Course: Deep Learning for Computer Vision 16 December 2019
Name: Zeynep Cankara
Student Id: T08902113

Homework 04

Problem 1:

Part 1.1:

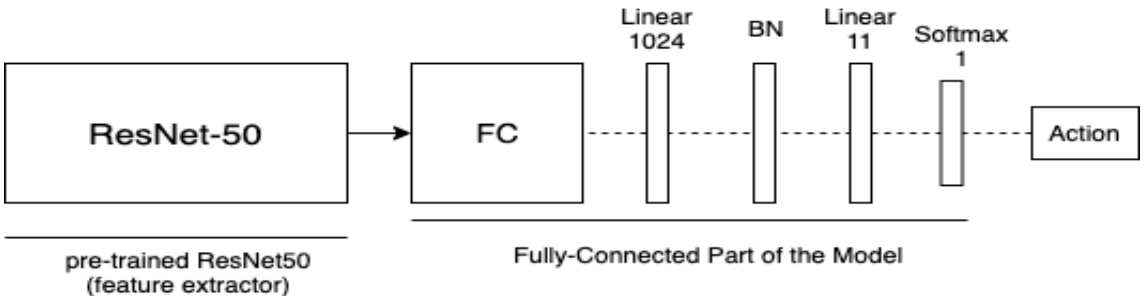

Figure 1: CNN Feature Extractor Model

CNN Feature Extractor Details:

- The model uses ResNet50 (excluding the last 2 layers of the architecture) as a feature extractor with freezed parameters which are pre-trained with ImageNet dataset. The model outputs features with size (2048 x 9 x 7).
- The action recognition part of the model contains 2 fully connected layers and a Batch Norm layer.
- I used Batch-norm and Relu functions for sustaining healthy gradient flow.
- I used Adam optimizer with learning rate of 0.00001. Together with the cross-entropy loss. The model trained with 20 epochs.
- **The Validation Set Accuracy of the model: 0.367 (37%)**

