CSC 400 ITRG Mini-proposal

Understanding Videos with Low Shot Learning

Wei Xiong
wxiongwhu@gmail.com
University of Rochester
September 12, 2016

1. Project Description

As the development of multimedia, a huge amount of videos have been produced on the web every day. Understanding the contents and motion patterns of videos automatically by computer systems are becoming increasingly crucial in many tasks, including video classification [1], unconditional video generation [2], future prediction [3] and robotics [4]. However, learning from the videos is a very challenging task for the two major reasons. First, most of the videos on the web are unlabeled. They are usually created without any detailed descriptions, or precise tags for the contents. Some videos may be tagged as a whole, but miss annotations for each clip of scenes. Second, the videos are usually noisy. The same contents can repeat for several times in a video. Some videos contain clips that are totally blank or show no meaningful scenes. The labels or descriptions can also be fake or containing incorrect information. All these obstacles prevent us from discovering useful patterns in videos.

To tackle these problems, we propose to understand the videos through low shot learning [5]. Low shot learning methods are a set of machine learning methods that can learn significant features from very limited number of training samples, or even only one sample per category. We propose to perform and develop low shot learning methods for two major tasks. The first is video classification. Video classification is a task to assign correct labels to a given video according to its contents. The second task is future prediction. Given one or a few starting frames, we will generate the long-term future frames automatically. Previous work [6-7] aiming to solve these problems usually require large quantities of labeled data. For video classification tasks, the conventional methods need correct labels that can best describe the video content. For future prediction, the videos are usually precisely clipped and no much noise data are allowed before sending to the conventional methods. However, labeled video data are extremely hard to obtain, making the previous work impractical to some extent.

We make attempts to correctly classify the videos and generate future scenes with very few training data. Some prior work have been done to tackle the small data problem, but lead to unsatisfactory results. Our goal is to develop methods based on low shot learning, annotating the videos with comparative accuracy with the strongly supervised methods, as well as generating future videos that are as realistic as the methods that are trained on a large amount of data.

2. Broader Impacts

The result of this project can be used for various applications. Since our goal is to learn from limited data, or even one sample, we will make the procedure of learning features from videos become easier, and also reduce the training time. As a consequence, the efficiency of training the vision systems of an automatic driving car will be greatly improved. For an ideal result, the automatic driving car is even able to learn and understand the traffic environment in real-time. Another important application is robotics. If the robot can train itself with a single look at the current scenes, it will become intelligent enough to deal with the current affairs.

3. Methods

3.1 Video Classification

For most vision areas, the data is limited. But there are some public large scale video datasets. We adopt the transfer learning mechanism, first train a model on a large labelled video dataset, then transfer the features learned from the dataset to the target videos. In this way, we are able to classify videos with a few target data with annotations.

3.2 Future Prediction

We use the generative adversarial network [8] to learn motion patterns from a large amount of unlabeled videos. We don't require the videos to be very clean. We don't need to annotate the video clips manually. Then we adopt low shot learning methods to enhance the performance of the generative adversarial network.

4. Research Plan

a) We first collect a video dataset, which contains raw videos without dedicately clipping or other processing. Less than 10% of the data will be manually annotated by us.
b) We use low shot learning methods to accomplish the video classification task on our video dataset.
c) We use combine the low shot learning methods and generative methods to predict the future of given scenes.
d) We evaluate our method by compare it with other low shot learning methods. We also compare our results with the models trained on strongly labeled video data.

Reference

[1] Karpathy, Andrej, et al. Large-scale video classification with convolutional neural networks.

Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.
[2] Vondrick, C.; Pirsiavash, H.; and Torralba, A. 2016. Generating videos with scene dynamics. In Advances In Neural Information Processing Systems (NIPS), 613–621.
[3] Mathieu, M.; Couprie, C.; and LeCun, Y. 2015. Deep multiscale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440.
[4] Ruvolo, P. 2017. Dude, where's my robot?: A localization challenge for undergraduate robotics. In AAAI Conference on Artificial Intelligence (AAAI), 4798–4802.
[5] Fei-Fei, Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*28.4 (2006): 594-611.
[6] Villegas, R.; Yang, J.; Zou, Y.; Sohn, S.; Lin, X.; and Lee, H. 2017b. Learning to generate long-term future via hierarchical prediction. International Conference on Machine Learning (ICML).
[7] Vondrick, C., and Torralba, A. 2017. Generating the future with adversarial transformers. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[8] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Advances In Neural Information Processing Systems (NIPS), 2672–2680.