

# Deep Learning for Computer Vision HW1

Pin-Han, Huang: R10946003

October 8, 2022

## 1 Problem1: Image Classification

### 1. Draw the network architecture of method A

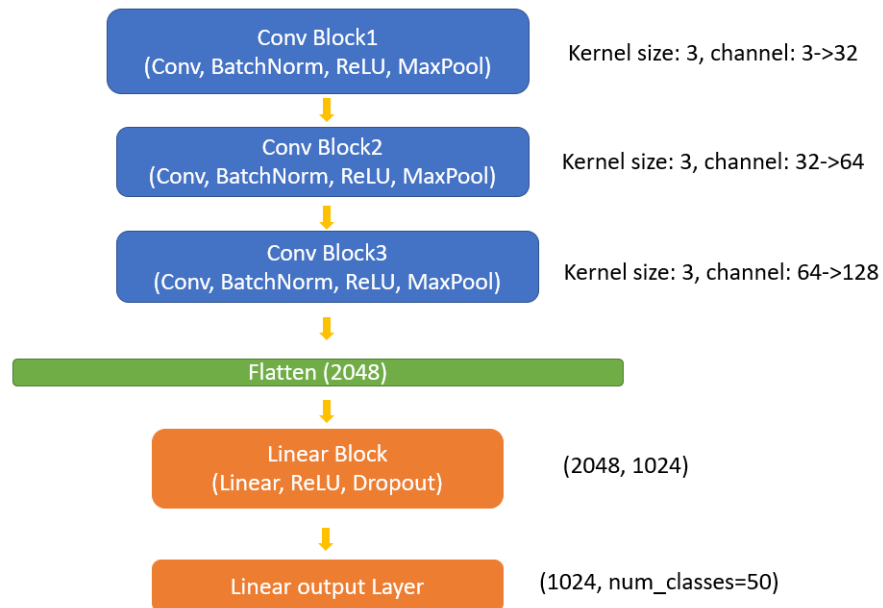


Figure 1: Three convolution blocks with two layers MLP.

2. Report accuracy of model A and B on the validation set.

- The accuracy of model A is **0.62** and the accuracy of model B is **0.88**.

3. Report the implementation detail of model A.

- Model A is trained 50 epochs with batch size = 250, learning rate =  $5e-4$  with weight decay =  $1e-5$  and cosine annealing learning rate decay. Early stop is implemented with 3 patient epochs. Lastly, focal loss is used as loss function as the replacement of cross entropy.

4. Report your alternative model and its difference from model A.

- Pretrained Efficientnetv2s is an improved version of the efficientnetB0 to efficientnetB7. The second version applies Fused-MBConv block to accelerate training time since the depthwise convolutional layers in original MBConv is time consuming in training models. Neural architecture search is performed to a baseline model to find the optimal width, depth, and resolution of the model. Efficientnetv2s contains a large amount of Fused-MBConv and MBConv blocks, a smaller kernel size ( $3 \times 3$ ). In comparison with other state-of-the-art models, Efficientnetv2s has 0.873 top-1 accuracy in ImageNet21K, with smaller model size and faster training time.
- The difference between model A and model B is that model A is simply a VGG-like basic CNN model while model B is a SENet,

MobileNet, ResNet back-boned with neural architecture search, efficient-wise model.

5. Visualize the learned representations of model A via PCA.

The PCA result of second last layer of model A is shown as follows:

