Reverse Video Captioning with Attention

Wenqi Zheng
Syracuse University
wzheng11@syr.edu

Abstract

The explosive amount of data of video today is requiring more cost-efficient and effective way to generate the video captions automatically. Previous works has proven language decoder using Recurrent Neural Network(RNN) including Long short-term memory(LSTM) and Gated Recurrent Units(GRU) suitable for this task, however, these methods does not consider the nature dependency of words. In this paper, Reverse Video Captioning is proposed as an improvement of the decoding stage based on the characteristics of English language. By reversing the decoding direction, the result of video captioning could focus more on more important content of the generated sentences. Besides, an attention method is introduced to this model to improve the performance of the model. Extensive experiments on dataset YouTube-to-Text (MSVD) show this method is shown the effectiveness and improvement of the baseline method.

Introduction

Early video understanding tasks such as scene understanding and action recognition focus on generating a few labels for videos. After that, video captioning has been developed to generate more specific descriptions of videos. Two-dimensional Convolutional Neural Networks (CNN) have exhibited good performance in still image tasks such as classification or detection(Simonyan and Zisserman 2014). Video captioning methods were typically developed as an extension of image captioning (Vinyals et al. 2015; Karpathy and Fei-Fei 2015), aiming to linking video elements to words in a caption. However, video usually contains more information than images, for they no only contains different actions, but also contains multiple interactions between the objects shown up in the video. Such methods discard the important temporal information that has been shown to provide important cues in videos.(Wang et al. 2011)

Many methods have been proposed to utilize this feature.In (He et al. 2016), ResNet and and 3D ConvNets (Tran et al. 2015) have been introduced into video captioning to summarize motion information in short videos. The Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), which has been proved successful in natural language processing (NLP), has also been applied to capture the temporal information in video in addition to the sentence generation. Moreover, (Chung et al. 2014) has proven that some variants of Recurrent Neural Network (RNN) like Gated Recurrent Unit (GRU) (Cho et al. 2014; Kang, Zhang, and Liu 2016) are more suitable for short-sequence data as these structures are of fewer parameters and can avoid the suffer of overfitting. While most of the sequence-to-sequence video captioning model is generated in a forwarding way in both encoding and decoding process, it should be noticed that in English language, the dependency between words is not usually in from left to right. In contrast, in the language structure of the combination such as preposition and noun, article and noun, the dependency is actually from right to left. For example, in the phrase "on the playground", the word "the" depends more on the word "playground" than the other way around. We can see that the most center words in a sentence are often nouns, such as shirt, competition, cartoon. However, usually there are some auxiliary words such as in, on, blue, a, the before that. These words are not so explicit in the extracted features and usually they are attached to the center words, therefore, a generating process from left to right is not so helpful as that from right to left.

To solve this problem, we propose a reverse language decoder with attention for video captioning. It will generate the description of video in a reversed way to give more details about the video as a improvement based on the language feature of English. We further add a soft attention mechanism to achieve better performance. We conduct experiments on widerly-used video captioning datasets YouTube-to-Text (MSVD) and also compare with the tradition sequence-to-sequence method. The experiment results show that our method can achieve very competitive performance and produce better captions than the traditional model.

Related Work

Video captioning is a hot topic in computer vision and natural language processing communities and many methods have been proposed as we summarize below.

Template-based approaches. Early works in video captioning (Guadarrama et al. 2013; Krishnamoorthy et al. 2013; Thomason et al. 2014) treat the problem as a template-matching problem, focusing on generating video descriptions based on the identification of (subject, verb, object) triplets with visual classifiers. However, such the template-

based approaches have limited ability to generalize to unseen data, and the generated sentences cannot satisfy the richness of natural language.

