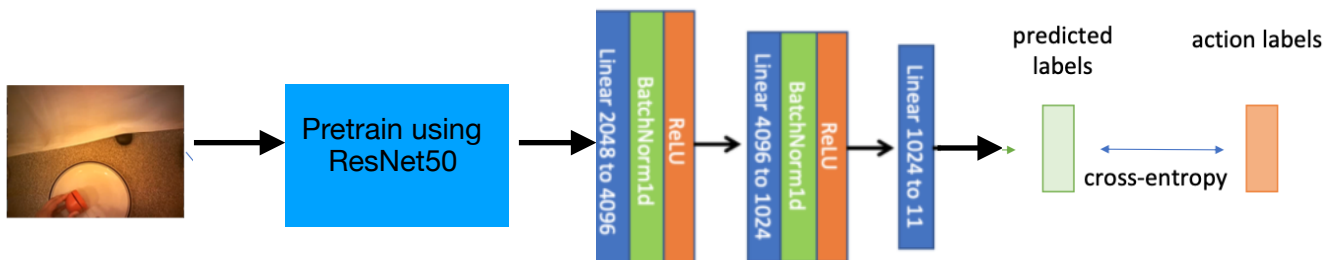


Report:HW4 DLCV

Student ID: R07943158

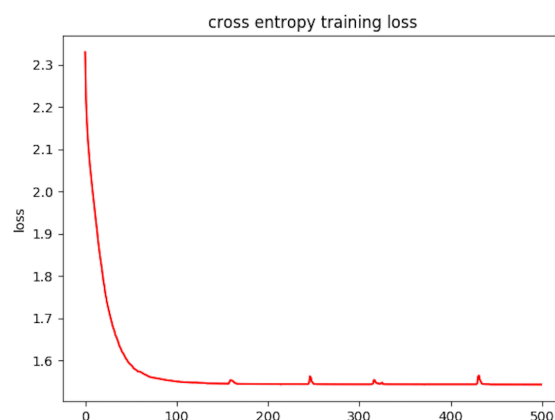
Problem 1 : Trimmed action recognition w/o RNN (20%)

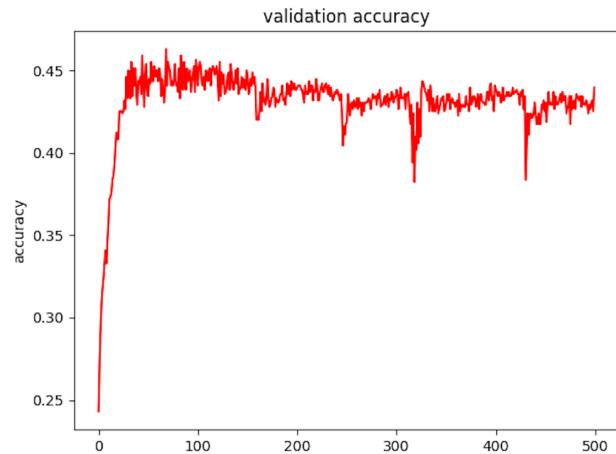
1. Describe your strategies of extracting CNN-based video features, training the model and other implementation details (which pre-trained model) and plot your learning curve (The loss curve of training set is needed, others are optional). (5%)



During Training, I use ResNet50 pre-trained on ImageNet. This way I extract the features. They are 2048 in number. Then I feed the features in my model. This helps us get predicted label. Cross-Entropy Loss function is used to calculate the loss. I train my model for 500 epochs with bath size of 64.

For valid dataset I get accuracy as - **0.4629** . Also I save the predicted labels in file: p1_valid.txt

