + S_{lk} - S_{kk}) \rangle, \end{aligned} \quad (9)$$

where S_{kk} is positive image and sentence pair, S_{kl} and S_{lk} are negative image and sentence pair. Our first approach BARNN uses this loss function but they use γ instead of 1. However both of them are inspired by (Karpathy et al., 2014).

3.3 Adversarial Training

The model generating paragraph from stream photos with adversarial training is the newest model in this field to the best of my knowledge. The basic idea for this model is that they create a hierarchical story generator with Convolutional Neural Network (CNN) and Hierarchical Recurrent Neural Network (HRNN) to generate story from sequential photos inputting, and then they build two discriminators which critic determines whether the image and generated sentence are similar and critic determines whether the generated sentence is human-level story. Finally, by adversarial training, the model would be more and more powerful.

Hierarchical Story Generator To generate story-style narrative paragraph, they use train two RNN as encoder and decoder which both are built on Gated Recurrent Units (GRUs) (Chung et al., 2014). In order to train Hierarchical Story generator, they treat the generator as agent, and they build the loss function 10 as below:

$$\begin{aligned} L(\theta) &= -\sum_{n=1}^{N}\sum_{t=1}^{T} p_\theta(y_{n,t}|x_{1:n}; y_{n,1:t-1})R(y_n) \\ &= -\sum_{n=1}^{N} \mathbb{E}_{y_n \backsim p_\theta}[R(y_n)], \end{aligned} \tag{10}$$

where $R(x)$ is reward function which will be mentioned later, generator parameter θ defined a policy $p_\theta(y_{n,t}|x_{1:n}; y_{n,1:t-1})$, x means observed image and y represents generated sentence.

Multi-modal Discriminator This discriminator is to tell how similar between image and generated sentence is. The multi-model discriminator contains fusion mechanism is inspired by (Antol et al., 2015) and a fully connected layer followed by a softmax layer, the equation 11 see below:

$$\begin{aligned} v_{x_n} &= W_x \cdot x_n + b_x \\ v_{y_n} &= W_y \cdot LSTM_\eta(y_n) + b_y \\ f_n &= \tanh(v_{x_n}) \odot \tanh(v_{y_n}) \\ C_n^m &= softmax(W_m \cdot f_n + b_m), \end{aligned} \tag{11}$$

where x_n is image x, y_n is sentence, v_{x_n} denotes embedded image feature by linear layer, v_{y_n} denotes sentence vector by sentence inputting into Long-short Term Memory network, f_n denotes fused vector, C_n^m is our multi-model discriminator where $C_n^m(c|x_n, y_n)$ can be seen probability and $c \in paired, unpaired, generated$. For example, $C_n^m(paired|x_n, y_n)$ means what is the probability of image and generated sentence are paired.

Language-style Discriminator The other discriminator is language-style discriminator, which is aimed to check if the generated sentence is human-level understanding sentence. See the equation 12 as follow:

$$\begin{aligned} v_p &= LSTM_\phi(\overline{p}) \\ C^s &= softmax(W_p \cdot v_p + b_p), \end{aligned} \tag{12}$$

where v_p denotes the last hidden state as the encoded paragraph vector, $\overline{p}$ denotes paragraph embedding and C^s denotes the probability $C^s(gt|y)$ which means how similar with $groundtruth(gt)$. Finally, they create the reward function 13 defined below:

$$R(y_n|\cdot) = \lambda C_n^m(paired|x_n, y_n) + (1-\lambda)C_n^s(gt|y), \tag{13}$$

where λ denotes tradeoff parameter.

4 Experiment

5 Evaluation

6 Discussion

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Comput. Linguist.*, 34(1):1–34, March.

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.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, April.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 1889–1897, Cambridge, MA, USA. MIT Press.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. cite arxiv:1411.4555.

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.