18-794 Project Progress Report

Harsha Chivukula
hchivuku@andrew.cmu.edu

Rishi Khanna
rishik@cmu.edu

Pratik Prakash
pratikpr@andrew.cmu.edu

Markus Woodson
mwoodson@andrew.cmu.edu

Abstract

Recurrent and convolutional networks have been extensively applied to image recognition, image segmentation, natural language processing, and other retrieval tasks. In this work we propose and evaluate several deep neural network architectures that combine image and sound information across time in videos to accurately classify human actions. In our approach we integrate detail, motion, and sound information into our classifier by constructing 3 deep networks and pooling their feature output for classification. We evaluate our results with both convolutional and recurrent networks using Long Short-Term Memory (LSTM) cells. Our current results give us a 54% accuracy on the UCF-101 dataset with only using 1 non-fully trained network which is admirable given other's results.

1. Introduction

Convolutional neural networks have had great success in static image recognition problems such as MNIST, CIFAR, and ImageNet Large-Scale Visual Recognition challenge[15,21,28]. By using a deep neural network, we get both a trainable feature extractor and classifier which are both automatically learned. Recently there has been a surge to apply these networks to other problems in fields such as natural language processing and video classification tasks [2, 3, 4, 5].

Video classification tasks introduce 2 new dimensions to the problem that should be exploited: the time and the sound component. Though we now have new information to exploit, the task itself is very computationally expensive as there may be thousands of frames per video, yet not all of them are useful. An approach that has been used before is to only use the raw frames from the video as input into a CNN in order to classify the video. This not only fails to take advantage of the new information we have obtained, but it also completely disregards the time component of the data.

Thus, we predict that a description of a video does not only need to come from the frames but also from the temporal relation between these frames and the sound associated with each video as well. To that end, we propose a 3-fold network approach using both recurrent and convolutional neural networks. Taking inspiration from previous work [3, 4] which split the processing into 2 streams, a fovea and a context stream, we follow a similar pattern. The CNNs will extract the local frame information while the recurrent connections will calculate the temporal features. We employ Long Short-Term Memory (LSTM) cells in order to implement a recurrent neural network. LSTMs operate similarly to CNNs but are able to integrate information in time.

Naturally we will need to incorporate motion information into our model to achieve better performance. There has been success [4] using optical flow to incorporate motion information into the model. The motion flow serves to provide motion information so we can use a lower fps when training on the raw video frames. However the optical flow was integrated over all 3 channels (RGB). To save some computation time and keep the accuracy loss minimal, we propose computing optical flow on gray scaled versions of the videos.

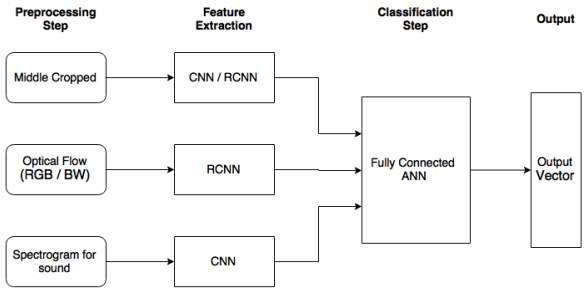

Figure 1. Proposed action classification architecture

We incorporate the sound information similarly to the motion and context information by means of a deep neural network. We train on spectrograms of the sound in each video so we can exploit CNN architectures to find complex frequency information that might not be found in a simple

ANN. Using the features from sound, motion from optical flow, and context from raw image frames taken from videos, we hope to achieve close to state-of-the-art performance while still keeping computational intensity modest.

2. UCF-101 Dataset

The UCF-101 Action Recognition Dataset contains 13,320 videos in a total number of 101 action classes divided into five main types covering a wide range of actions: Body-Motion Only, Human-Human Interaction, Human-Object Interaction, Playing Musical Instruments and Sports [1, 3]. We are given three training and testing partitions, to select videos to train and test on, and we follow the given protocol for cross-validation between the three training partitions of videos. Figure 2 shows sample frames from 6 action classes of the UCF-101 dataset [1].

