Please use this report template, and upload it in the PDF format. Reports in other format will result in ZERO point. Reports written in either Chinese or English is acceptable. The length of your report should NOT exceed 8 pages.

Name: 鍾勝隆　Dep.:電信碩一　Student ID:R06942052

[Problem1]

1. (5%) Describe your strategies of extracting CNN-based video features, training the model and other implementation details.

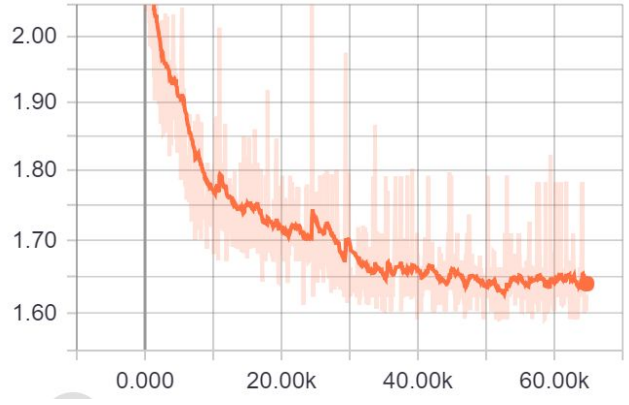
   First, I downsampled the video to 2 fps and used the ResNet50 to extract the features of each frame represented by 1000 dimension vector. For each video, I averaged the frame features. Thus, videoes were embedded into 1000 dimension vector.

   To classify the video category, I used four fully connected layers and add batch normalization for the output of the layers. The loss function is the cross entropy of the softmax output and the ground truth tag.

2. (15%) Report your video recognition performance using CNN-based video features and plot the learning curve of your model.

   The accuracy of the validation data, 517 videoes, is **42.36%**. The learning curve is shown in Fig. (1).

   Fig. (1) The loss curve when training. The darker color line is the smoothed curve and the lighter color is the raw data. x-axis is the training step.



[Problem2]

1. (5%) Describe your RNN models and implementation details for action recognition.

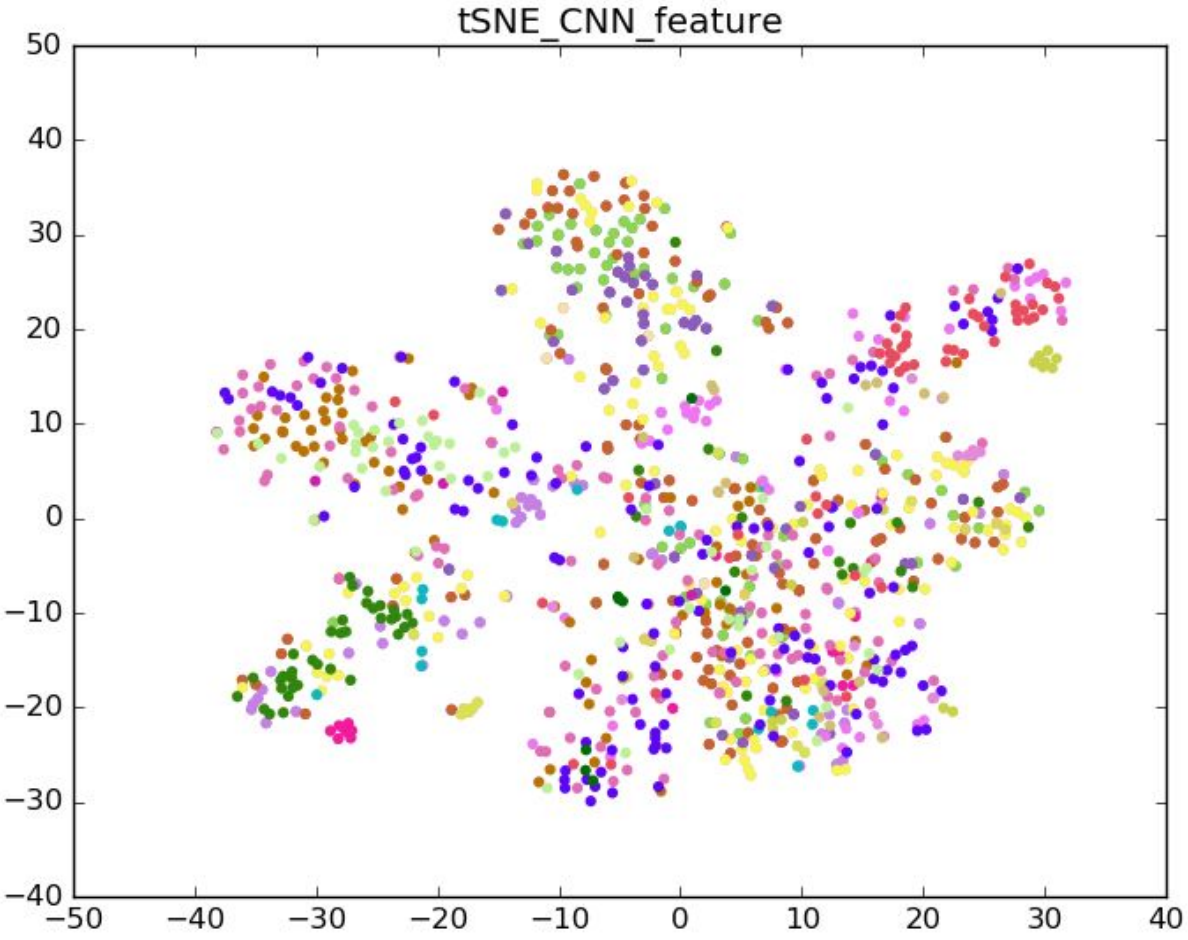
   Using the same feature extract strategy, I turned the videoes into Resnet50 features sequences as the inputs of the RNN. For the RNN part, I used the **LSTM model** with 1 layer whose hidden size is 1000, which can be seen as re-extracting the

