Transformer-based Self-Supervision on MNIST

Anand Kumar

Chapter 1

Masked Autoencoder Vision Transformer (MAE ViT) on MNIST Dataset

1.1 Introduction

This project implements a Masked Autoencoder Vision Transformer (MAE ViT) to classify specific digits from the MNIST dataset. The goal is to leverage the self-supervised learning capability of MAE for feature extraction and subsequently fine-tune the model for supervised classification.

1.2 Dataset

The MNIST dataset consists of 70,000 grayscale images of handwritten digits (0-9), each of size 28x28 pixels. For this project, we selected only three classes (0, 1, and 2) and sampled 100 images per class for both training and testing, resulting in a total of 600 training samples and 600 testing samples.

1.2.1 Data Preparation

- **Transformations:** The images were transformed into tensor format.
- **Sampling:** A custom function sampled specific digits, ensuring a balanced dataset for the selected classes.

1.3 Model Architecture

The architecture comprises several key components:

1.3.1 Multi-Head Self-Attention

This component allows the model to attend to different parts of the input simultaneously, enhancing its ability to capture long-range dependencies in the data.

1.3.2 FeedForward Network

A simple feedforward network with GELU activation and dropout layers processes the output from the attention layer to add non-linearity.

1.3.3 Vision Transformer Block

Each block integrates multi-head attention and feedforward layers, with residual connections and layer normalization to stabilize training.

1.3.4 Masking Function

Images were divided into patches, with a specified ratio of patches masked (set to zero). This encourages the model to learn effective representations by reconstructing the masked patches during training.

1.3.5 Overall Model

The MAE ViT model includes a patch embedding layer, a series of transformer blocks, a reconstruction layer, and a classification layer.

1.4 Training Strategy

1.4.1 Pretraining

During pretraining:

- The model learns to reconstruct masked patches from unmasked ones using Mean Squared Error (MSE) loss.
- The training runs for 15 epochs with Adam optimizer.

1.4.2 Fine-tuning

The model is fine-tuned on the classification task:

- Cross-Entropy Loss is used to train the model on labeled data.
- The training process continues for another 15 epochs.

1.5 Results

1.5.1 Training Loss

The training loss decreased over epochs, indicating that the model effectively learned to reconstruct masked patches during pretraining and improved its classification accuracy during fine-tuning.

1.5.2 Test Accuracy

Upon evaluation, the model achieved a test accuracy of **99%**.

1.5.3 Confusion Matrix

The confusion matrix displayed the performance of the model across the three classes, highlighting areas of misclassification.

```
Epoch 1/15, Loss: 0.06652068980038166
Epoch 2/15, Loss: 0.021728174574673177
Epoch 3/15, Loss: 0.012168906349688768
Epoch 4/15, Loss: 0.008198781823739409
Epoch 5/15, Loss: 0.006216342234984041
Epoch 6/15, Loss: 0.005066391779109836
Epoch 7/15, Loss: 0.004266713256947696
Epoch 8/15, Loss: 0.0037377104396000504
Epoch 9/15, Loss: 0.003348561027087271
Epoch 10/15, Loss: 0.0030402599833905695
Epoch 11/15, Loss: 0.0027514635119587185
Epoch 12/15, Loss: 0.0025389221729710696
Epoch 13/15, Loss: 0.0023511125007644297
Epoch 14/15, Loss: 0.002171503659337759
Epoch 15/15, Loss: 0.0020181869389489294



Epoch 1/15, Train Loss: 2.6977170944213866
Epoch 2/15, Train Loss: 0.8205160439014435
Epoch 3/15, Train Loss: 0.36533886194229126
Epoch 4/15, Train Loss: 0.11246921829879283
Epoch 5/15, Train Loss: 0.060719408979639414
Epoch 6/15, Train Loss: 0.019734630733728407
Epoch 7/15, Train Loss: 0.013750673073809594
Epoch 8/15, Train Loss: 0.009313967626076192
Epoch 9/15, Train Loss: 0.010129339795093983
Epoch 10/15, Train Loss: 0.007930704415775836
Epoch 11/15, Train Loss: 0.0038162920449394734
Epoch 12/15, Train Loss: 0.0022549850720679386
Epoch 13/15, Train Loss: 0.000502196792513132
Epoch 14/15, Train Loss: 0.0006468929917900823
Epoch 15/15, Train Loss: 0.00040223480318672954
Test Accuracy: 99.00%

Confusion Matrix:
 [[100   0   0]
 [  0  98   2]
```

Figure 1.1:

Chapter 2

Vision Transformer with InfoNCE Loss on MNIST Dataset

2.1 Introduction

This project implements a Vision Transformer (ViT) for classifying specific digits from the MNIST dataset using InfoNCE loss. The model leverages self-supervised learning to extract features and fine-tunes for supervised classification.

