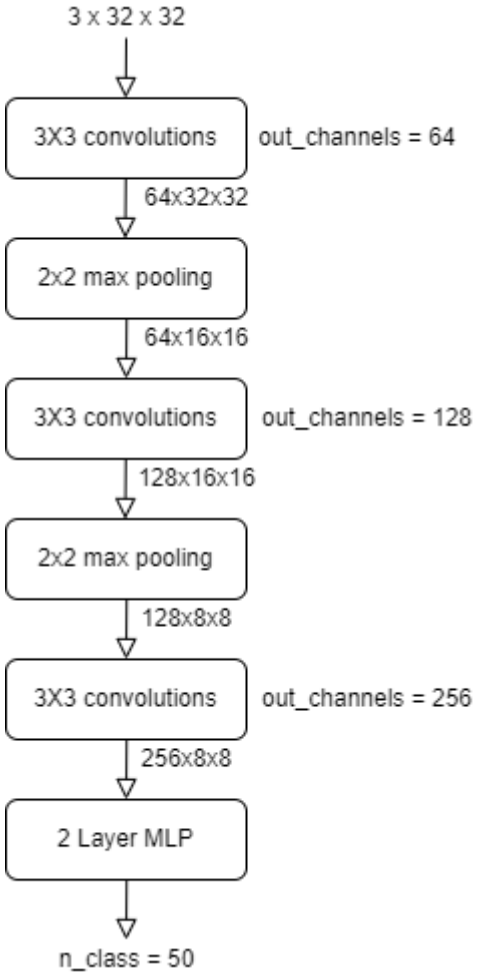
DLCV HW1 Report

資工所 呂兆凱 R11922098

Image classification

1. Draw the network architecture of method A or B.

   method A :



2. Report accuracy of your models (both A, B) on the validation set.

   Accuracy of model A : 50 %

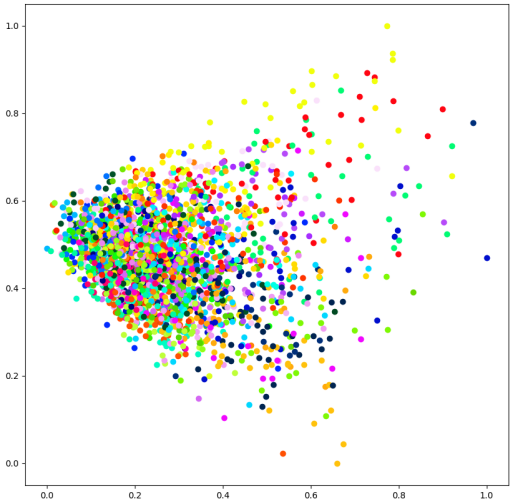
   Accuracy of model B : 96%

3. Report your implementation details of model A.

   - Model A中利用了三層的3x3 convolution和max pooling，以及最後加入2層的MLP
   - 訓練時使用Adam optimizer，learning rate為0.001，momentum為0.9
   - Loss function則使用Cross Entropy Loss

4. Report your alternative model or method in B, and describe its difference from model A.
   - Model B使用vgg 11，比起model A更為深，總共有8層的卷積層與3層的全連接層，其中也包含5個maxpool層。
   - 這次model B使用torch.hub裡面pretrained on CIFAR-100 dataset 的 vgg11 model，並用SGD optimizer與learning rate = 0.0001去對模型做fine-tune。
   - Pretrained- model reference : https://github.com/chenyaofo/pytorch-cifar-models

