Advanced Topics in Machine Learning

Assignment 0 - Report

8 September 2025

1 Introduction

This assignment covers foundational content for the course, starting from Skip Connections-based architectures of CNN models, i.e, ResNet-152, and moving forward with Vision Transformers (ViT), Generative Adversarial Networks (GAN), Variational Autoencoders, and lastly a modern multi-modal task-based model Contrastive Language-Image Pretraining (CLIP). This exercise helps in understanding the theoretical sides of these architectures and then implementing them with the framework PyTorch. Along the way, there were a few more things that came up, e.g., Upsampling of images to match the model's training dimensions, and looking into datasets like CIFAR-10. The core objectives of the first task revolve around fine-tuning the task-head of a model, either using pretraining weights or initializing random weights, and understanding how transfer learning works. Moving on with task 2, the core idea is to look inside the vision transformer, visualize attention maps, and play around with random and structured masking. Coming towards task 4, the objective of this task is to train a variational autoencoder on the MNIST dataset, visualizing reconstruction and generation, and posterior collapse investigation.

1.1 Task 1: Inner Workings of ResNet-152

The ResNet-152 is a deep Convolutional Neural Network (CNN) with 152 layers, with the last layer a Fully Connected Layer. The core idea of this architecture is Residual Connections or Skip Connections. The idea is to bypass a few layers and pass the input directly to the output of a layer. The goal behind this idea is to mitigate the problem of vanishing gradient by adding input to the output of a layer during forward propagation. When the back propagation calculates gradients, it does not face vanishing gradients. And thus, a very deep model is trained to achieve good accuracy. Once trained, the model can be utilized for tasks such as transfer learning. This task's core objectives include fine-tuning the task-head with pretrained weights and with randomly initialized weights to draw a comparison between the achieved accuracy, fine-tuning the fulll architecture, and disabling a few skip connections in the model architecture.

1.1.1 Methodology

- **Baseline Setup:** First of all, the model was imported from the PyTorch visionmodels module with defualt weights. After that, a dataset CIFAR-10 was imported from PyTorch and upsampled to 224. The CIFAR-10 dataset has images of dimension 32 by 32, while the model, RESTNET-152, is trained on the ImageNet dataset, which has images of dimension 224 by 224. As we train the classification head using the pretrained latent space, for the learned weights to map effectively to a new dataset, i.e., CIFAR10, the dimensions should be the same. Therefore, as seen in the 10 epochs, the performance is not really good. After that, the model's backbone was frozen, and the classification head was fine-tuned on this dataset.

- **Disable Residual Connections:** In this part of the task, the BottleNeck class responsible for Skip Connections implemntation in its forward funciton was inherited to define a new class having a forward function without skip connections. Then, in a few layers of the imported model, the Bottleneck was replaced with this new class to have a modified architecture.

- **Feature Hierarchies and Representations:** It involves getting intermediate features, i.e., early, middle, and late. For this purpose, one epoch of training was run on the validation set to retrieve features. And then they were plotted using the dimensionality reduction t-SNE technique.

- **Transfer Learning and Generalization:** This involved fine-tuning the model with default vs random weights. The fine-tuning was further branched into full fine-tuning vs classification head fine-tuning.

1.1.2 Results

- **Baseline Setup:** The class head was fine-tuned on CIFAR-10 with default weights. On (32, 32) dimensions of the image dataset, it gave poor results. Tweaking the dimensionality to (224, 244) resulted in good performance. The reason is that lower dimensionality causes a poor (very reduced) feature space, thus producing poor results.

- **Disable Residual Connections:** Changing the architecture resulted in worsening the model performance, which, though improved over 5 epochs slowly.

