Report

BY LU XIAOYANG

1 Mandatory Task

1.1 Architecture of the network designed by myself

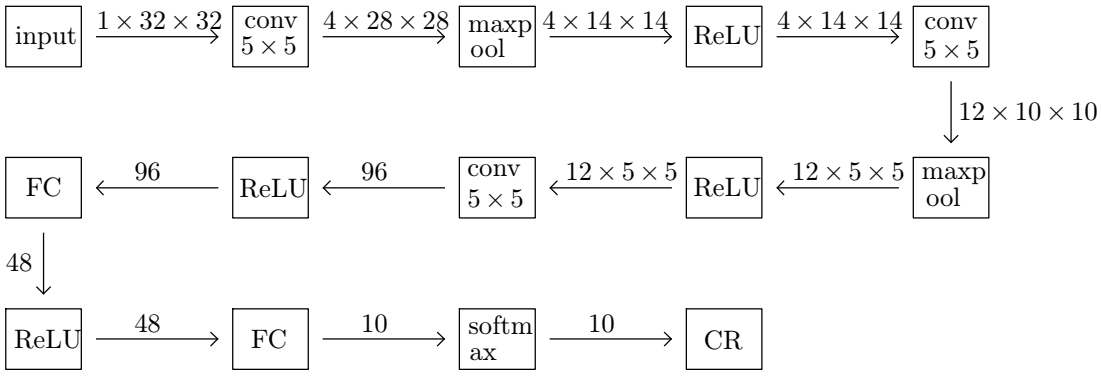

Figure 1. architecture

I just mimicked LeNet and use less parameters for faster training.