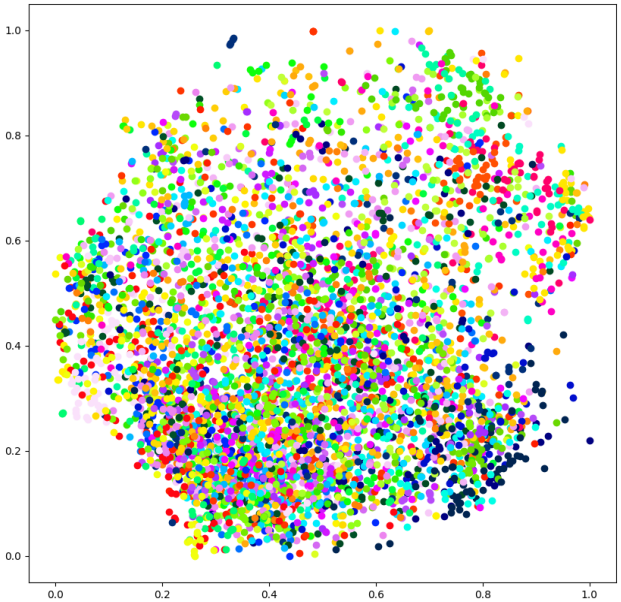
5. Visualize the learned visual representations of model A on the validation set by implementing PCA (Principal Component Analysis) on the output of the second last layer. Briefly explain your result of the PCA visualization.



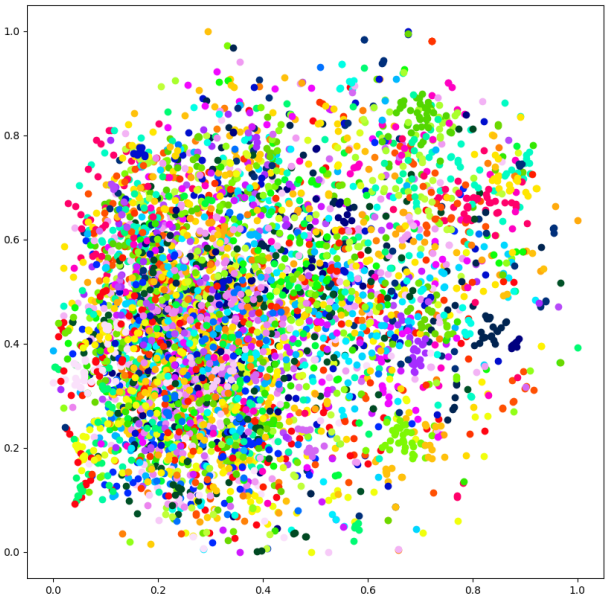
可以看到有些相同顏色的點會連續性的連成一塊區域，但群聚的成果不是非常明顯。與 t-SNE相比，在相同epoch上的結果，t-SNE的群聚成果較佳。

6. Visualize the learned visual representation of model A, again on the output of the second last layer, but using t-SNE (t-distributed Stochastic Neighbor Embedding) instead. Depict your visualization from three different epochs including the first one and the last one. Briefly explain the above results.