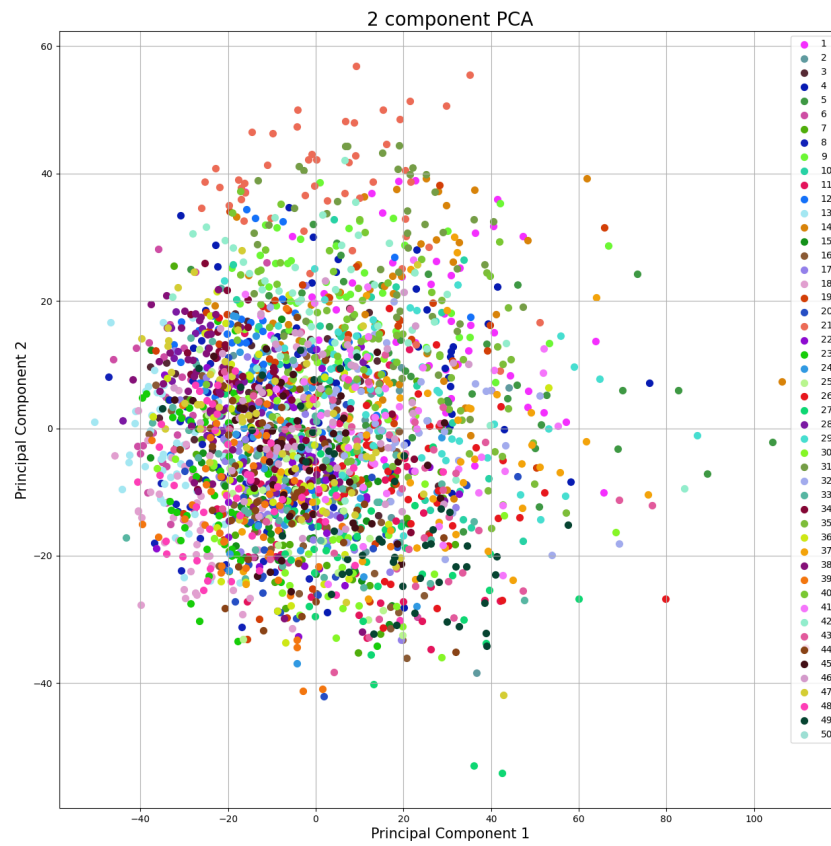


Figure 2: Second last layer PCA in model A.

- The original layer dimension is reduced to two dimension. We can observe that most of the points are overlapped and we can not visually distinguish different classes. We can merely notice that class 39 is relatively distributed in the top-right corner.

## 6. Visualize representations of model A via t-SNE.

The t-SNE result of second last layer of model A in three different epochs is shown as follows:

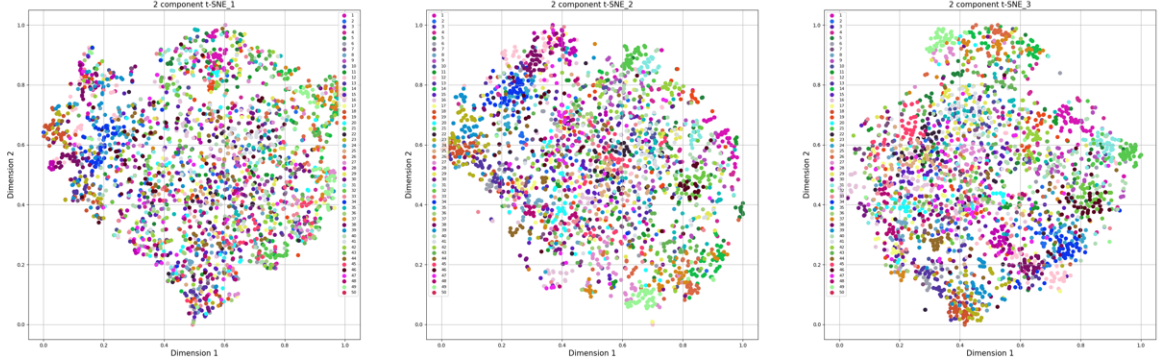


Figure 3: Second last layer t-SNE in model A.

- We can observe that the visually clustering result of t-SNE is better than PCA. In addition, as training epochs increases, same class of data are distributed relatively closer.