Encoder-decoder based methods. Inspired by the success of deep neural networks in neural machine translation (NMT) and image captioning (Vinyals et al. 2015; Karpathy and Fei-Fei 2015), the recent video captioning works have moved to utilize RNNs, which, given a vectored description of a visual content, can naturally generate sequences of words (Vinyals et al. 2015; Karpathy and Fei-Fei 2015). The first work applying RNNs to video captioning is (Venugopalan et al. 2014), in which they proposed an encoder-decoder based framework for video caption generation. They used Convolutional Neural Network (CNN) to extract features from each single frame, and took an average pooling over all the video frame features to get the entire video representation, which is then fed into an LSTM network for the sentence decoding. However, their mean pooling operation ignores the sequential nature of videos. After that, many works have tried to improve the video encoding process. In (Donahue et al. 2015), they encoded the input video with another LSTM network and employed CRFs to generate a coherent video description. Venugopalan et al. (Venugopalan et al. 2015) encoded a video with a stacked LSTM network, which was later improved by incorporating attention mechanisms (Yao et al. 2015a) in the sentence decoder, or combining with external knowledge (Rohrbach, Rohrbach, and Schiele 2015).

Other enhanced methods. To generate more accurate captions, attribute detections are introduced into the model to detect semantic attributes of a frame. Each element in an attribute vector represents the possibility of a particular semantic attribute appearing in the frame. Pan et al. (Pan et al. 2017) proposed a video captioning model combined with semantic attribute detection. They introduced a transfer unit to merge the temporal information with image attributes and video attributes. Later on, Gan et al. (Gan et al. 2017) further improved the method by proposing an individual LSTM structure using the semantic information to guide the weight matrix in LSTM. With the new structure, the semantic information is involved in language decoding process rather than being merged with the hidden state to generate a fused vector.

Recently, Pasunuru et al. (Pasunuru and Bansal 2017) proposed a video captioning framework which requires multiple datasets for training. Particularly, they added a video prediction process and an entailment generation process into the video captioning task to learn better representations of both video and language. They involved the UCF-101 (Soomro, Zamir, and Shah 2012) action videos dataset to train the video prediction model and the Stanford Natural Language Inference (SNLI) corpus to train the entitlement generation.

Approach

In this section, we will explain the two main parts of the proposed method:(i) video feature extraction, (ii) language decoder. The whole framework is shown in Figure 1. The two proposed approaches work together into an encoder-decoder video caption generator. The details of the proposed method is to be explained in the following part.

Gated Recurrent Unit(GRU) In end-to-end sequence-to-sequence training tasks, GRU structure is often adopted. It takes an update gate r_t, an reset gate z_t to generate the state h_t in time step t:

$$r_t = \sigma(W_r v_{f_t} + U_r h_{t-1} + b_r) \tag{1}$$

$$\hat{h_t} = \phi(W_h v f_t) + U_h(r_t \odot h_{t-1}) b_h) \tag{2}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{3}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \hat{h_t} \tag{4}$$

where σ is sigmoid function; ϕ is tanh as the activation function; $\odot$ denotes the element-wise multiplication; $W_r, W_r, W_h, W_z, U_r, U_h, U_z, b_r, b_h, b_z$ denotes the model weights and bias to be trained.

Video feature extraction As shown in Figure 1, a pre-trained CNN is used to encode each video frame $f_i \in \mathbf{f}$ into a feature vector v_{f_i} yielding video features $\mathbf{v}_f = (v_{f_0}, \cdots, v_{f_{N-1}})$. Let $\mathbf{f} = (f_0, \cdots, f_{N-1})$ be a sequence of video frames, where f_i is the image at time step i, and N is the length of the sequence. The video captioning task is to generate a text $\hat{\mathbf{Y}} = \{\hat{Y}_0, \cdots, \hat{Y}_{T-1}\}$ that describes the video, where $\hat{Y}_t \in \mathcal{D}$ is a word predicted from a vocabulary $\mathcal{D}$, and T denotes the number of words in the description. We obtain the image features v_{f_i} by performing a mean-pooling over the spatial image features in the final convolutional layer of a CNN that is pre-trained on a large image recognition dataset. Figure 1 illustrates our caption generation model from a given sequence of image features $\mathbf{v}_f = (v_{f_0}, \cdots, v_{f_{N-1}})$. The spatial image features are encoded into a temporal sequence of feature representations $\mathbf{h^v} = (h_0^v, \cdots, h_T^v)$ with a GRU-based RNN:

$$h_i^v = GRU(v_{f_i}, h_{i-1}^v) \tag{5}$$

Language decoder To change the dependency of the words while decoding, we propose a language decoder with reversed decoding direction. This is implemented by simply reverse the ground truth sentence from end to begining. The feature representations $\mathbf{h}^v = (h_0^v, \cdots, h_T^v)$ are decoded into the caption using a language model with a visual attention layer in (Yao et al. 2015b) that learns the ability to focus on subsets of frames to in order to capture global temporal structure.

