CS482/682 Final Project Report Group 3

Deep Learning on Video Classification

Jingxi Liu, Yuetong Liu, Yingkun Wang, Jiaqian Zhong
jliu238, yliu390, ywang601, jzhong16

1 Introduction

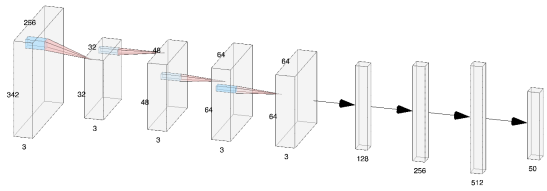
Background Video classification becomes a hot trend in current technological field. With deep learning methods, we could analyze human actions, gestures, and do AR interactions in videos. We learnt and practiced many concepts about image classification in class, and we further think videos are a collection of a set of images arranged in a specific order. As a result, we want to do a video related project to both solid what we've learnt in this semester and to have some innovative ideas go beyond. In this project, we are going to choose the sports videos in UCF101 dataset to perform motion classification task. We first pre-processed videos to get important features and representations. Then we chose appropriate models to train and validate the data. Finally, we used the test data to find out the most accurate model. Our ultimate goal is to accurately classify different kind of sports.

Related Work Soomro et al, used the standard bag of words method to classify the UCF101 dataset and achieved an overall accuracy of 44.5%. UCF101 contains five different types of videos, which are Human-Object Interaction, Body-Motion, Human-Human Interaction, Playing Musical Instruments and Sports. In Soomro's study, Sports actions achieve the highest accuracy (50.54%) in their baseline model. Therefore, our project will focus on sports video classification. Moreover, given UCF101 dataset contains a large number of classes and clips, we would use deep learning neural networks to improve the classfication accuary.

2 Methods

Dataset UCF101 is an action recognition dataset of realistic action videos, collected from YouTube, having 101 action categories. In this project, we selected 50 actions, focusing on Sports. Each action category is grouped into 25 video groups and each group contains 4-7 videos. For the data preprocessing part, we used OpenCV to select about 20 frames each video and resize the images to standardize the input(256, 256, 3). After that, the dataset was split into training, validation, and testing set with a ratio of 3:1:1.

Setup, Training and Evaluation We used **2D-CNN** as our baseline model. During the training, we used pre-trained model VGG16 for feature training and saved the weights. Then we loaded the weight into another CNN models that contains four fully connected dense layers. For each video, we used CNN to classify all frames individually. The label that has the largest probability was chosen as our predicted output. For **3DCNN** as our improved model we applied a simple approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks. Compared to 2D CNN, 3D CNN is able to model temporal information better owing to 3D convolution and 3D pooling operations. The model takes multiple frames as input and after the first convolution layer, temporal information is collapsed completely. Our model contains four convolutional layers which are mixed with Relu activation, maxpooling and batch normalization. Three fully connected layers were added after to output the predicted classes. Below is the flowchart of the 3DCNN model.



Model	Training Epoch	Testing Accuracy
2D-CNN	200	54%
3D-CNN	70	72.49%

Figure 3: Accuracy of Models

3 Results

For 2D-CNN, we trained the model for 200 epochs, the below plots are the accuracy and loss value w.r.t number of epochs. According to the plots below, the model converges in 25 epochs.

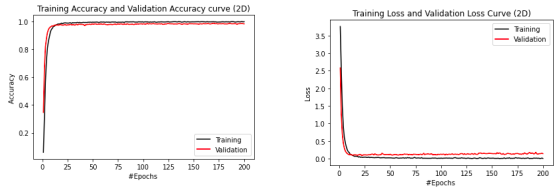


Figure 1: Loss and Acc for 2D

For 3D-CNN, we trained the model for 70 epochs with batch 20 and learning-rate 0.0001. Plots below are the loss and accuracy w.r.t number of epochs.

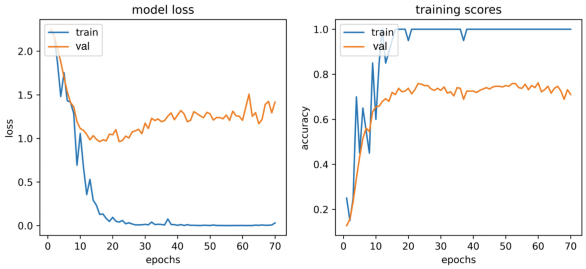


Figure 2: Loss and Acc for 3D

The final accuracy of the 2-D CNN and 3-D CNN are as in figure 3.

4 Discussion

One of the problem exists in our current project is that the input size of our dataset is relatively small. Many people dealing with the same dataset choose to use video directly as input. However, what we did was that during pre-processing, we selected around 20 frames from each video and changing each video as a sequence of frames. Also, we cropped the frame size from 320 $\times$ 240 to 256 $\times$ 256, and removed colors from images. Those pre-processing steps will decrease the final accuracy. Though the 3DCNN model didn't achieve a great improvement in terms of the model loss and accuracy compared to the pre-train VGG16 model, the 3DCNN model is still able to deliver a promising model with much smaller dataset within 4 hours on 1 GPU. The VGG16 is trained on ImageNet, which contains over 15 million labeled high-resolution images belonging to roughly 22,000 categories. The 3DCNN model is trained on about 31,250 images on 50 categories.

5 Future work

To improve the current project, we would like to include more images to increase the size of our dataset. In terms of the 3DCNN, we would like to add more layers and possibly modify it as a 3D-ResNet model. Also, we think about making the current sport video classification task be more creative in the future. Instead of just classifying the sport types, we could further predict which team is the winner of a game; we could use deep learning methods to classify actions happening in the videos and notify athletes when they violates the sports' rules.

References

[1] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015. APA

[2] Wang, Limin, et al. "Appearance-and-relation networks for video classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[3] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild., CRCV-TR-12-01, November, 2012.

[4] https://github.com/HHTseng/video-classification

[5] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv preprint arXiv:1212.0402 (2012).