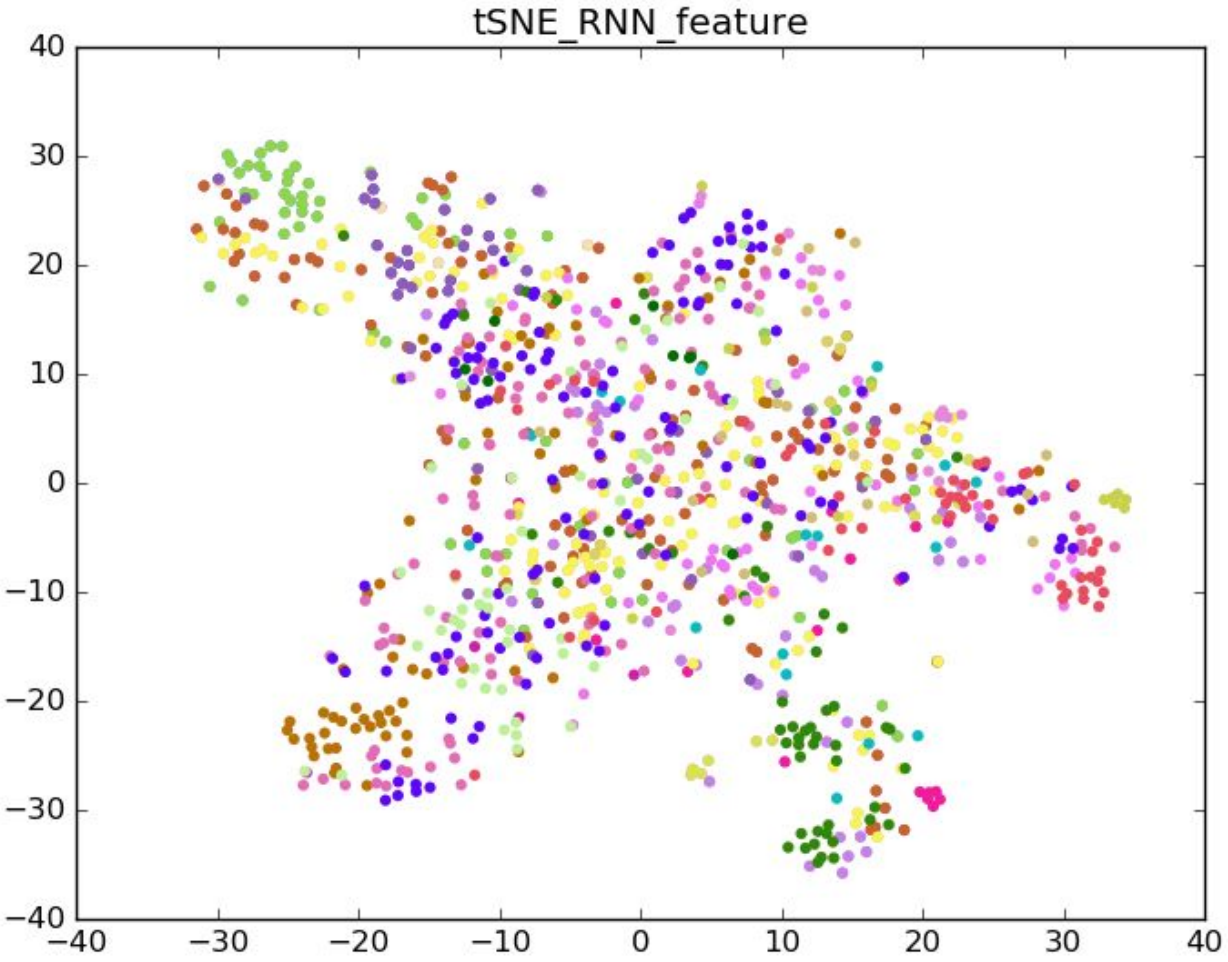
feature of the relationship between frames. I used the **zero padding** technique for the sequence with different length to speed up the training. Last, I fed the last hidden state to a fully connected network.

