Video Description

Project Report, Group 9, 8 Nov 2016
Ishita Ankit (13316), Lakshay Garg (13373), Shubham Jain (13689)
CS698N: Recent Advances in Computer Vision, Jul–Nov 2016
Instructor: Gaurav Sharma, CSE, IIT Kanpur, India

1 Problem Definition

In recent times, video has become one of the most important forms of communication. With the advent of video sharing sites like YouTube, hundreds of hours worth of video content is produced every minute. Cameras are also ubiquitous in modern security and surveillance systems. Such a huge amount of video content requires us to automate our processes to manage humongous amounts of information. The problem of describing videos is one of the long-standing goal of computer vision research. Being able to generate automatic descriptions of videos enables us to index large and heterogeneous collection of video material, automatically generate video summaries and is a step closer to the ultimate goal, automatic understanding of video content.

Motivated by the present state of video description techniques and their increasing need, we proposed to tackle the challenge of generating video descriptions and developing a method which could incorporate knowledge from sources other than the paired training data to improve the prediction performance.

2 Literature review

The problem of description is not a new one and the description quality of proposed methods has improved over time. Video description and the closely related task of activity recognition has been approached by learning frame-to-frame representations augmented with optical flow [11]. CNNs have been used to learn temporal dependencies using convolutions in time [9]. The problem has been modeled as a problem of machine translation and approached using sequence learning methods like RNNs and the more recent LSTMs [14, 7]. Recent work in related fields such as image captioning [6, 16], visual question answering [17] has demonstrated improved performance by incorporating knowledge from external sources.

3 Our Approach

We develop a method to generate video descriptions using external knowledge by combining two state-of-the-art methods for sequence learning [15] and image captioning using external knowledge [6]. The paper by Hendricks et al. [6] incorporates external knowledge in the task of image captioning. It does so by dividing the task of captioning into three models: lexical model, language model and the multimodal unit. The lexical model is trained on large object recognition datasets and is a convolutional neural network used to extract visual features from images. The language model is an LSTM based model which learns to predict the next word given earlier words. These two models are combined into a single framework by introducing the multimodal unit which takes into account the output of these models and learns an embedding which is then used for generating image captions.

Venugopalan et al. [15] have exploited the sequence modeling capabilities of LSTMs and used it to translate one sequence to another. They have demonstrated its ability by using it for generating video descriptions. In this project, we have tried to combine these two models into a unified framework for generating video descriptions.

The model that we worked on can be subdivided into three broad sections: the sequence model, the language model and the fusion unit. The architecture of our network is shown in Figure 1.

3.1 Sequence Model

The sequence model is an adaptation of the architecture presented in [15]. Specifically we take the video encoding stage of the S2VT model and use it as a video feature extractor in our model. The CNN used here is the vanilla VGG-16 [12] network trained for object detection task on the ImageNet [4] dataset. We take the output of its pre-softmax layer and pass it on to the LSTM. The outputs of the LSTM are mean-pooled and combined into a single vector which is passed on to the fusion unit. Since the model contains a CNN for extracting frame features which are then passes to a sequence modeling LSTM, we expect the output features of the sequence model to account for both the spatial and temporal information in the input.

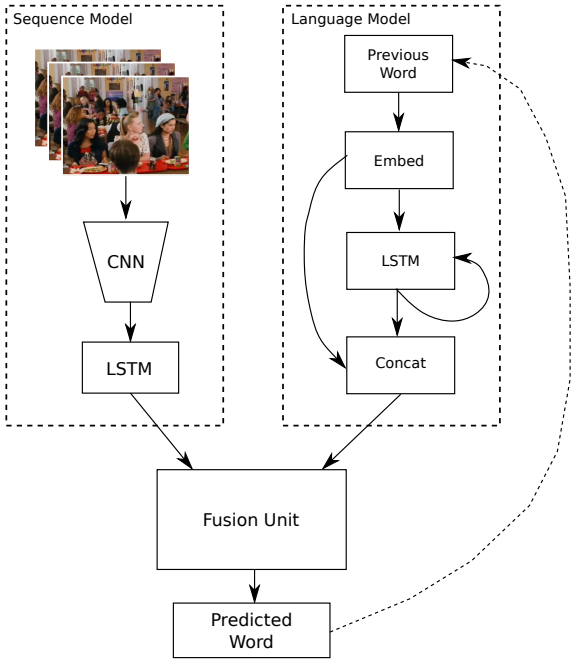


Figure 1: Overview of the proposed architecture

3.2 Language Model

The language model used in our model has been borrowed from [15]. The language model takes the word as a one-hot vector which is encoded using Word2Vec [10]. This representation is passed to a LSTM which learns to predict next word given earlier words in the sequence. The final word generated by the fusion unit is passed back into the language model. The language model is trained on large text corpora such as the British National Corpus [3] and Wikipedia.

3.3 Fusion Unit

The fusion unit is an adaptation of the multimodal unit in [15]. This unit takes the video and language features from sequence and language models and learns to embed them in a shared space. The probability of word is predicted using $p = \text{softmax}(W_V f_V + W_L + f_L)$. The parameters W_V and W_L are video and language features learnt by training this model on paired video caption dataset. Further, we employ the direct transfer method presented in [6] for transferring knowledge to the fusion unit.