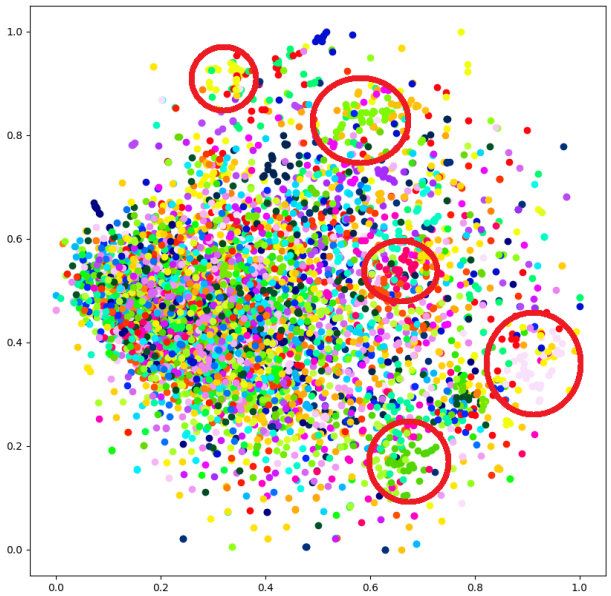
1st epoch : Accuracy : 21%



10-th epoch : Accuracy : 40%



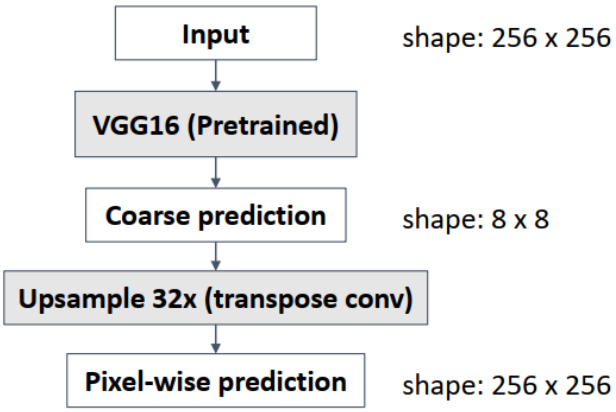
20-th epoch : Accuracy : 50%



可以看到第一個epoch所做出來的結果，雖然有些相同顏色的點已經開始有些群聚，但這些群聚的範圍都是較大的，甚至更多顏色的點是分散各處的。而做到最後一個epoch時，有些相同顏色的點已聚得很近，表示這時相同的圖片所被cnn層輸出的representation之間距離是相近的。

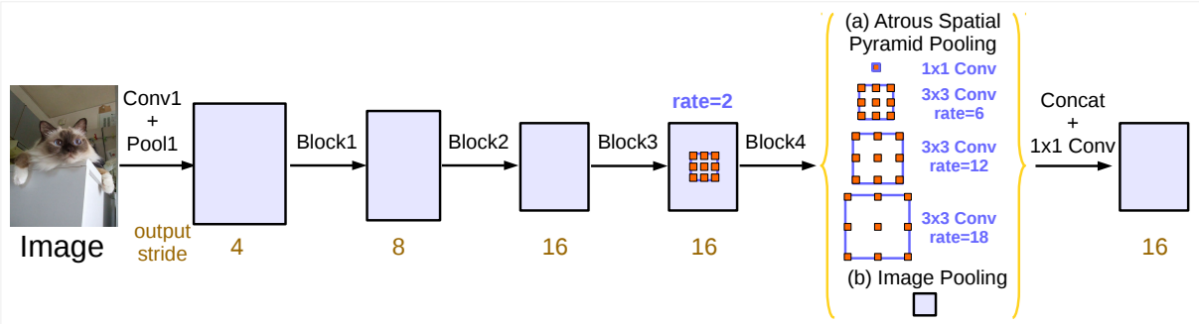
Semantic segmentation

1. Draw the network architecture of your VGG16-FCN32s model (model A).



2. Draw the network architecture of the improved model (model B) and explain it differs from your VGG16-FCN32s model.

model B : DeepLabv3

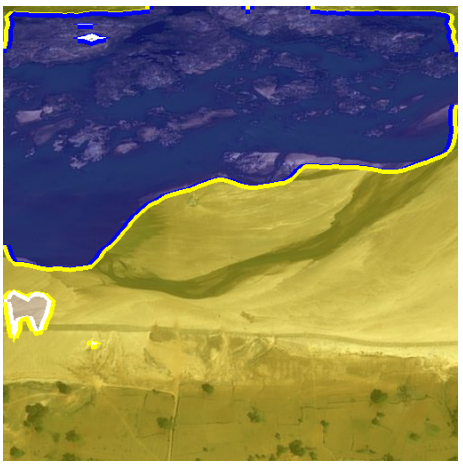


VGG16-FCN32s使用Encoder-Decoder的做法，而DeepLabv3則是使用Atrous Convolution。
Encoder-Decoder : 採用先downsample再upsample的方法，其計算量會較使用Atrous Convolution的方法來得少，但由於pooling會導致信息丟失，因此其效果較差。
Atrous Convolution : 空洞捲積，利用在kernel之中補零的方式，保持kernel的參數量卻又可以增大其Receptive Field，也就是可見範圍變廣。而DeepLabv3會再透過Atrous Spatial Pyramid Pooling (ASPP) 的方式，組合不同dilated rate的結果，以增進最後的分類效果。