1.2 Learning curves

1.2.1 LeNet

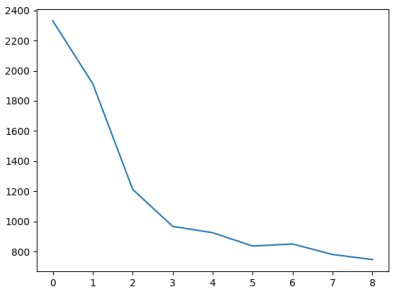

Figure 2. LeNet loss

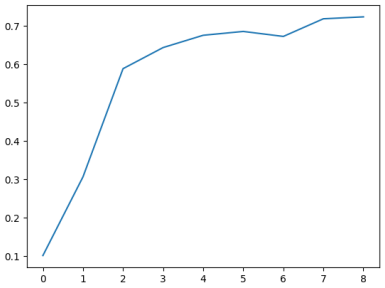

Figure 3. LeNet accuracy

1.2.2 Mine

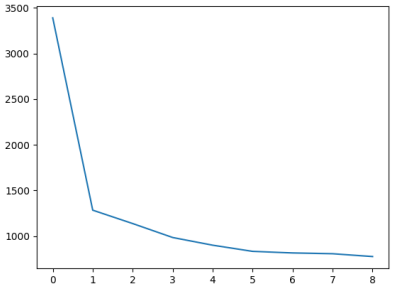

Figure 4. loss of mine

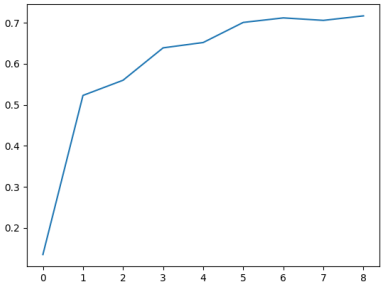

Figure 5. accuracy of mine

1.3 Visualization

1.3.1 PCA

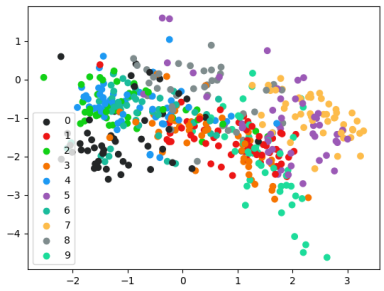

Figure 6. output of LeNet convolution layer

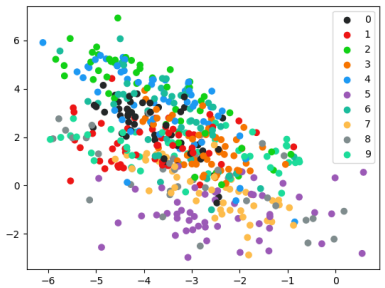

Figure 7. output of LeNet FC layer

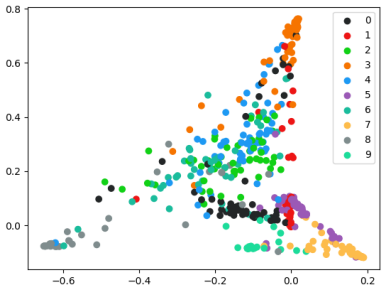

Figure 8. output of LeNet output layer

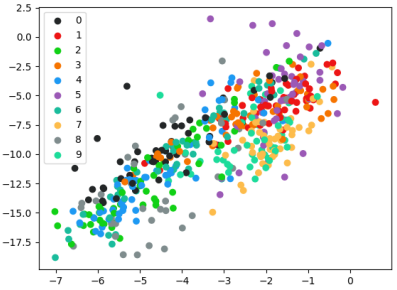

Figure 9. output of my convolution layer

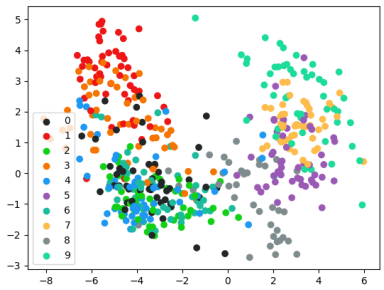

Figure 10. output of my FC layer

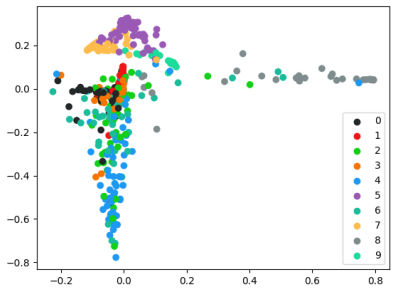

Figure 11. output of my output layer

1.3.2 t-SNE

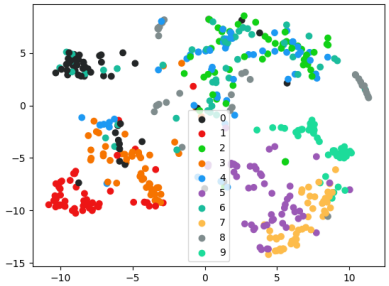

Figure 12. output of LeNet convolution layer

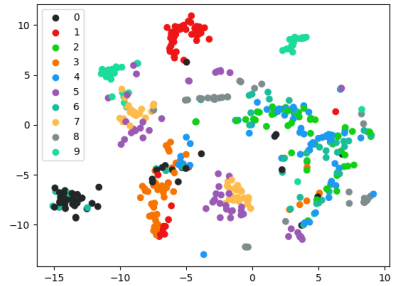

Figure 13. output of LeNet FC layer

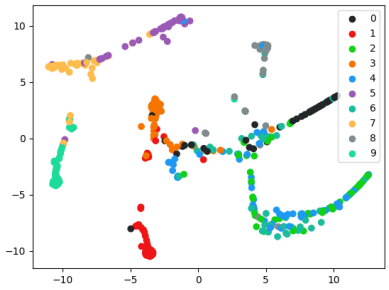

Figure 14. output of LeNet output layer

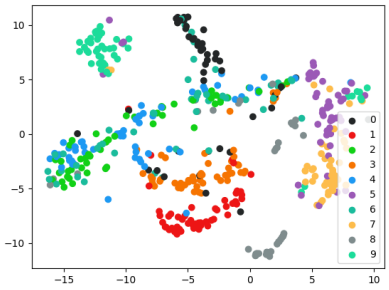

Figure 15. output of my convolution layer

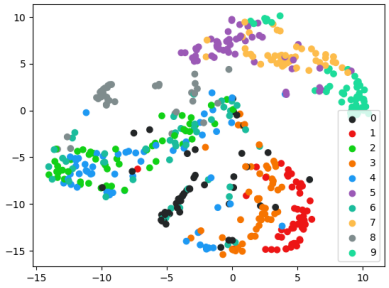

Figure 16. output of my FC layer

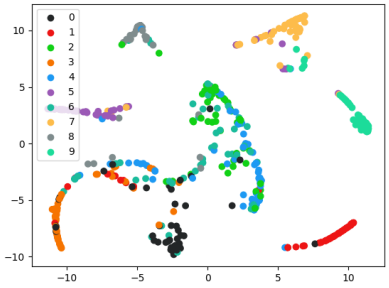