- **Feature Hierarchies and Representations:** After extracting early, middle, and late features and plotting the late ones, here is the result:

- **Transfer Learning and Generalization:** In this part, two different approaches for fine-tuning are implemented. One is full-finetuning and the other is fine-tuning the classification head only. Furthermore more for each approach, one-time default weights are used, and one-time random

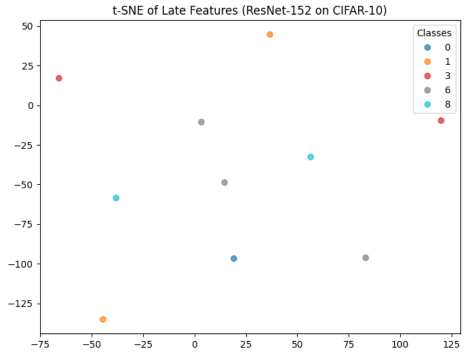

Figure 1: Late Features

weights are used. In both approaches, default weights fine-tuning gave good results. It shows the model's strength in generalizing to a new dataset once pre-trained fully on another dataset.

1.1.3 Discussion

Resizing CIFAR-10 images to 224×224 improved results, showing the need for inputs that match the pretrained model. Removing residual connections lowered accuracy, proving their importance in training deep networks. Transfer learning with pretrained weights worked better than random initialization, highlighting the value of prior knowledge. However, the study is limited by dataset size, few epochs, and partial removal of skip connections, so further testing is needed.

1.2 Task 2: Understanding Vision Transformers (ViT)

The Vision Transformer (ViT) is a deep learning model that applies the transformer architecture, originally built for natural language processing, to image data. Instead of using convolutional layers like CNNs, ViT breaks an image into small patches, flattens them, and treats each patch as a sequence input similar to words in a sentence. These patch embeddings are then processed through transformer layers with self-attention, which allows the model to capture both local and global relationships in the image. This design makes ViT highly effective for image classification and other vision tasks, especially when trained on large datasets, as it can learn powerful and flexible representations beyond what traditional CNNs provide.

1.2.1 Methodology

I selected the DeiT-Small model, pre-trained on ImageNet-1k from HuggingFace. The model works with inputs of size 224×224. After loading the model, inference was performed on sample inputs, and the predictions were found to be reasonable. Attention weights were then extracted, aggregated across heads using the mean, and converted into a clean attention map. Since the input is split into 16×16 patches, a 224×224 image forms a grid of 14×14 = 196 patches. The CLS token provides attention scores for each patch, resulting in a 196-length vector that can be reshaped back into a 14×14 grid to recover the spatial layout. Further analysis with masking showed that the model was still able to classify correctly, which indicates robustness to missing patches. This is because Vision Transformers use global self-attention and spread attention across many patches, allowing correct classification even when some regions are hidden. In this case, both the CLS token and mean pooling gave the same accuracy (0.964), showing that the model has learned strong and consistent global representations. The CLS token is designed to gather overall information during pretraining, while mean pooling averages over all patch embeddings. Since both perform equally well, it suggests that the pretrained ViT distributes information effectively across the CLS token and the patch tokens.

1.2.2 Results

When center masking was applied, the model still classified the input correctly as Saluki. This shows that the Vision Transformer is robust to missing patches, as it can correctly predict even when important regions are hidden. Both the CLS token and mean pooling achieved the same accuracy of 0.964, indicating strong and consistent performance.

1.2.3 Discussion

These results highlight the ability of Vision Transformers to use global self-attention effectively, distributing focus across multiple patches instead of relying on single regions. The equal performance of the CLS token and mean pooling suggests that the pretrained model captures global information in both representations. This shows that ViTs not only generalize well but also maintain robustness when parts of the input are missing.

2 Conclusion

This report explores different machine learning models, including ResNet-152, Vision Transformers, and explains how these models work, their strengths, and how they were tested on datasets like CIFAR-10 and MNIST. Experiments showed that pretrained models work better than training from scratch, and residual connections are important for deep networks. Vision Transformers were found to be robust even when parts of images were hidden. Overall, the report highlights how modern architectures learn strong and flexible representations for different tasks.

Repository: GitHub Repository