## 2 Problem2: Semantic Segmentation

### 1. Draw the network architecture of method A

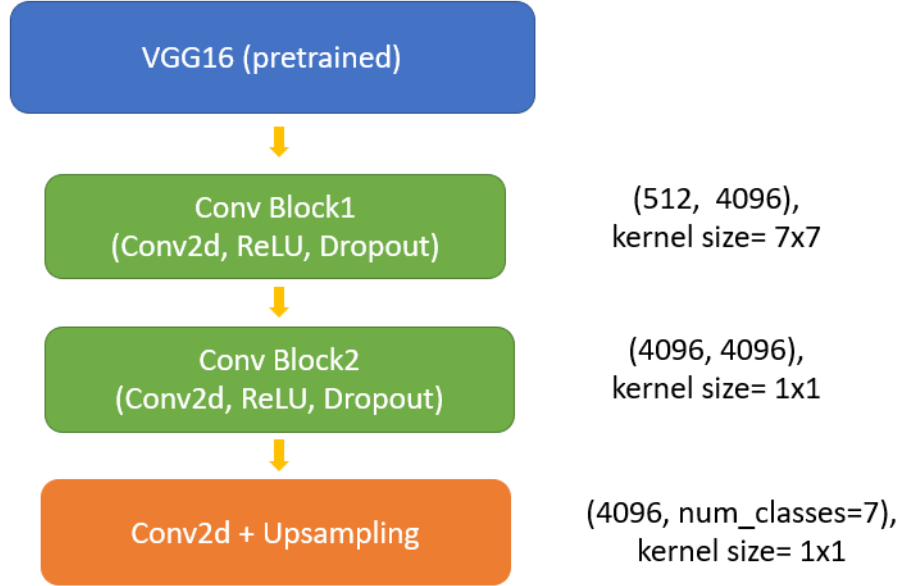


Figure 4: VGG16+FCN32s.

2. Draw the network architecture of method B

- The improved model B used is DeepLabV3 (shown in figure 5) with ResNet50 as backbone. Instead of using FCN, dilated convolution is applied after stage 2 to have a larger receptive field. Atrous Spatial Pyramid Pooling is implemented and is concatenated with a  $1 \times 1$  convolution before the output block.

3. Report mIoUs of two models on the validation set.

- The mIoU of model A is **0.57** and the mIoU of model B is **0.73**.

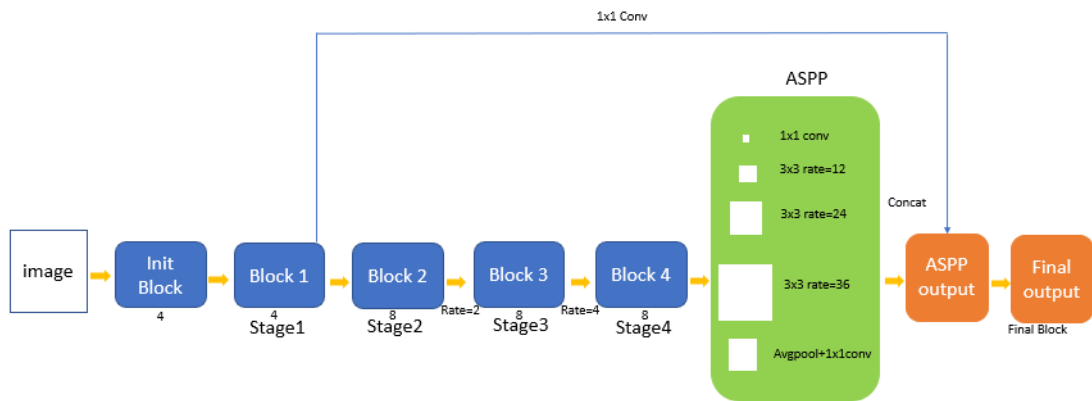


Figure 5: DeepLabV3

4. Show the predicted segmentation masks.

- The predicted segmentation mask of 0013, 0062, and 0104 during early, middle, and final stage of training process.

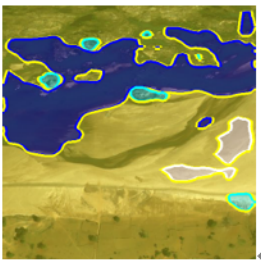
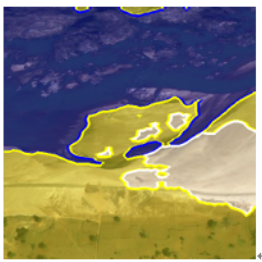
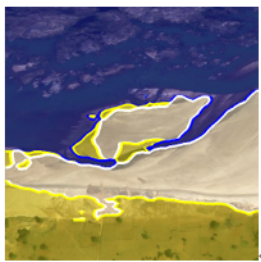






	Early Stage	Middle Stage	Final Stage
0013			
0062			
0104			

Figure 6: Prediction segmentation mask.