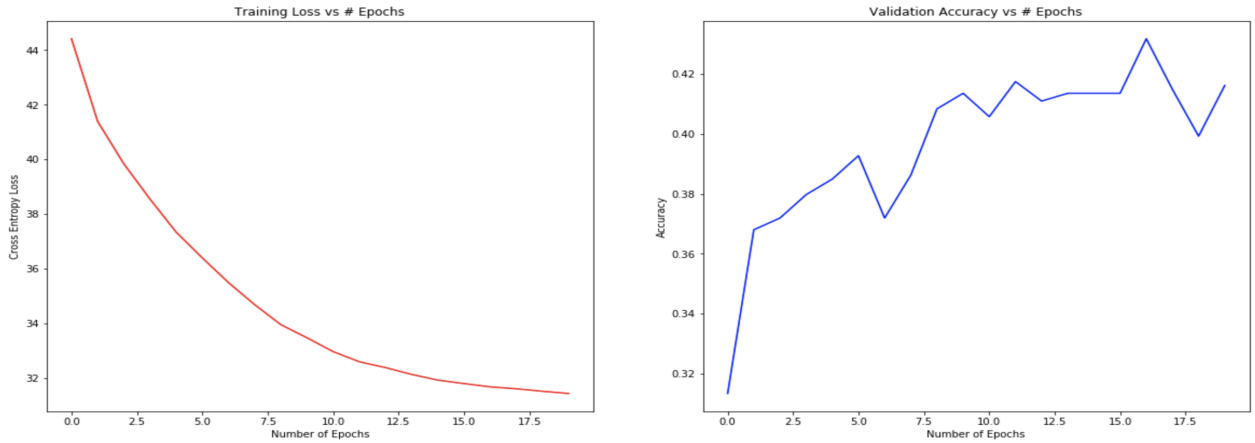

Figure 2: Plot of Training Loss and Validation Set Accuracy

Part 1.2

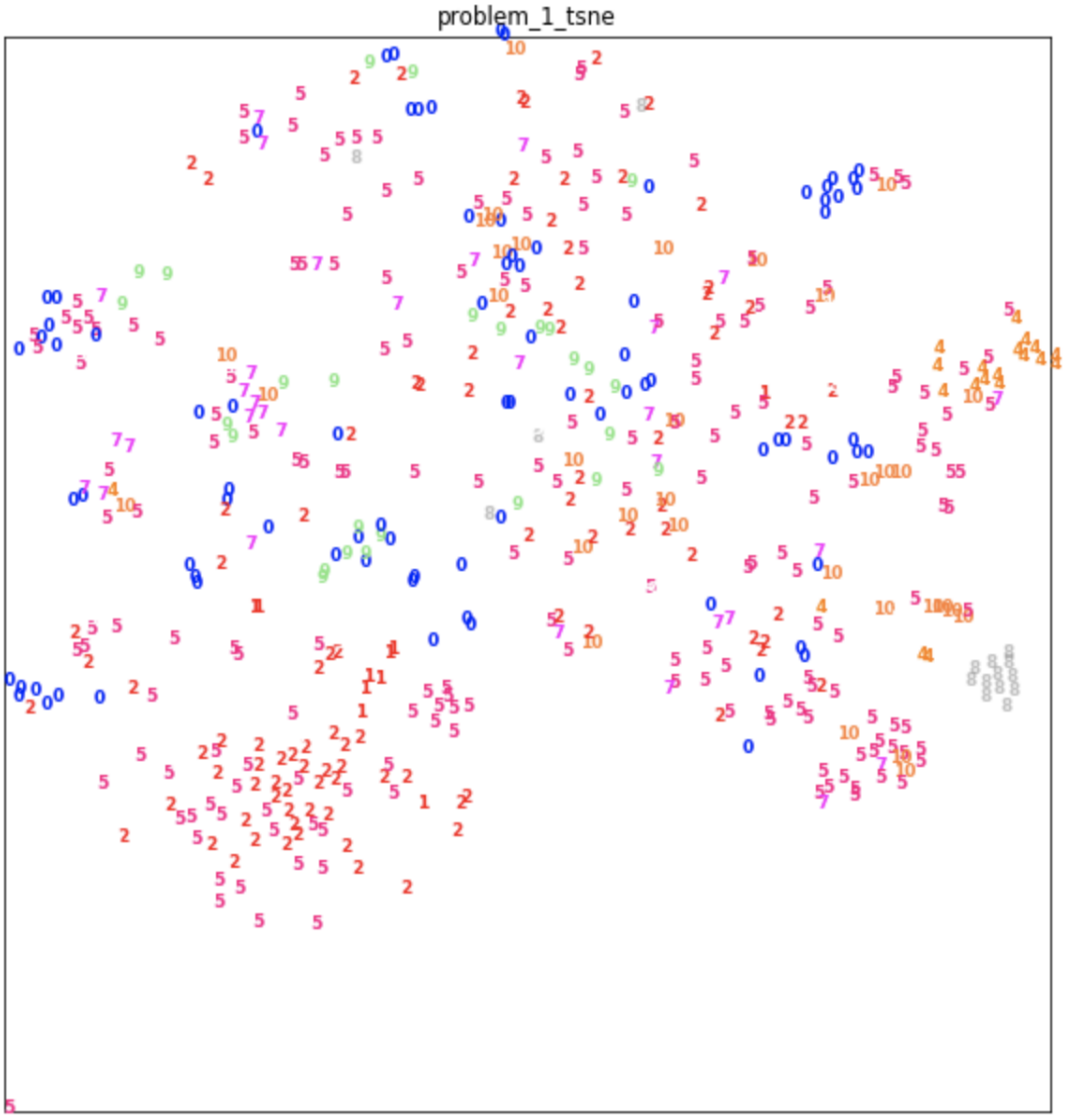

Figure 3: CNN-based video features on 2D space via t-SNE visualization

In the figure each color represents a different action where there are 11 actions possible in total. Actions are encoded with unique number and color values which are displayed on the Figure 3. The data belongs to the validation set features.