4 Implementation

The basic structure of our model consists of extracting features from the S2VT model [15] and feeding it into DCC's multimodal unit in place of image features. We implemented our networks using the Caffe [8] deep learning framework. The entire approach is summarized below:

- The S2VT model concatenates features extracted from every frame (sampled at 5 frames per second) and passes them through a double-stacked LSTM architecture to finally predict the caption for the video.
- The DCC model takes as input a feature representation from the lexical model and passes it into the multimodal unit which generates a caption by utilizing the language model.
- To connect the S2VT and DCC models, we mean-pooled the output of LSTM in the S2VT model which was then passed to DCC.

The presented model is not end-to-end and needs to be trained in stages. The sequence model was taken directly from [15] and the language model was taken from [6]. To train the fusion unit, we mean-pooled features from the sequence model before and then passed it to the fusion unit. To make the training and testing process faster, we extracted the VGG-16 features for the training and validation sets of the Microsoft Video Description dataset [2]. These features were separately passed into the LSTM and mean-pooled.

5 Results

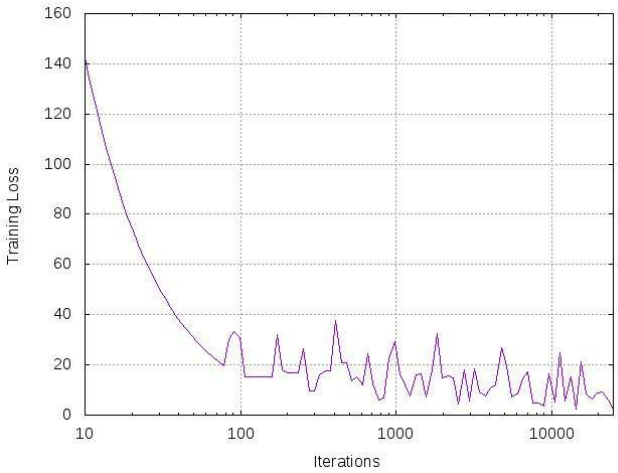
5.1 S2VT (Trained on entire training dataset)

We trained the baseline S2VT model on the entire Youtube training data and tested on the test data to obtain close to pretrained model results. The results are mentioned in the table below:

Model	CIDEr	Bleu_4	Bleu_3	Bleu_2	Bleu_1	ROUGE_L	METEOR
Pretrained	0.515	0.367	0.476	0.588	0.735	0.651	0.293
Trained : 25000	0.483	0.325	0.436	0.554	0.712	0.629	0.286
Trained : 11000	0.501	0.342	0.453	0.642	0.642	0.642	0.287

5.2 S2VT-DCC Fusion Model

The features from S2VT were fed into the DCC model to enhance the captioning results by incorporating the information retrieved from unpaired text. We successfully trained the fusion model but were unable to evaluate it due to a bug in the deployed network which we haven't been able to track yet. Since the model was made by combining the S2VT and DCC model definition files, we believe that the bug is due to mismatch in the dimension of feature tensor which had to be passed to one of the layers.



5.3 Experimenting with audio

We evaluated the results obtained for activity recognition using audio.

- The random forest model attained an accuracy of 62.43% when evaluated on 7 classes namely : laugh, cry, clap, breathing, sneeze, singing, whistling.
- 630(90 per class) audio files obtained from Freesound were used for training and testing was done on 70 audio files.
- Best accuracy was obtained for an ensemble of 10 classifiers and length of Bag of Audio Words (BoAW) features as 70.

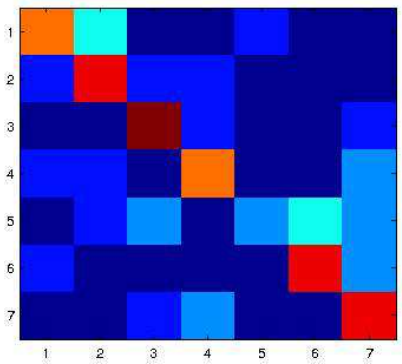


Figure 1: The figure on left shows the confusion matrix obtained by using an audio based classifier on the dataset obtained from Freesound. The figure on right is a from a YouTube video, the vanilla S2VT model classifies it as talking whereas audio classifier classifies it as laughing. We therefore believe that incorporating audio into video description models can improve performance

References

[1] *Daily Activity Recognition Based on DNN Using Environmental Sound and Acceleration Signals*. Zenodo, Aug. 2015.

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, June 2011.

[3] J. H. Clear. The digital word. chapter The British National Corpus, pages 163–187. MIT Press, Cambridge, MA, USA, 1993.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

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

[6] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. J. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. *CoRR*, abs/1511.05284, 2015.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[11] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014.

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

[13] J. A. Stork, J. Silva, L. Spinello, and K. O. Arras. Audio-based human activity recognition with robots. In *International Conference on Social Robotics (ICSR'11)*, Amsterdam, The Netherlands, 2011.

[14] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[15] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[16] Q. Wu, C. Shen, A. van den Hengel, P. Wang, and A. R. Dick. Image captioning and visual question answering based on attributes and their related external knowledge. *CoRR*, abs/1603.02814, 2016.

[17] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.