Attention:

$$e_i^t = W_v h_i^v + W_l * h_t^l + b \tag{6}$$

$$a^t = softmax(e^t) \tag{7}$$

$$H^t = \sum_{i=1}^{k} a_i^t h_i^v \tag{8}$$

where k is the total frames of the video.

Decoding:

$$h_t^l = GRU([w_{t-1}, H^t], h_{t-1}^l) \tag{9}$$

$$Y_i = argmax(W_w h_t^l + b^l) \tag{10}$$

$$w_t = \begin{cases} embed(< bos >) & \text{t=0} \\ embed(Y_t) & \text{t>0} \end{cases} \tag{11}$$

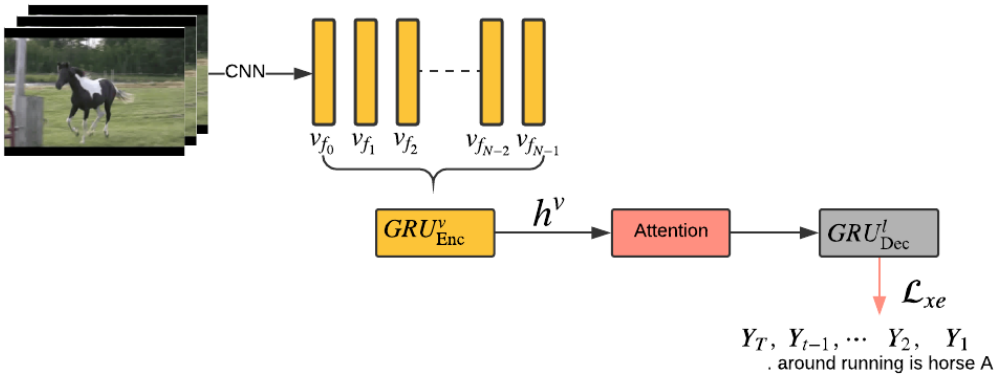

Figure 1: The main frame work of our method. Noted that the ground truth is reversed in order to focus on the center word.

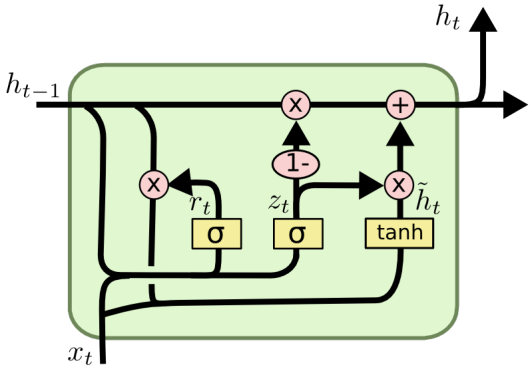

Figure 2: The structure of the Gated Recurrent Unit.

Training

Given the video features $\mathbf{v}_f = (v_{f_0}, \cdots, v_{f_{N-1}})$ and the ground-truth caption $(Y_0, Y_1, ..., Y_{T-1})$, the language decoder is conditioned step by step on the first t words of the caption (already generated) and on the corresponding video descriptors in $\mathbf{v}_f$, and is trained to produce the next caption word. We train the caption generation process by minimizing the cross entropy (XE) loss:

$$\mathcal{L}_{\text{xe}} = -\sum_{t=0}^{T-1} \log p_{\theta_t}(Y_t|Y_{0:t-1}, v_{f_{0:N-1}}; \theta) \qquad (12)$$

where $p_{\theta_t}(Y_t|Y_{0:t-1}, v_{f_{0:N-1}})$ is the output probability of the predicted word Y_t with the model parameters θ. We share the weights of the model across all time steps.

Experiments

Datasets and Metrics

Microsoft Video Description Corpus (MSVD) (Chen and Dolan 2011). MSVD contains 1,970 Youtube video clips, 85K English descriptions collected by Amazon Mechanical Turkers. As done in previous works (Guadarrama et al. 2013; Venugopalan et al. 2014), we split the dataset into contiguous groups of videos by index number: 1,200 for training, 100 for validation and 670 for testing.