Problem 2:

Part 2.1:

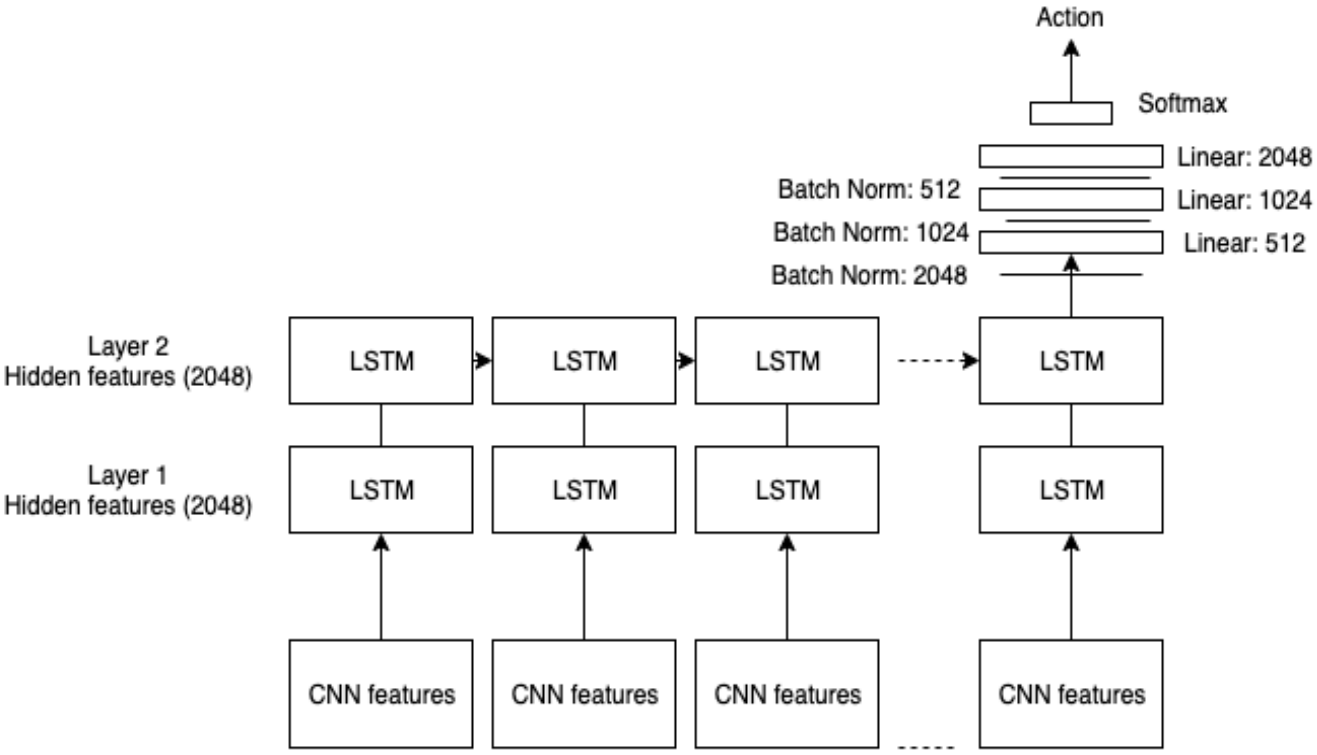

Figure 4: RNN Feature Extractor Model

RNN Feature Extractor Details:

- The model uses ResNet50 as a feature extractor (excluding the last layer) with freezed model parameters which are pre-trained on ImageNet dataset. The model outputs features with size (2048 x 2).
- The Max-pool layer at the end of the connected layers of ResNet50 is restored to keep feature space small.
- I used Batch-norm to sustain healthy gradient flow.
- I used Adam optimizer with learning rate of 0.00001. Together with the cross-entropy loss. The model trained with 50 epochs.
- Since the videos differ in length I needed to pad the videos before feeding to the RNN.
- **The VALIDATION SET ACCURACY: 0.471: (47%)**