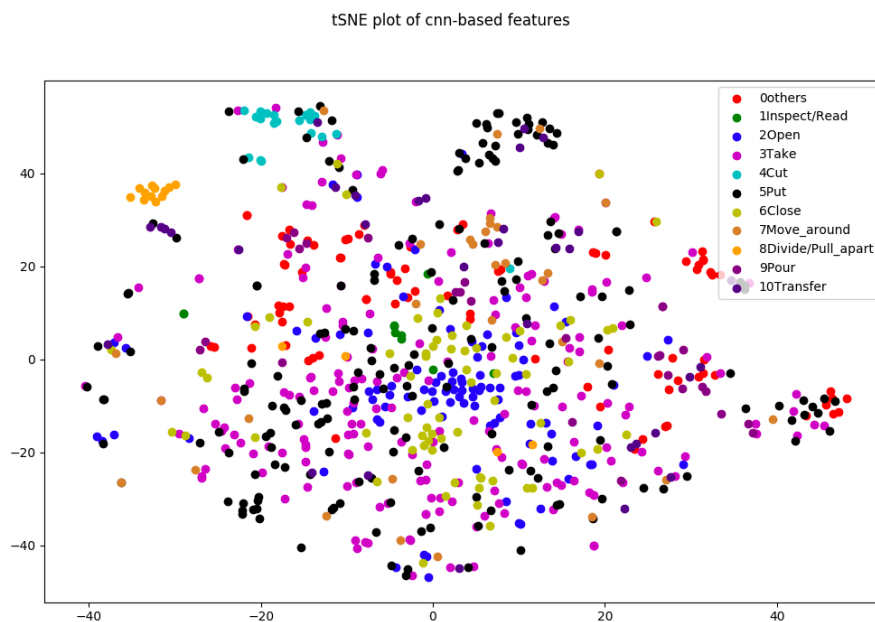




2. Report your video recognition performance (valid) using CNN-based video features and make your code reproduce this result. (5%)

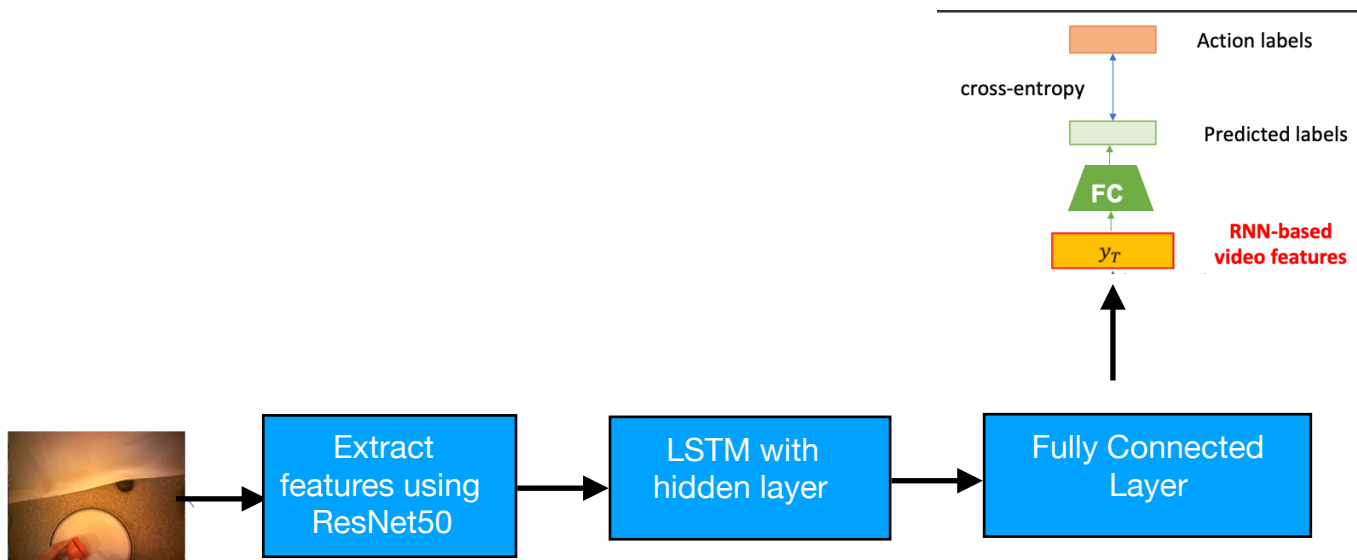
```
evaluation ans...
accuracy: 0.46293888166449937
```

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%)



