CS 8803DL: Deep Learning Project Proposal

Casey Battaglino,[*] Min-Hung Chen,[†] Chih-Yao Ma,[‡] and Hao Yan[§]

1 Proposal

Motivation

Deep convolutional neural networks are a leading technique for image classification. However, *video* classification is less developed. While it is certainly possible to simply run a neural network on each individual frame of a video, this sacrifices a wealth of temporal attributes such as movement, gestures, gait, etc. There are indeed methods of preprocessing temporal information into a single 2-dimensional input [1], but a more attractive research goal is to develop a neural network that can discover temporal relationships on its own. There are multiple recent approaches towards this goal, but as of yet no consensus on which is superior.

The difficulties in training neural networks on video input include the following: memory requirements (particularly if 3D convolutions are used), fewer public data sets, size of the data sets (for instance, the Sports-1M data set is $\mathcal{O}(()\,1TB)$ large), and lack of consensus on which approach is most effective.

Our goal is to develop a deep neural network that classifies videos. We have settled on two main approaches, discussed below: 3-dimensional convolutional neural networks (3D-CNNs) and recurrent neural networks that incorporate long short-term memory (LSTM).

Related Work

One popular approach is to apply a 2D-CNN to each frame of the video, followed by an RNN with LSTM [3].

Large-scale video classification: [4]

Learning Spatio-temporal features: [6]

Beyond short snippets: [5]

[*] cbattaglino3@gatech.edu

[†] cmhungsteve@gatech.edu

[‡] cyma@gatech.edu

[§] yanhao@gatech.edu

Approach and Techniques

We have decided on combining two possible approaches, which address the third dimension (time) in different ways: 3D Convolutional Neural Networks (3D-CNNs) [6, 4] and Multi-Dimensional Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) [3].

Using a 3D-CNNs, it is possible to build a classifier on a moving window of frames, which extract useful local movement information shared in a few frame. We would like to use the slow fusion model [4] due to the GPU memory constrain. On the other hand, LSTM/ RNNs is able to build the classifier with long term memory using much larger number of the frames in the video. Combining both 3D-CNNs and LSTM is able to combine the local movement information with long-term memory, which can further increase the classification accuracy.

Furthermore, the current LSTM model is built on the fully-connected layer, which no spatial information is preserved. We would like to try to plug-in the LSTM at earlier layers to test its performance. Especially, if applying LSTM directly after the convolution layer, the input to LSTM model is actually multi-dimentional tensor. To address this challenge, we would like to apply the multi-dimensional LSTM [2] to efficiently take advantage of the spatial information after the convolution layer.

Data Set

Describe the video data set...

Experimental Methodology

For training we plan to use the Jinx cluster at Georgia Tech. Each node is equipped with 2 nVidia Tesla M2090 “Fermi” GPU cards, and CPU nodes with large memory are available.

- Move 2D-LSTM into CNN
- Replace 2D-CNN → 2D-CNN with slow fusion
- Replace 2D-CNN → Shallow ResNet
- Replace 2D-CNN → Shallow ResNet with slow fusion
- 2D-CNN incorporated with Optical Flow *and* slow fusion

Group Tasking

- Proposal writing - Steve, Hao, Casey
- Proposal presentation - Yao
- Read and understand papers on 2D-LSTM - Everyone

- Read and understand ResNet papers - Everyone
- Implement LSTM: Yao, Casey
- Implement CNN: Steve, Hao
- Preparation of UCF data (incl. optical flow) - Casey
- Preparation of Sports-101 data - (? . . . tentative)

References

[1] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, 2011.

[2] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3547–3555, 2015.

[3] 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. *CoRR*, abs/1411.4389, 2014.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1725–1732, Washington, DC, USA, 2014. IEEE Computer Society.

[5] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *CoRR*, abs/1503.08909, 2015.

[6] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *CoRR*, abs/1412.0767, 2014.