Figure 2. Sample frames from videos of 2 UCF-101 classes

UCF-101 contains videos of length 10-15 seconds on average, and the videos of each class are subdivided into 25 groups of 4-7 videos each. The videos have a fixed frame rate of 25 FPS and frame resolution of 320 x 240 pixels[9]. All the videos also preserve audio, which we plan to leverage in our next steps to form a spectrogram network to learn feature information from the sound of each video as well. UCF-101 is the most challenging action recognition dataset because it features several pattern recognition challenges encountered in video classification. Furthermore, it contains a much larger number of unconstrained clips from YouTube than other action classification datasets[9].

3. Previous Work

Traditional approaches to video recognition have used hand-crafted features such as Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) to obtain both spatial and temporal information [12]. CNNs are able to automatically learn such features for action classification in videos, and perform better than traditional approaches [3]. [3] suggests two approaches, namely feature pooling and recurrent neural networks. To account for the loss of motion information due to processing only one frame per second for a reduced computational intensity, the use of optical flow images computed over adjacent frames is incorporated.

Long Short Term Memory (LSTM) is used to counter the problem of standard recurrent neural networks of learning spatio-temporal features over long time periods, by using memory units to alter internal state [3]. [3] uses AlexNet and GoogLeNet to process individual video frames, and we follow the AlexNet CNN architecture due to it having lesser parameters to deal with and a shorter time to train, despite GoogLeNet outperforming AlexNet. These suggested approaches in [3] yield the highest performance measured for the the Sports-1M benchmark when incorporating LSTMs in both image frames as well as optical flow.

Two topics of interest described by [10] are the usage of different CNN network structures to describe time information in videos and the pre-processing of frames to speed up computation. The authors investigated four different kinds of CNN fusion techniques in an attempt to fuse information over the time domain.

The baseline used was a simple single frame network representing a single frame of time. Next, they implemented a technique called Late Fusion where two single frame networks were built 15 frames apart and then merged before the first fully connected layer. Individually the single frame networks were not able to describe any motion but after the first fully connected layer the motion could be described by comparing the outputs of the two networks. The third technique used was called Early Fusion. This was done by modifying the first layer of the single frame network to accept a fourth dimension, T, that represented the number of consecutive frames. This comes at the cost of reduced dimensionality in the actual image. This technique performed the worst excluding the baseline. The last technique that was implemented was called Slow Fusion. This involved the combination of the Early and Late fusion techniques where each successive layer has access to more and more time information.

In pre-processing, the frames were split into two streams - a fovea stream and a context stream. The fovea stream acted as the human eye's fovea and cropped the middle area of each frame. The context stream subsampled the original image to the same resolution as the fovea stream. The reasoning behind this was the low resolution image could still represent motion information just as well and the middle cropped image would represent the important features of the frame due to camera bias.

In order to accurately conduct video action classification, both spatial and temporal features should be consid-

ered. As a result, [4] proposes a two-stream convolutional neural network which incorporates both spatial and temporal networks. In the temporal network, the team also performs optical flow computations on the raw video frames in order to retrieve the temporal features from the videos. In the spatial network, the team used a CNN on the raw video frames to retrieve spatial features. After retrieving features from both the spatial and temporal dimensions, the team averaged the class score from both streams to classify the action. Fusion by averaging the results of the two streams resulted in an accuracy of approximately 85% [4].

4. Approach

The approach we used in constructing our solution was motivated by the biological principles described in [10]. We construct three deep neural networks representing detail, speech, and motion, and then we pool the features generated by these networks to derive the final classification.

In the human eye, the fovea centralis is responsible for the sharpest, and most focused vision. The first network we construct applies this principle by first pre-processing the frames by cropping the central region then using this as input to the network. We chose to model this network using an AlexNet CNN architecture because AlexNet has a history of performing well on ImageNet[11]. We also test a recurrent architecture using LSTM on the cropped images to introduce time information. We hypothesize that this will lead to moderate improvement in performance as our motion information should pre-dominantly come from optical flow as described below.

Next we attempt to incorporate the sound information of the videos into our model. The sound is extracted from the videos and an image representing the spectrogram is created. We considered implementing this network using an AlexNet architecture however we hypothesize it will sufficient to use a simpler model of a CNN. The spectrogram features are unlikely to be as complex as features that would be found in an image classificaton network, thus a complex network like AlexNet will not be neccesary. A CNN approach is preferred so we can exploit CNN architectures to find complex frequency information that might not be found in a simple ANN.