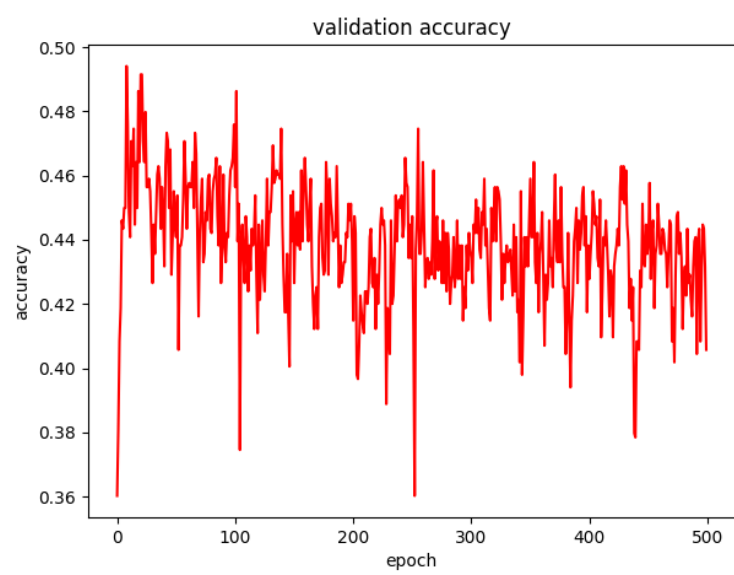
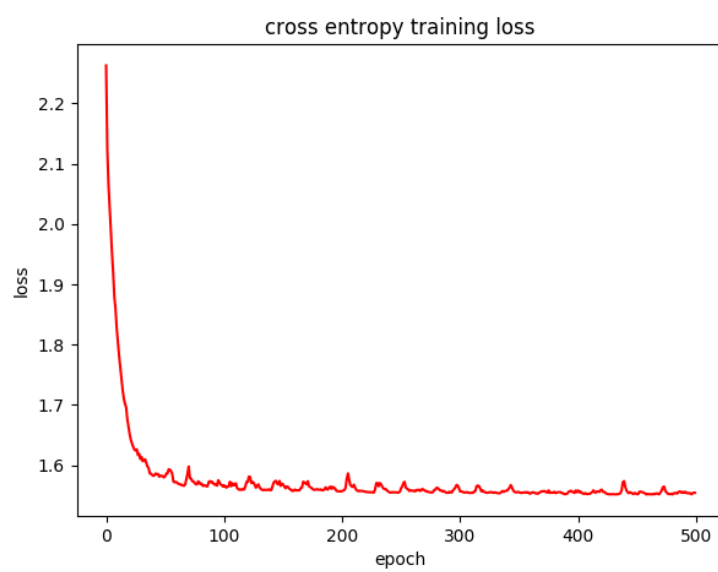
Problem 2 : Trimmed action recognition w/ RNN (40%)

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%)



During Training, I use ResNet50 pre-trained on ImageNet. This way I extract the features. The extracted features are fed into LSTM network with 512 hidden layers. I use batch size as 32. Hence I need to do zero padding to have same input length. After LSTM network, I apply three fully connected layers followed by a softmax function. Thus I get predicted labels. I use cross entropy loss function to compute loss.

For valid dataset I get accuracy as - 0.49414. Also I save the predicted labels in file: p2_result.txt

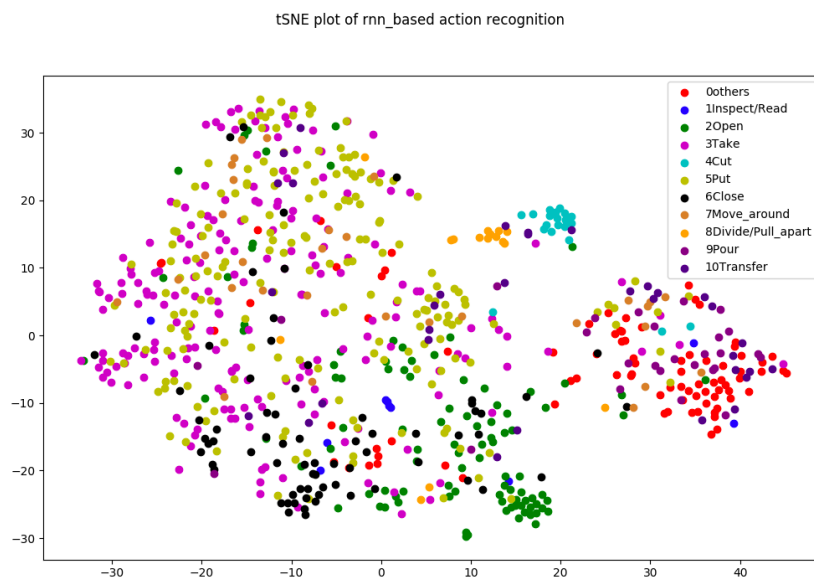


2. Your model should pass the baseline (valid: 0.45 / test: 0.43) validation set (10%) / test set (15%, only TAs have the test set).