Figure 17. output of my output layer

From the diagrams I guess classes 2 and 6 are hard to distinguish.

2 Optional Task 1

2.1 Architecture

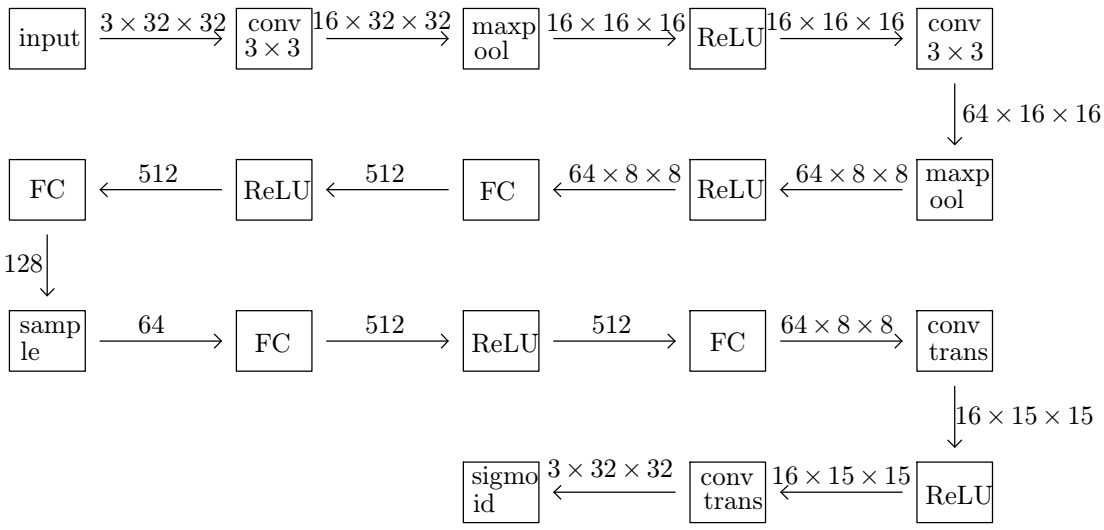

Figure 18. architecture

The structure of decoder is (almost) the inverse of that of encoder.

2.2 Implementation

I tried to reuse my code in **Task 1** but the training speed is unacceptable ($10 \sim 20$ hours; please refer to my code in `legacy/` for more details), so I turned to PyTorch. The structure of my code is based on the tutorial on their website (https://pytorch.org/tutorials/beginner/basics/quickstart_tutorial.html). Several observations:

- momentum in gradient descent can greatly speed up training. But in my original implementation (pure NUMPY), a large momentum can lead to undesirable results, for instance images in one pure colour (looks like a local minimum). A similar issue in LeNet is gradient vanishing if I choose sigmoid as activation functions.

- the factor to multiply KL divergence (or in fact, how strict do you measure the probability of the generated image conditional on the original image; or, the variance of the normal distribution corresponding to MSE loss) is important. The VAE will reconstruct images perfectly but perform badly on generating new ones if it's too small.