3. Report mIoUs of two models on the validation set.
   mIoUs of model A : 0.584775
   mIoUs of model B : 0.734971

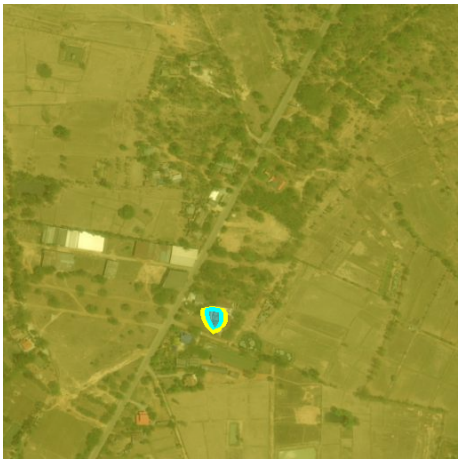
4. Show the predicted segmentation mask of “validation/0013_sat.jpg”, “validation/0062_sat.jpg”, “validation/0104_sat.jpg” during the early, middle, and the final stage during the training process of the improved model.

   Epoch : 1
   0013_sat.jpg



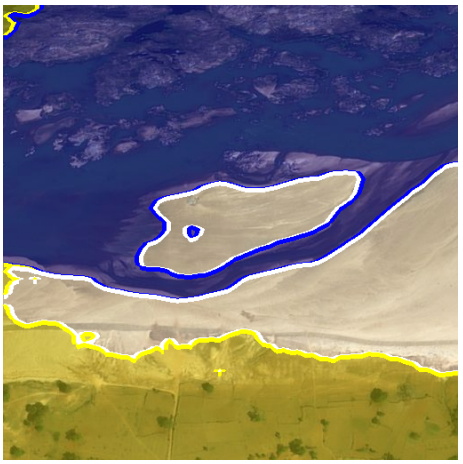
0062_sat.jpg

0104_sat.jpg



Epoch : 10

0013_sat.jpg

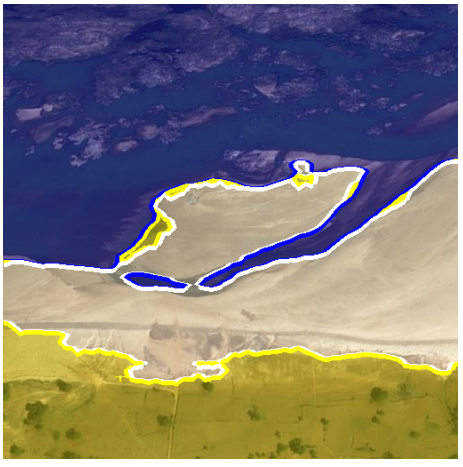


0062_sat.jpg

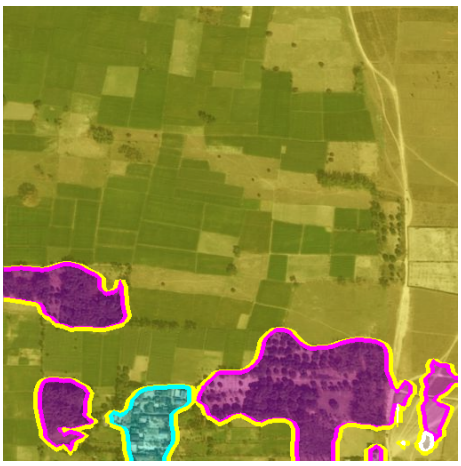
0104_sat.jpg

Epoch : 20

0013_sat.jpg



0062_sat.jpg



0104_sat.jpg