FDU Participation in TRECVID 2019 VTT Task

Shaoxiang Chen, Yu-Gang Jiang
Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
Shanghai Insitute of Intelligent Electroics & Systems
{sxchen13,ygj}@fudan.edu.cn

Abstract

This notebook paper presents the system design of the FDU team in the TRECVID 2019 [1] VTT task. Our approach adopts temporal concept prediction as an auxiliary task to assist caption generation. The concept prediction module generates a context sequence that contains latent semantic features, which are later fused into the captioning module. We demonstrate the effectiveness of our designed auxiliary task as well as the whole captioning system.

1. Introduction

The TRECVID VTT task (video description generation) asks the participants to generate one sentence to describe a video in the testing dataset. The dataset contains 1,010 videos from Flickr and 1,044 videos from Vine. Video description (captioning) is an important task in the computer vision literature. Popular methods [5, 10, 11] of video description can generally be divided into two sub-tasks: video representation learning and language generation. In recent years, there are suitable deep neural networks for both of these sub-tasks, namely CNN (Convolutional Neural Network) for learning high-level visual representation and RNN (Recurrent Neural Network) for sequence generation. Our system design follows the approach of [11], with an additional module that learns visual concepts in the video to boost the visual representation that's fed to the captioner.

2. Our System

2.1. Visual Representation

CNN learns high-level visual representation by learning to recognize the content of images. The quality of CNN visual representation depends on the architecture (depth, connection types) of the CNN and the scale of training dataset it is trained on. For better performance, we choose the Inception-Resnet-V2 [8] CNN pre-trained on the ImageNet-1M [3] dataset. The visual representation is a vector of length 1,536 and is the activation of its last pooling layer computed with a video frame as input into the CNN. Thus the whole video is represented as a sequence of vectors, denoted as $\{v_1, ..., v_N\}$.

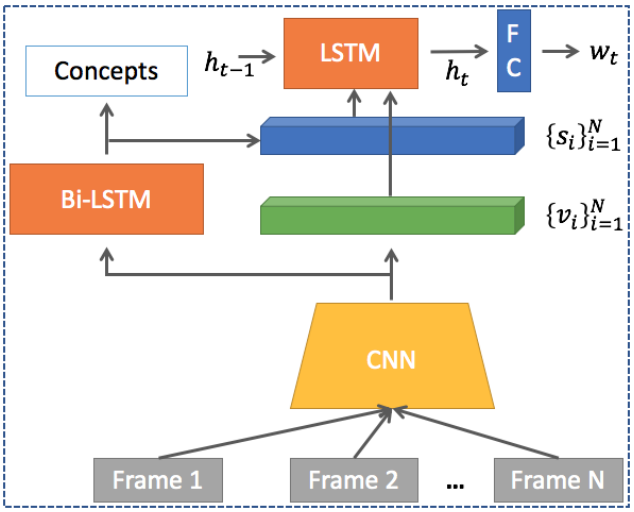

Figure 1. The system design of our approach.

2.2. Concept Prediction

Although the visual representation extracted from CNN is powerful, there is still a large domain gap between the feature representation domain and the text domain. Thus we designed a module that predicts the visual concepts (such as cat, vegetable and bowl) in the video to close this domain gap. As shown in the Figure 1, concept prediction is an auxiliary task in the video captioning system, and through this task the module learns a latent semantic representation that can be fed to the captioning module to assist caption generation. The visual representation sequence is encoded by a bi-directional LSTM and then a fully connected layer is used to predict concepts in the videos. The concept label is extracted from the sentence annotations. The latent semantic representations are also a vector sequence $\{s_1, ..., s_N\}$ which is concatenated with the visual representation.

2.3. Sentence Generation

The goal of this module is to generate a sequence of words $\{\hat{w}_1, \hat{w}_2, ..., \hat{w}_T\}$ one-by-one. We use a standard LSTM unit that takes an input (aggregated) visual representation at each time step t, and outputs a hidden state that's later projected into a word distribution. We adopt the dynamic temporal attention mechanism to aggregate the representations when generating each word. It first computes a set of attention weights for each element of the visual representation sequence.

$$a_i^t = W_a tanh(W_v[v_i, s_i] + W_h h_t + b), \tag{1}$$

$$\alpha_i^t = \frac{e^{a_i^t}}{\sum_{i=1}^{N} e^{a_i^t}}, \tag{2}$$

where the Ws and b are learnable parameters. The the feature aggregation is a weighted-sum of the feature sequence with $\{\alpha_i^t\}_i^N$ as the weights.

$$C_t = \sum_{i=1}^{N} \alpha_i^t [v_i, s_i]. \tag{3}$$

The C_t is then fed to LSTM for word generation.

$$\begin{aligned} h_t &= \text{LSTM}(h_{t-1}, C_t), \\ w_t &= W_w h_t + b_w. \end{aligned} \tag{4}$$

The final word prediction $\hat{w}_t$ is then the one with maximum probability (likelihood):

$$\hat{w}_t = \arg\max_i w_t. \tag{5}$$

The model training is then performed by maximizing the log-likelihood of the groundtruth words w.r.t the model parameters.

3. Experiments

We train our model on the TGIF [6] dataset, which contains 100K videos and 120K sentence annotations. The description performance is evaluated by common metrics such as BLEU [7], CIDEr [9] and METEOR [4]. We use the TREVID 2017 test set as our validation set. We first present our results in the validation set and compare with TRCVID 2018 winner's approach.

Method	BLEU@4	METEOR	CIDEr
2018 Winner [2]	8.06	13.85	32.53
Ours w/o concept pred.	7.56	12.75	31.12
Ours	8.04	13.23	32.00

Table 1. Experimental results on the TREVID 2017 test set.

As shown in Table 1, it is clear that concept prediction is very helpful in video captioning, and our final system is comparable with the winner of TRECVID 2018 on the validation set.

Method	BLEU@4	METEOR	CIDEr
Ours	2.7	24.4	42.8

Table 2. Experimental results on the TRECVID 2019 test set.

Our system's performance on the TRECVID 2019 test set is shown in Table 2. There is in fact only one submitted run for our system, and our final rank in the TRECVID 2019 VTT task is 3.

References

[1] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Qunot. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In *Proceedings of TRECVID 2019*. NIST, USA, 2019.

[2] Jia Chen, Shizhe Chen, Qin Jin, and Alexander Hauptmann. Informedia@trecvid 2018 video to text description. In *TRECVID 2018*, 2018.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[4] Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, 2014.

[5] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *ICCV*, 2017.

[6] Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*, 2016.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[8] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[9] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[10] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. Learning multimodal attention LSTM networks for video captioning. In *ACM MM*, 2017.

[11] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.