For valid dataset, the accuracy is 0.49414

```
evaluation ans...  
accuracy: 0.494148244473342
```

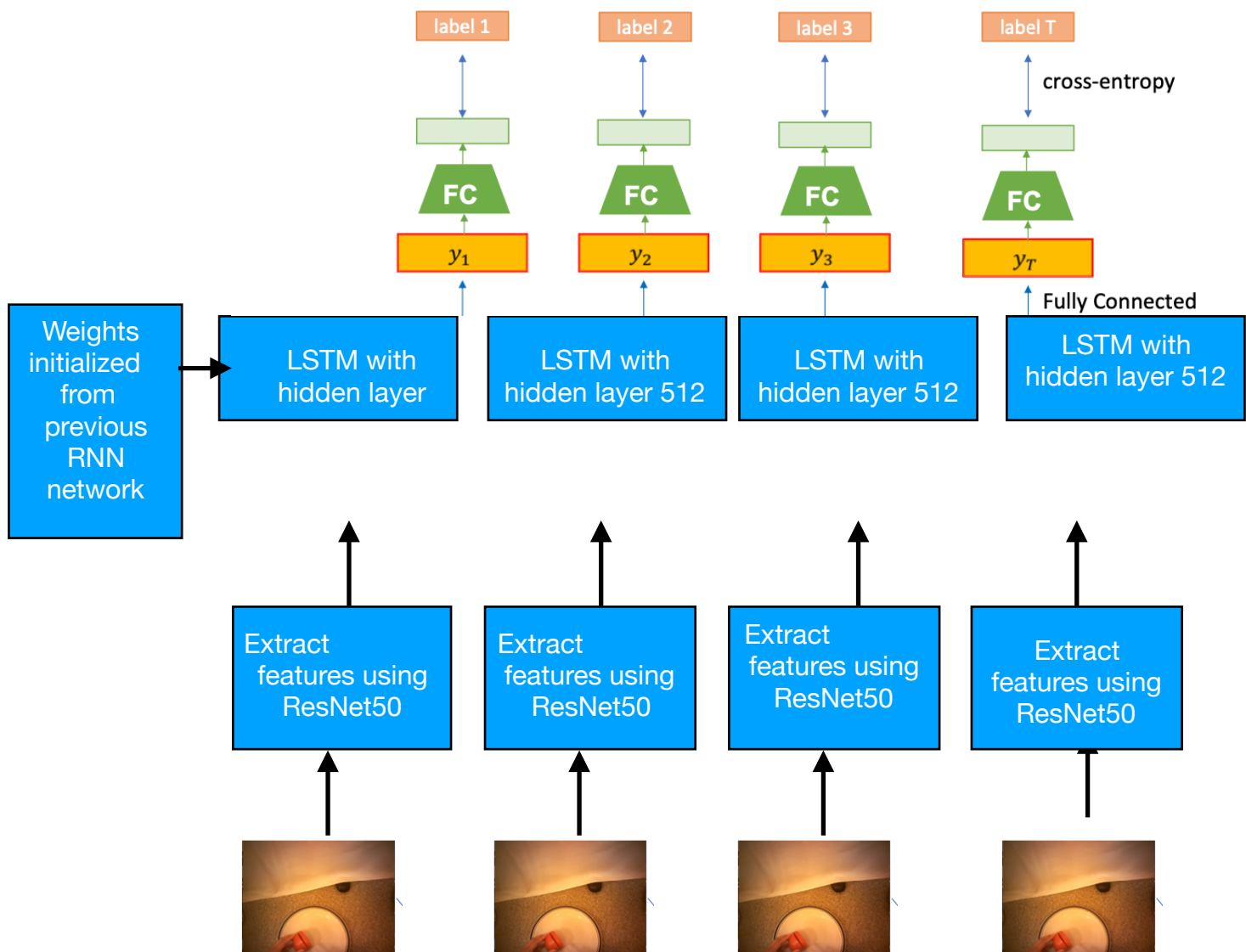
3. Visualize RNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels. Do you see any improvement for action recognition compared to CNN-based video features ? Why? Please explain your observation (10%).



Yes, there is slight improvement in performance of action recognition in this case as compared to CNN-based. The reason is the that LSTM network. LSTM retains long range information as compared to CNN. So it helps in better recognition.