Table 1: Comparisons of the video captioning results of the proposed method and the baseline method on the MSVD dataset.

Model	BLEU 4	METEOR	CIDEr
Enc-Dec(Basic)	38.7	28.7	44.8
Enc-Dec + SA + Reverse	45.1	31.1	66.0

Evaluation Metrics Three popular evaluation matrics are adopted to evaluate the result: BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). BLEU is a precision-based metric, calculated by matching the n-grams in candidate and reference captions. METEOR evaluates the captions by matching n-grams in different forms such as surface, stem and paraphrase found in the generated caption and in references. CIDEr measures consensus in video descriptions by performing a Term Frequency-Inverse Document Frequency weighting for each n-gram.

Results

To further explore the effectiveness of our model, we compare the result with a baseline model, which is a conventional encoder-decoder video captioning model with GRU as the video encoder and GRU to decode langauge in a forwarding way. The model is trained using the same loss function and errors in Eq. (12). Table 1 shows the comparison results of the baseline method and our method on the MSVD dataset. Figure 3 gives some examples of the generated captions. It can be seen that the proposed model provides more specific and accurate contents about the video. In the first example, out model is able to give the key word “competition”,

Video 1. GT1: 2 men are in a wrestling match GT2: a group of boys are wresting two on two in matches GT3: boys are wrestling in front of a crowd GT4: there are two people wrestling on the floor GT5: two boys wrestling matches are shown	
Enc-Dec(Basic): two men are wrestling **Enc-Dec + SA + Reverse** : two men are wrestling in a competition	
Video 2. GT1: Ingredients are put in a pan with oil to boil it GT2: a person is cooking a egg curry on a fry bowel GT3: a man is using a mental spatula to stir around eggs in a hotwok he is using on the stove GT4: an egg is smashed in a tawa and the rice noodles are added	
Enc-Dec(Basic): there is a man is making a dish in the kitchen **Enc-Dec + SA + Reverse** : a person doing a cooking show and mixing the ingredients for the recipe	
Video 3. GT1: an animate character are talking GT2: a cartoon character is jumping from flower to flower GT3: a cartoom charactor jumps on flowers GT4: a cartoon involving animals GT5: a cartoon pig jums between flowers	
Enc-Dec(Basic): a cartoon is being played **Enc-Dec + SA + Reverse** : cartoon characters are dancing in a cartoon	

Figure 3: Examples of the generated captions by our final model and a baseline.

while the baseline model can only return a simpler answer. The noun after the word "the" or "a" is hard to give a stable predict using forward language decoder, while in the second example, out method successfully gives the word "recipe" after the word "the" correctly.

Conclusion

In this paper, we have proposed a video captioning framework that integrates a reverse language decoder and a shared soft attention module. Test results shows that is method achieves better accuracy than the baseline method and effective in generating detailed information.

References

Chen, D. L., and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 190–200. Association for Computational Linguistics.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *ACL*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; and Deng, L. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R.; Darrell, T.; and Saenko, K. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2712–2719. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*.

Kang, J.; Zhang, W.-Q.; and Liu, J. 2016. Gated recurrent units based hybrid acoustic models for robust speech recognition. In *Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on*, 1–5. IEEE.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Krishnamoorthy, N.; Malkarnenkar, G.; Mooney, R. J.; Saenko, K.; and Guadarrama, S. 2013. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, volume 1, 2.

Pan, Y.; Yao, T.; Li, H.; and Mei, T. 2017. Video captioning with transferred semantic attributes. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Pasunuru, R., and Bansal, M. 2017. Multi-task video captioning with video and entailment generation. *arXiv preprint arXiv:1704.07489*.

Rohrbach, A.; Rohrbach, M.; and Schiele, B. 2015. The long-short story of movie description. In *German Conference on Pattern Recognition*, 209–221. Springer.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *ICLR*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Thomason, J.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; and Mooney, R. J. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Coling*, volume 2, 9.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, 4489–4497. IEEE.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; and Saenko, K. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, 4534–4542.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, H.; Kläser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 3169–3176. IEEE.

Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015a. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, 4507–4515.

Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015b. Video description generation incorporating spatio-temporal features and a soft-attention mechanism. *arXiv preprint arXiv:1502.08029*.