DLCV HW4

R08922050 Chih-Chun YANG

December 2019

1 Problem 1: Trimmed action recognition w/o RNN (20%)

1.1 Describe your strategies of extracting CNN-based video features, training the model and other implementation details (which pretrained model) and plot your learning curve (The loss curve of training set is needed, others are optional). (5%)

- For image processing, I only use the middle 10 frames and use **average pooling** to merge the video sequence into video embedding.
- I use ResNet50 as the feature extractor and does not fine-tune it since the model is too large. During training, I use Adam as the optimizer and reduce the learning rate if the loss does not improved for 3 epochs.
- In the classifier, I use multiple linear and SELU with dropout ratio 0.5 or 0.1. For each linear layer, I reduce the dimension to 0.25 times of previous layer.

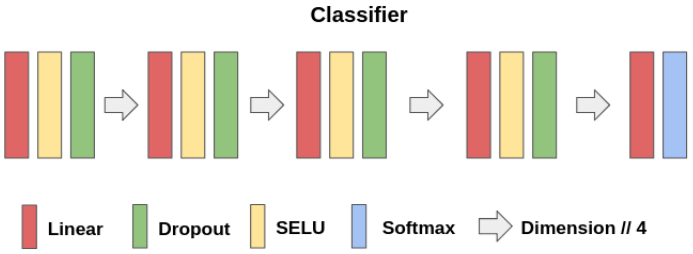
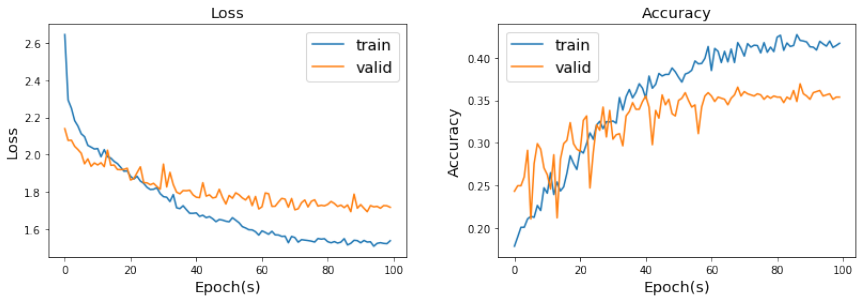


Figure 1: Classifier



(a) (b)

Figure 2: Loss and Accuracy

1.2 Report your video recognition performance (valid) using CNN-based video features and make your code reproduce this result. (5%)

- Validation Accuracy: 0.3602

1.3 Visualize CNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels.(10%)

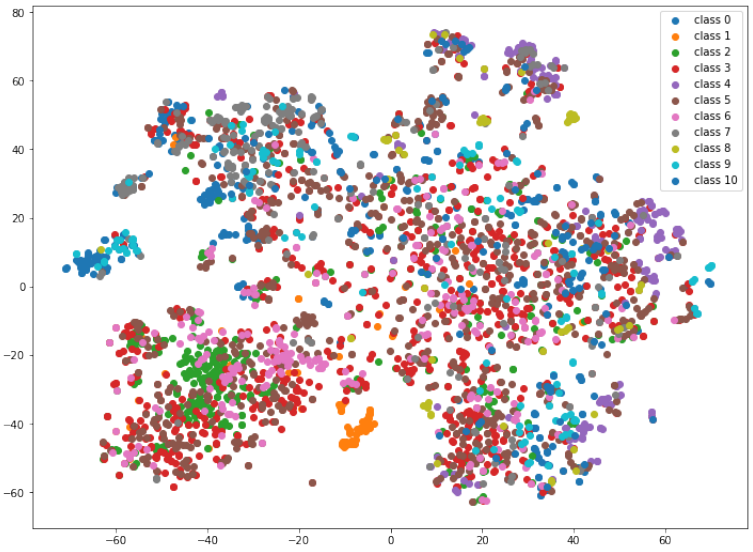


Figure 3: TSNE Visualization

2 Problem 2: Trimmed action recognition w/ RNN (40%)

2.1 Describe your RNN models and implementation details for action recognition and plot the learning curve of your model (The loss curve of training set is needed, others are optional). (5%)

- I use ResNet50 plus a linear layer which convert the dimension from 16384 to 256 as feature extractor.