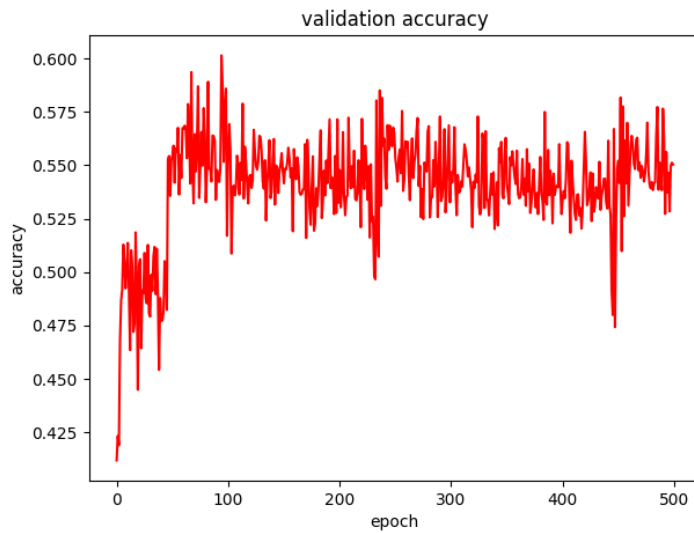
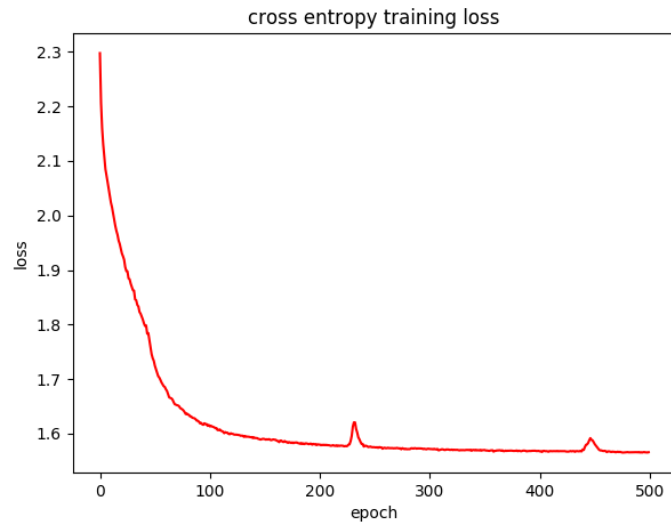
Problem 3 : Temporal action segmentation (40%)

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



I use the same method as described in HW slides. I initially extract the features using ResNet-50. Then I design LSTM model using the weights from previous RNN model. The last layer of is

fully connected. The loss function used is cross entropy. Hidden size of LSTM is 512. Batch size is 32.














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

Text File	Accuracy
OP01-R02-TurkeySandwich.txt	0.50197628458498
OP01-R04-ContinentalBreakfast.txt	0.617107942973523
OP01-R07-Pizza.txt	0.565358154593282
OP03-R04-ContinentalBreakfast.txt	0.528683914510686
OP04-R04-ContinentalBreakfast.txt	0.629493087557604
OP05-R04-ContinentalBreakfast.txt	0.543248945147679
OP06-R03-BaconAndEggs.txt	0.595744680851064
Mean Accuracy	0.568801858602688

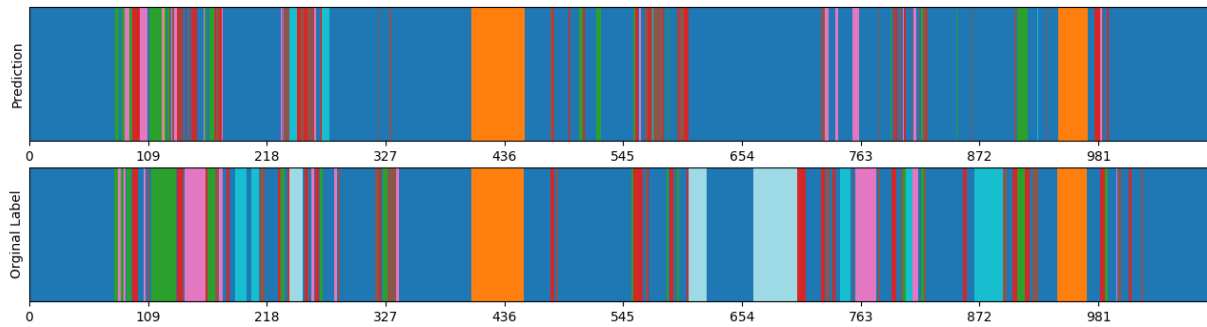
- 3. Choose one video from the 7 validation videos to visualize the best prediction result in comparison with the ground-truth scores in your report. Please make your figure clear and explain your visualization results (You need to plot at least 500 continuous frames). (15%)**

Using `matplotlib.colorbar.ColorbarBase` to plot the visualizer, and choosing color by color map in `matplotlib.pyplot`. ("tab20")

										
others	Inspect/ Read	open	take	cut	put	close	Move around	Divide/ pull apart	pour	transfer

OP04-R04-ContinentalBreakfast

Accuracy: 0.629



We visualize the predicted label against original label. This is the best accuracy result. These are the frames of the videos.

Reference: <https://github.com/CheeAn-Yu/DLCV2019/tree/master/hw4-Chee-An-Yu>