CS231n Project Proposal

Jiajun Sun, Jing Wang, Ting-chun Yeh

April 26, 2017

Problem Description

In *Deep Learning for Video Classification and Captioning*, Zuxuan Wu, Tian Yao, Yanwei Fu and Yu-Gang Jiang state that video has been a new method for Internet users to spend passtimes. Increasing users and their need generate tremendous amount of data. Thus, deeper video understanding application is encouraged.

Video classification is the main focus in this project. Video captioning is also considered if time is available. Compared to image classification, long-time and sequential image classification problem on video classification is a new challenge. In *Beyond Short Snippets: Deep Networks for Video Classification*, Joe Yue-Hei Ng, and the other authors proposes that obvious accuracy improvement has been achieved by CNN combined with LSTM compared to pure CNN with max pooling architectures.

In our project, video datasets from Microsoft Multimedia challenge will be applied to train and test. Our baseline is going to adopt CNN to classify each image of video and get the final video type by majority vote. Moreover, Better performance might be seen from improved version: CNN + LSTM. With respect to the video captioning, a new LSTM after video classification can be a solution.

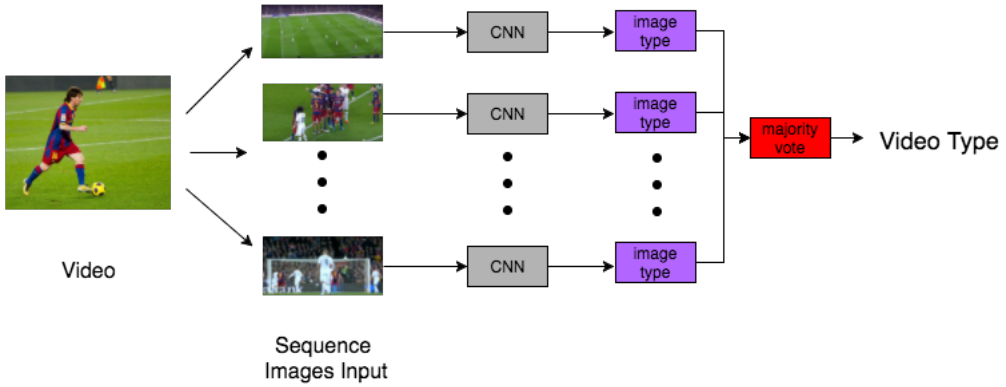
Datasets

The video description dataset is provided by Microsoft Multimedia Challenge. This dataset is based on MSR-VTT(A Large Video Description Dataset for Bridging Video and Language) and it has been split into training, testing and validation set according to 65%:30%:5% split ratio (Xu, J. et al, 2016).

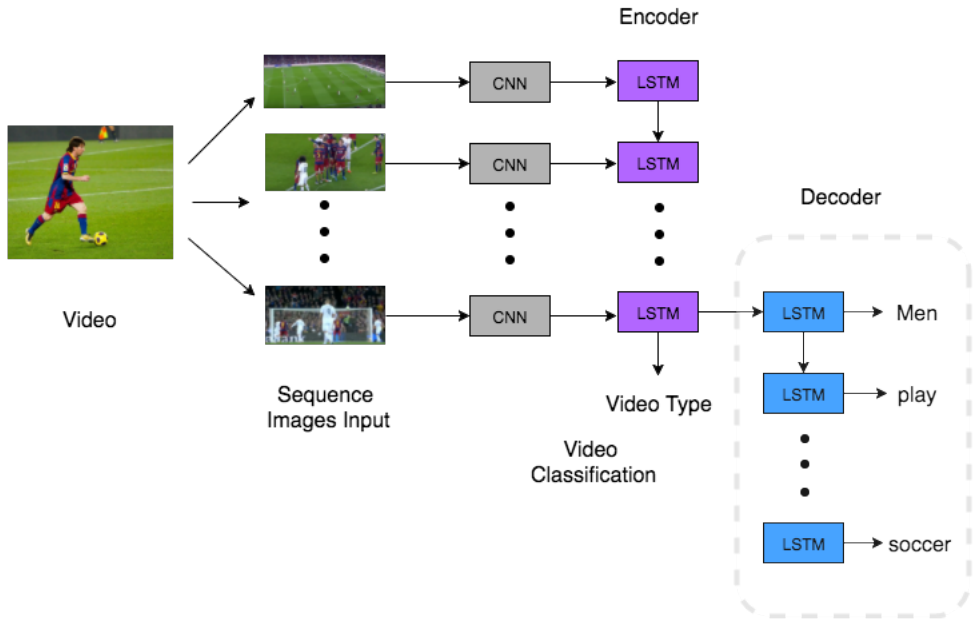
This dataset in total has 10,000 video clips with a total duration of 41.2 hours. The video clips come from 20 different categories. Annotated description has in total of 200,000 sentences (Xu, J. et al, 2016). Labels for this dataset, including clip category and annotated description, are generated by AMT workers. Each clip comes with 20 natural sentences by 1327 AMT workers (Xu, J. et al, 2016).

Methodology

In order to show the advantages of combined model: CNN + LSTM, the pure CNN for sequential images input is considered as the baseline in our project. The basic idea here is that the video type could be obtained from majority vote since every image can be assigned a type by CNN. The framework is as follows:



Since video input is the sequential images, LSTM is considered to combine with CNN to classify the video. Each LSTM receive the output matrix from CNN and the output of previous LSTM cell. This series of LSTM cell is the encoding layer. Output of the encoding layer contain all the sequential information. In order to produce description sentences, a decoding layer is used here. This encoder-decoder structure is similiar to the architecture used in Machine Translation task. The decoder will keep outputing words until it produces a stop token. The framework is shown as below:



Evaluation Metrics

At the first stage of this project, F1 score is applied to evaluate video classification accuracy. We will also use confusion matrix to plot out our accuracy. The confusion matrix can help us the visualize our accuracy and easily identify any bias.

At the next stage, We will use METEOR as a metric to evaluate our annotation result. METEOR is a machine translation scoring metric by aligning generated description with one or more reference description. We will compare our result with state of the art result. CMU & RUC has achieved a score of 0.282 on the same dataset in 2016. We will print out some of the results and manually compare them with natural description to identify where might the error come from.

Reference

Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Wu, Z., Yao, T., Fu, Y., & Jiang, Y. G. (2016). Deep Learning for Video Classification and Captioning. arXiv preprint arXiv:1609.06782.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5288-5296).

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4694-4702).