The last network we construct characterizes the time information of the videos. The optical flow information is derived based on movement in the video. There are two key components of this network that still need to be finalized. Firstly, when representing the optical flow information it is common to use all 3 RGB channels. The main purpose of this approach is to gain context in the motion. But as we are getting our context from our fovea stream we hypothesize it would be sufficient to compute optical flow on grayscale frames. Using this approach we also save computation time and space as our input only has 1 channel now. To encurage the network to find generalized feautures we sample the video at different frame-rates. Sampling too frequently could potentially lead to learning of features that are too specific so it is likely we will sample at intervals between 6-30 frames apart as described in [3]. The architecture of choice for this stream will be recurrent using LSTM to emphasize the temporal information in optical flow.

5. Preliminary Results

We test how our networks perform on the UCF-101 dataset using the 3-way cross-validation as proposed by the UCF-101 providers[1]. We begin by pre-processing our data to extract the optical flow and cropped images. Currently we have only done this for 3-channel RGB images from each of the videos and have them stored on disk to avoid long computation time when training. We train AlexNet on our cropped frames and evaluate its performance without using any optical flow or temporal information.

Currently our CNN for classifying the fovea stream of cropped images is in the training process. After pausing training at 5 epochs we can report that we have achieved 54% accuracy on UCF-101. This is a good first step as without incorporating any other temporal or sound information, as well as without a complete network, we have achieved a little above average performance compared to others. However, since our testing process is most likely not as entirely rigorous as that of an official lab group, it would be fair to say that we are currently overestimating our current performance.

6. Future Work

We have yet to fully implement and train the complete neural network architecture for classifying actions in videos using both video frames and sound as features. First, we plan to finish implementing the Recurrent Convolutional Neural Network (RCNN) that learns features based on the pre-processed optical flow computations. Afterwards, we will train and test this RCNN for both optical flow computations over all three RGB channels and over gray-scaled images. We expect training the RCNN using computations over the gray-scaled videos to be more computationally efficient with trivial loss in accuracy compared to training using computations over RGB.

As mentioned above, we have already implemented and trained our AlexNet CNN for the cropped frames. Our next step is to test this AlexNet using the cropped images of the UCF-101 database and note down the testing error. Depending on this testing error and the accuracy of the AlexNet, we would like to determine whether a recurrent convolutional neural network architecture would better represent the middle-cropped frames than an AlexNet CNN and per-

form slightly better because temporal integration for any network should perform better than just raw frames. As a result, if we have additional time, we would like to implement a RCNN to test this hypothesis.

In order to learn features from the sound data extracted from the videos, we first need to convert the sound into spectrogram images. After we have obtained spectrograms for each video in the data set, we will implement another CNN to learn features from these processed images. We expect sound features by itself will be insufficient to accurately classify videos by actions. Nevertheless, when sound features are combined with the previously trained image features, we expect the network to perform better than just relying on image features alone.

Once we have implemented, trained, and tested the individual CNNs / RCNNs for optical flow, cropped and spectrogram images, we will integrate all 3 of the networks using a simple fully connected ANN. This fully connected network will take in the outputs of our 3 feature computation networks and perform the video classification.

References

[1] UCF101 - Action Recognition Data Set, `http://crcv.ucf.edu/data/UCF101.php`

[2] A. Graves, A. Mohamed and G. E. Hinton. Speech Recognition with Deep Recurrent Neural Networks.

[3] J. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici. Beyond Short Snippets: Deep Networks for Video Classification.

[4] K. Simonyan and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos.

[5] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, pages 17641772, Beijing, China, 2014. 2

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 10971105, Lake Tahoe, Nevada, USA, 2012. 1, 2, 3

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CoRR*, abs/1409.4842, 2014. 1, 3, 4

[8] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818833, Zurich, Switzerland, 2014. 1, 3

[9] K. Soomro, A. R. Zamir and M. Shah. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. CRCV-TR-12-01, November, 2012.

[10] 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.

[11] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." Web.

[12] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In Proc. CVPR, pages 18, Anchorage, Alaska, USA, 2008. 2, 3