2.3 results

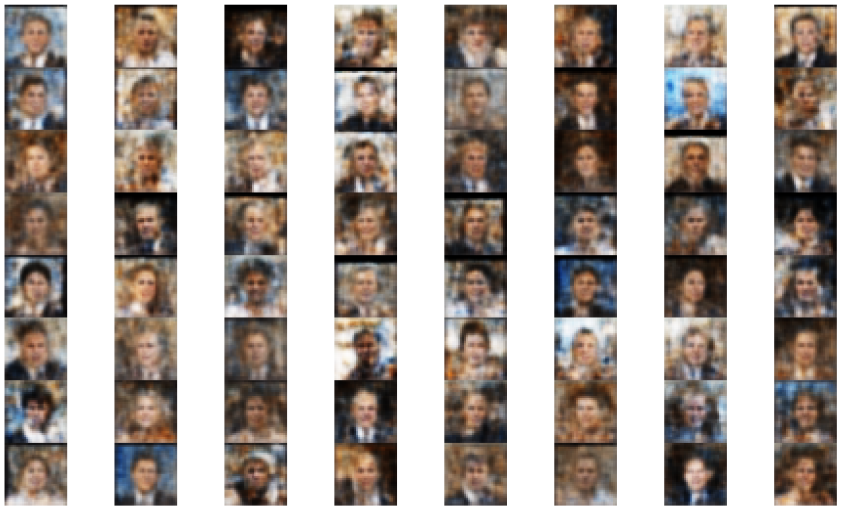

Figure 19. fake images

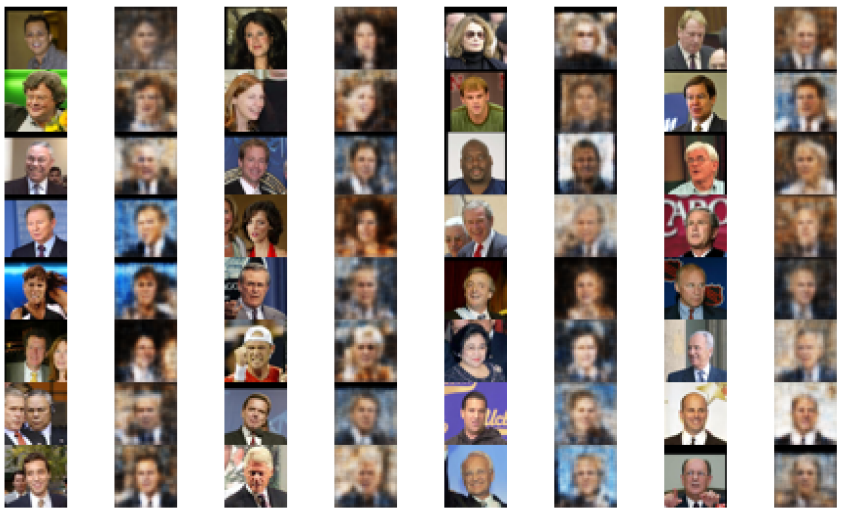

Figure 20. reconstructed images

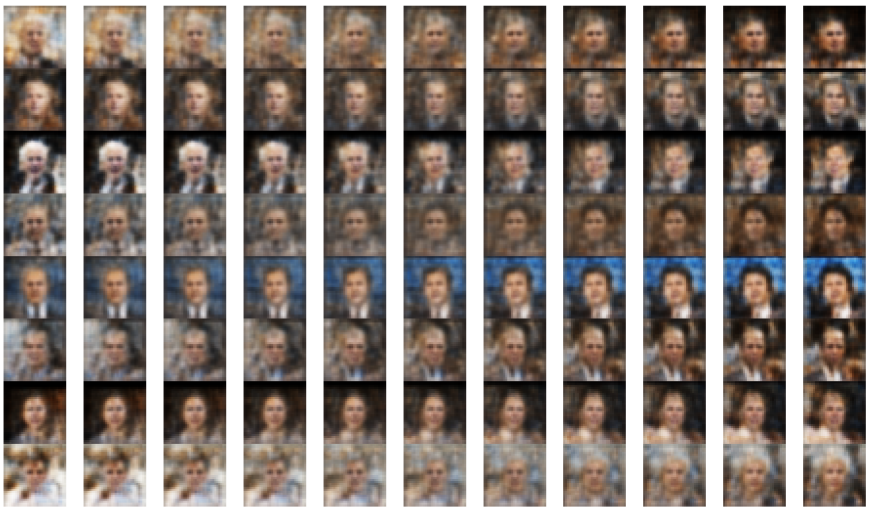

Figure 21. linear interpolation