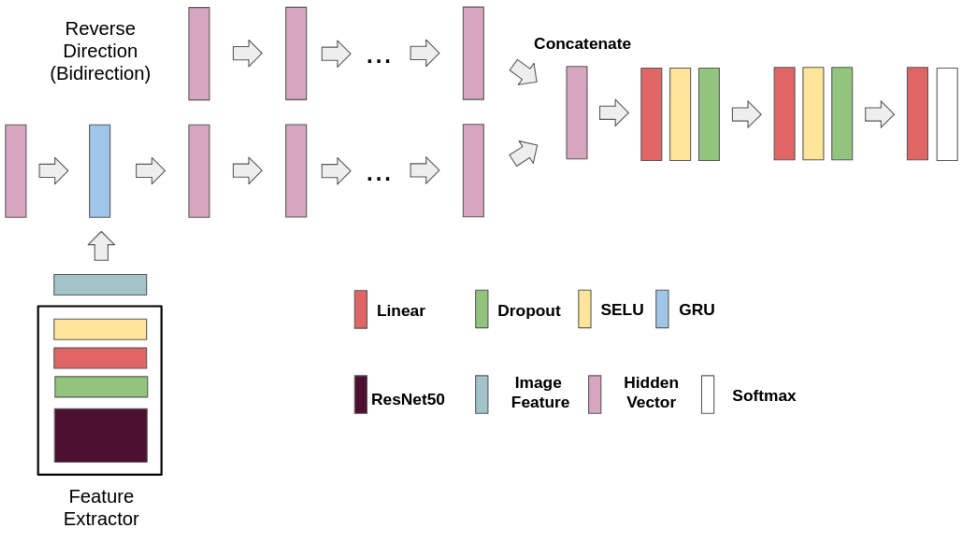
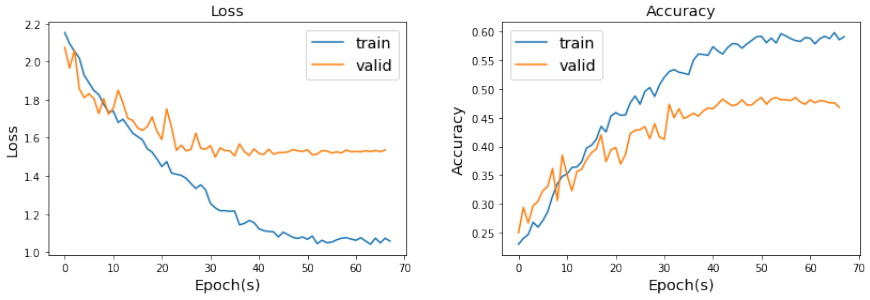


Figure 4: Modle Architecture



(a) (b)

Figure 5: Loss and Accuracy

2.2 Your model should pass the baseline (valid: 0.45 / test: 0.43) validation set (10%) / test set (15%, only TAs have the test set).

- Validation Accuracy: 0.4772

2.3 Visualize RNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels. Do you see any improvement for action recognition compared to CNN-based video features ? Why? Please explain your observation (10%).

- RNN-based video feature is classified better than CNN-based. It is probably because RNN can capture the sequential information and does not view all the frame equally important, compared with average pooling in problem 1.

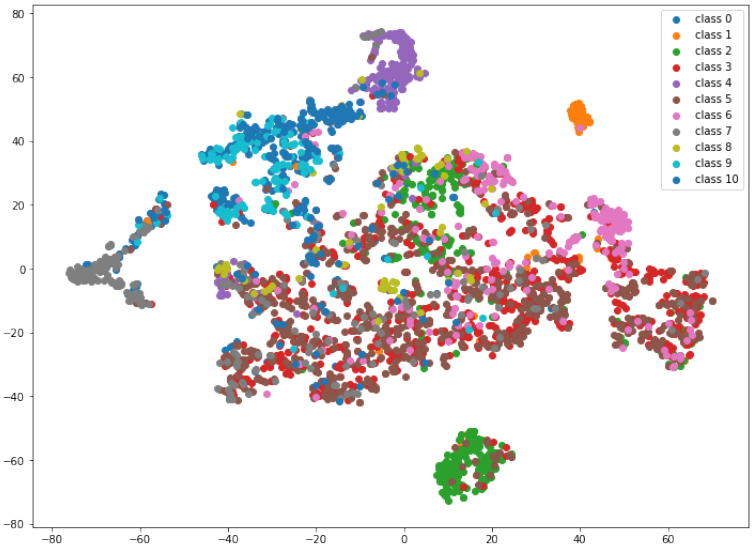


Figure 6: TSNE Visualization

3 Problem 3: Temporal action segmentation (40%)

3.1 Describe any extension of your RNN models, training tricks, and post-processing techniques you used for temporal action segmentation. (5%)