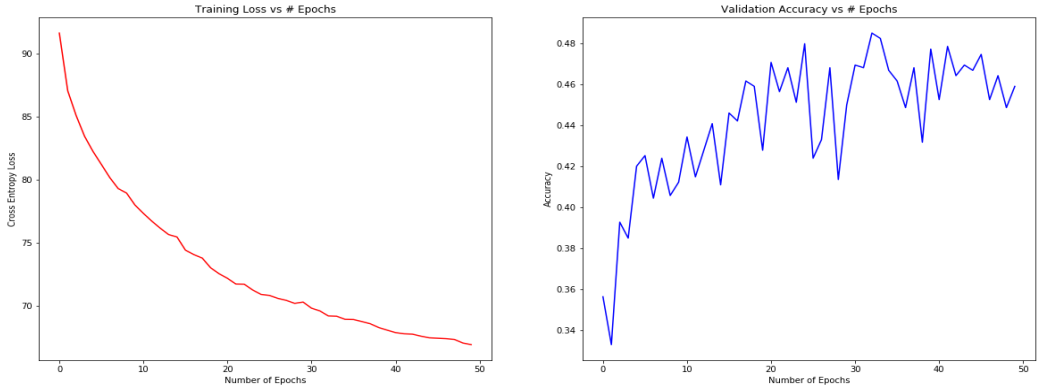

Figure 5: Plot of Training Loss and Validation Set Accuracy

Part 2.2:

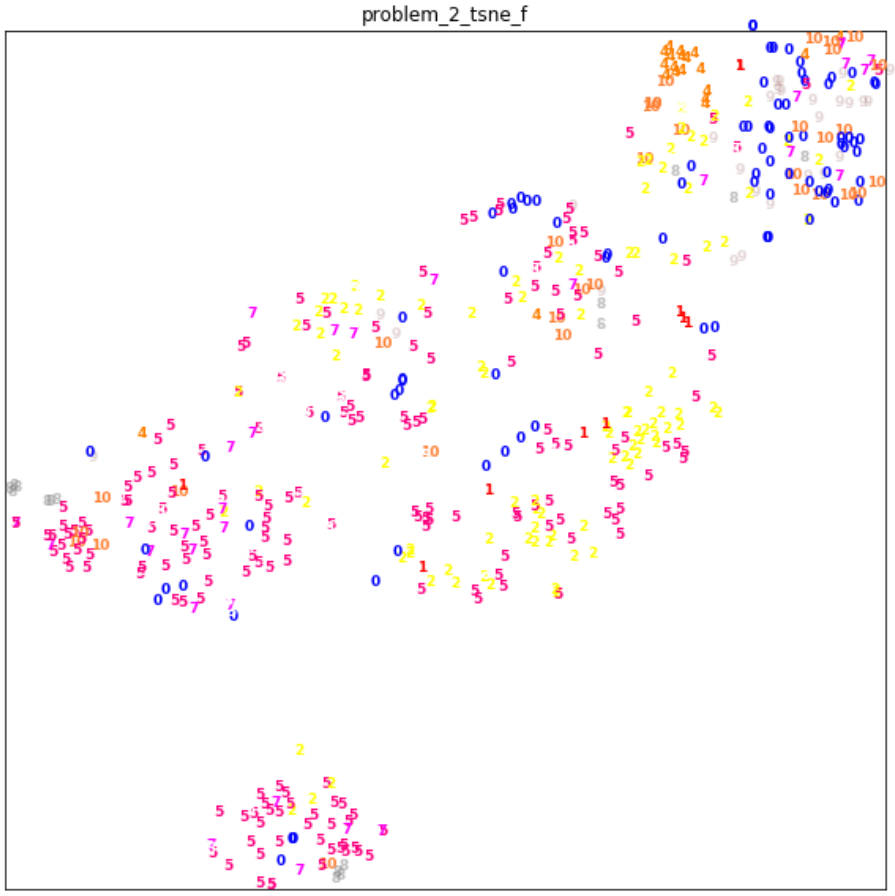

Figure 6: RNN-based video features on 2D space via t-SNE visualization

When compared with the CNN t-SNE visualization different actions seem to clustered closer to each other. The RNN model is better at classifying images since it can take account into dependencies coming with sequential nature of the data.

References and Collaborators

I used Google Colab's GPU: Tesla K80

1. https://zhuanlan.zhihu.com/p/34418001
2. https://github.com/eriklindernoren/Action-Recognition