2. (15%) Visualize CNN-based video features and RNN-based video features to 2D space (with tSNE). You need to generate two separate graphs and color them with respect to different action labels. Do you see any improvement for action recognition? Please explain your observation.

The validation accuracy for my RNN-based model is **48.74%**. In the Fig. (2), I found CNN based features are denser with the different tags, which makes the fully connected network hard to classify the data, like using a line to separate different category.

Fig. (2) The t-SNE graphs of the CNN-based and RNN-based features for validation data





[Problem3]

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

   I changed the LSTM from unidirectional to **bidirectional** and added one more LSTM layer, which made the hidden state dimension become 2000. Then, I trained the fully connected network with these new features. For the issue that training data is too long, I **segmented the frame sequence with 400 frames and sampled it from the total video every 20 frame**. Thus, I generated 1398 video sequence from the training data. Originally, I used random segmented, but in this way, the training data might not represent the real data distribution because the random segmenting missed some part of sequence.

2. (10%) Report validation accuracy and plot the learning curve.

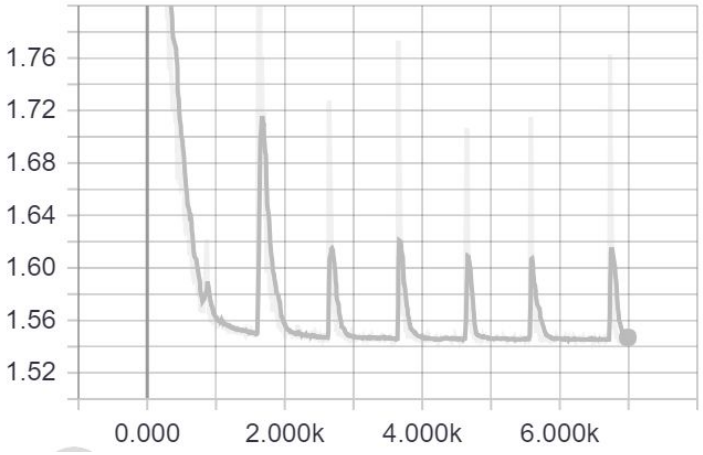
   My RNN sequence to sequence works best in the first video - BaconAndEggs, shown in Table. (1). In Fig. (3), although the loss dropped quickly, there are some

peaks after it decreased to around 1.56. I think it is because I used mini batch for the training, which is a technique to make the model escape from the local minimum.

Table. (1) The accuracy of the validation video

Video Category	Accuracy
BaconAndEggs	**57.804%**
ContinentalBreakfast	57.249%
TurkeySandwich	52.742%
Pizza	46.354%
Cheeseburger	56.765%
Average	55.259%

Fig. (3) The learning curve of the RNN sequence to sequence model



3. (10%) Choose one video from the 5 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 300 continuous frames (2.5 mins).

I choose the video, BaconAndEggs, and plot the 400 tags of it. I found my prediction tend to be noisy, like the part of "Move Around". Moreover, my prediction lagged if it is compared with the ground truth, which I think is due to the characteristic of RNN. It need to take more frames to assure its tags.