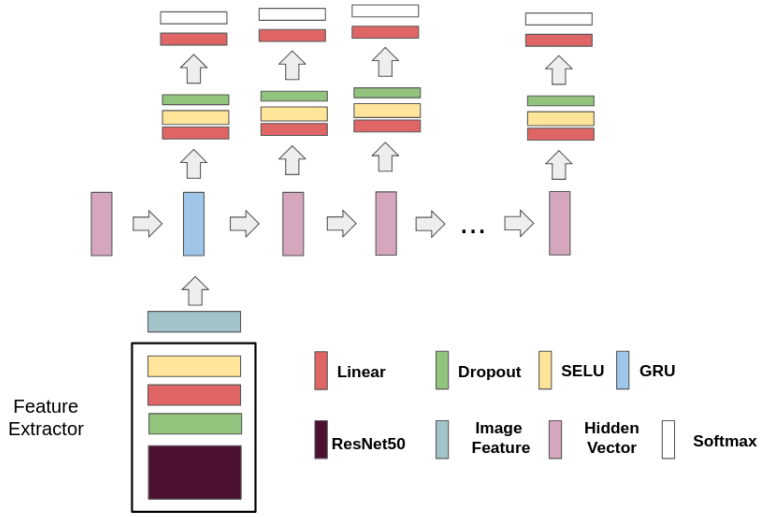
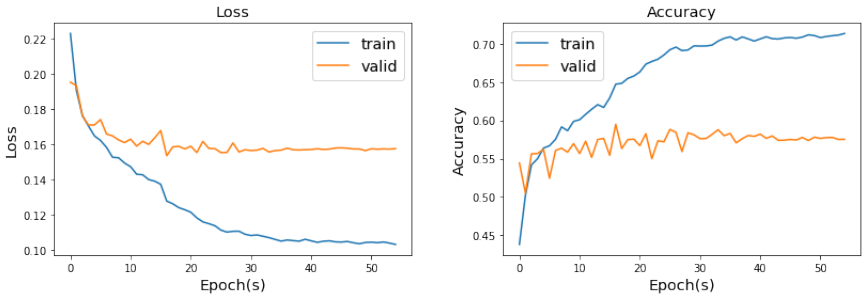


Figure 7: Modle Architecture



(a) (b)

Figure 8: Loss and Accuracy

- I use GRU as the encoder and split the full length video into short video with 20 frame each. Training process is similar to problem 2.

3.2 Report validation accuracy in your report and make your code reproduce this result. (20%)

- Validation Accuracy: 0.5974

3.3 Choose one video from the 7 validation videos to visualize the best prediction result in comparison with the ground-truth scores in your report. Please make your figure clear and explain your visualization results (You need to plot at least 500 continuous frames). (15%)

- I choose category "BaconAndEggs" to visualize. The model know where to output "0" (Other) most clearly. It is probably because 0 is the most common class in each video. So the model learns well on distinguishing it from other class.

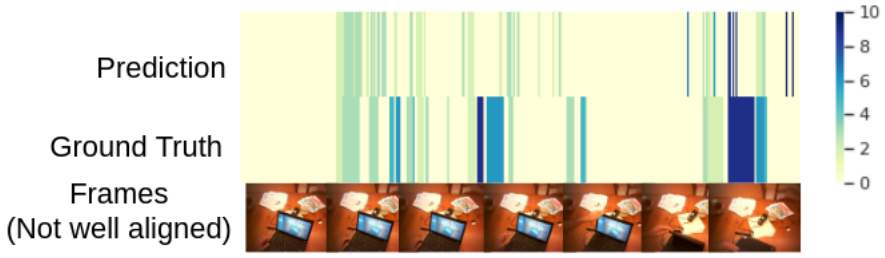


Figure 9: "BaconAndEggs" Visualization

3.4 Collaborators

- 林子淵 R07921100
- 黃孟霖 R07922170
- 張緣彩 R07922141