CALIFORNIA STATE SCIENCE FAIR
2016 PROJECT SUMMARY

Name(s)

Mihika Nadig

Project Number

36466

Project Title

Application of a Deep Learning Architecture Using Convolutional and Recurrent LSTM Networks to Video Classification

Abstract

Objectives/Goals

Currently, the use of neural networks has been restricted to image classification. Video classification has been a challenge due to the difficulty of fusing spatial-temporal information. My project aims to construct a deep learning architecture involving convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with LSTMs as a viable solution for video classification.

Methods/Materials

To perform feature computation, two CNNs are used to process raw individual frames and optical flow images from the UCF-101 Video, CIFAR-10, and Microsoft COCO Datasets. After filter learning, feature aggregation is performed via feature pooling or RNNs with a deep LSTM architecture. Three feature pooling architectures were experimented with: Conv Pooling, Intermediate Pooling, and Neighborhood Pooling. A softmax classifier, the final layer in the feature pooling architecture, returns the prediction. Training is done by coupling stochastic gradient with momentum at 0.9 for optimization processes, and weight decay of 0.0005 was used with a learning rate of N frames ∗ 10−5. The LSTM Architecture provides output by returning the prediction at the last time step.

Results

Conv Pooling had the best results at 89.6% when using the three datasets. Comparing the Conv Pooling with the deep RNN using LSTMs, the second feature aggregation method, LSTMs outperformed Conv Pooling at 90.5%. These are comparable results for modern usage and demonstrate that identifying long range temporal relationships between video frames is crucial to video classification.

Conclusions/Discussion

I was able to surpass the benchmark of 2% by over 65% instead of using the naive method of performing 3D convolutions over each single frame. Because Conv Pooling was optimal, it was concluded that max-pooling over the outputs of the final convolutional layer is important. It is clear that LSTMs performed better due to its ability to identity long range temporal relationships. By coupling optical flow and raw image (300 frames/video) as input, I improved previous work that only sampled 60-120 frames.

Summary Statement

I created a deep learning architecture coupling CNNs and RNNs with LSTMs to aid in the problem of video classification, which can be extended to applications ranging from military assistance to navigational technology for the blind.

Help Received

I would like to thank my dad for providing support and resources for my project.