Assignment 4 Transfer Learning, Image-Class Saliency Map Visualization

CSD4999 - John Protopsaltis

May 10, 2025

1 Transfer Learning

1.1 What happened

1.1.1 Case 1: Just replacing the last layer

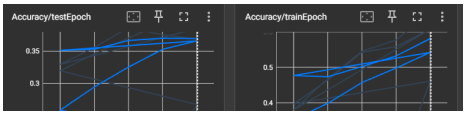

Figure 1: Training loss, accuracy, and test accuracy plots for Case 1

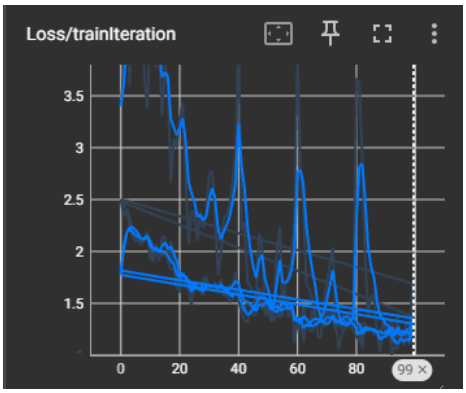

Figure 2: Training loss, accuracy, and test accuracy plots for Case 1

1.1.2 Case 2: Replacing the last two layers

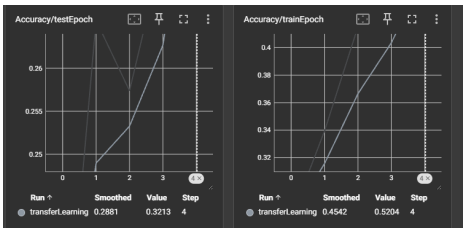

Figure 3: Training loss, accuracy, and test accuracy plots for Case 2

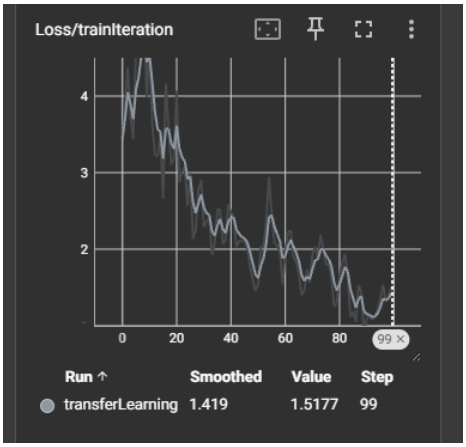

Figure 4: Training loss, accuracy, and test accuracy plots for Case 2

1.2 What I learned

1.2.1 Different approaches, slightly different results

When I just replaced the final FC layer (keeping that middle 4096-4096 layer intact), I got around 38% test accuracy. When I ditched both final layers and just used a single new 4096-10 layer, accuracy dropped a bit to about 32%.

This makes sense - the extra layer in the first approach gives the network more parameters to play with, allowing it to learn more complex relationships.

1.2.2 Why is the accuracy low?

- 4000 images is tiny for training a deep neural network
- Art styles are way harder to classify than regular objects - there's tons of variation within each style

- AlexNet was trained on normal photos, not paintings, so the features it learned might not transfer super well

Given all this, I think the 32-38% accuracy is actually reasonable.

1.3 Theoretical Questions

1.3.1 My dataset is small but similar to the original dataset. Should I fine-tune?

If your new dataset is small but similar to the original training data, you probably shouldn't do much fine-tuning. Better to just use the pre-trained network as a feature extractor and only train a simple classifier on top. Otherwise, you'll overfit.

1.3.2 My dataset is large and similar to the original dataset. Should I fine-tune or train from scratch?

With loads of similar data, fine-tuning makes sense. The pre-trained weights give you a head start, and you have enough data to safely update more layers.

1.3.3 My dataset is different from the original. Should I fine-tune?

When your new data is quite different, you should probably fine-tune more layers, maybe even the whole network. Those pre-trained weights still help as a starting point, especially for detecting basic features like edges and textures.

2 Image-Class Saliency Map Visualization

One problem I noticed is how neural networks don't always "look" at what we expect them to. When examining my saliency maps, I discovered some interesting behavior patterns in the network's attention:

1. For the flamingo image, the network got it wrong – it said "seashore" instead of "flamingo." The network was paying more attention to the sandy beach and water.
2. The cat image worked perfectly
3. With the husky image, the network got confused and classified it as an "Eskimo". The network was giving too much importance to the white background rather than the dog's features.
4. The Doberman classification went smoothly.
5. For the teddy bear, while the classification was correct, the saliency map showed something concerning – the network was relying heavily on background cues. This suggests our model might fail if we placed the same teddy on another environment.

This highlights a common issue in neural networks – they can get stuck in what we call "local minima," where they learn shortcuts based on correlations that don't actually matter.