2.2 Dataset

The MNIST dataset consists of 70,000 grayscale images of handwritten digits (0-9), each of size 28x28 pixels. For this project, only three classes (0, 1, and 2) were selected, with 100 images sampled per class for training and testing, resulting in 600 training samples and 600 testing samples.

2.2.1 Data Preparation

- **Transformations:** Images were converted to tensor format.
- **Sampling:** A custom function was used to sample specific digits, ensuring a balanced dataset.

2.3 Model Architecture

The architecture consists of several key components:

2.3.1 Vision Transformer (ViT)

The model includes:

- **Patch Embedding:** A convolutional layer to extract patch embeddings from input images.
- **CLS Token:** A learnable token representing the whole sequence.
- **Position Embedding:** Adds positional information to the token sequence.
- **Transformer Blocks:** Sequential layers of multi-head self-attention and feedforward networks.
- **Classification Layer:** A linear layer used for fine-tuning the model.

2.3.2 InfoNCE Loss

The InfoNCE loss encourages the model to distinguish between similar and dissimilar instances in the embedding space by normalizing the CLS token embeddings and calculating the cosine similarity.

2.4 Training Strategy

2.4.1 Pretraining with InfoNCE Loss

During pretraining:

- The model learns to maximize the similarity of CLS tokens from augmented versions of the same image while minimizing the similarity with tokens from different images.
- Training runs for 15 epochs with the Adam optimizer.

2.4.2 Fine-tuning for Classification

The model is fine-tuned using cross-entropy loss for the classification task:

- Cross-entropy loss is computed using the outputs from the classification layer.
- The fine-tuning process also lasts for 15 epochs.

2.5 Results

2.5.1 Training Loss

The InfoNCE loss decreased over epochs, indicating effective learning during pretraining.

2.5.2 Test Accuracy

Upon evaluation, the model achieved a test accuracy of **98.67%** (to be filled in with actual results).

2.5.3 Confusion Matrix

The confusion matrix illustrated the model's performance across the three classes, showing areas of misclassification.

2.6 Introduction

This project implements a Vision Transformer (ViT) for classifying specific digits from the MNIST dataset, incorporating Masked Autoencoding (MAE) and InfoNCE loss to enhance feature representation and classification accuracy.

2.7 Dataset

The MNIST dataset consists of 70,000 grayscale images of handwritten digits (0-9), each sized at 28x28 pixels. For this project, only three classes (0, 1, and 2) were selected, with 100 images sampled per class for training and testing, resulting in 600 training samples and 600 testing samples.

2.7.1 Data Preparation

- **Transformations:** Images were transformed into tensor format.
- **Sampling:** A custom sampling function ensured a balanced dataset of selected classes.

2.8 Model Architecture

The architecture comprises several essential components:

2.8.1 Vision Transformer (ViT) with MAE and InfoNCE

The model is structured as follows:

- **Patch Embedding:** A convolutional layer that extracts patch embeddings from input images.
- **CLS Token:** A learnable token representing the entire sequence.
- **Position Embedding:** Provides positional information to the token sequence.

```
Epoch 1/15, InfoNCE Loss: 1.9085472762584685
Epoch 2/15, InfoNCE Loss: 1.7500728726387025
Epoch 3/15, InfoNCE Loss: 1.7175899982452392
Epoch 4/15, InfoNCE Loss: 1.7027308344841003
Epoch 5/15, InfoNCE Loss: 1.6804448306560515
Epoch 6/15, InfoNCE Loss: 1.6905959486961364
Epoch 7/15, InfoNCE Loss: 1.6814386427402497
Epoch 8/15, InfoNCE Loss: 1.6814820528030396
Epoch 9/15, InfoNCE Loss: 1.677879160642624
Epoch 10/15, InfoNCE Loss: 1.6728041827678681
Epoch 11/15, InfoNCE Loss: 1.6604488492012024
Epoch 12/15, InfoNCE Loss: 1.6577719748020172
Epoch 13/15, InfoNCE Loss: 1.658681285381317
Epoch 14/15, InfoNCE Loss: 1.6373381555080413
Epoch 15/15, InfoNCE Loss: 1.6481409788131713



Epoch 1/15, Train Loss: 0.39963886328041553
Epoch 2/15, Train Loss: 0.03929458229104057
Epoch 3/15, Train Loss: 0.006768366030883044
Epoch 4/15, Train Loss: 0.0009841733990469947
Epoch 5/15, Train Loss: 0.00018410865486657714
Epoch 6/15, Train Loss: 0.00014942258385417518
Epoch 7/15, Train Loss: 0.00010025337260231027
Epoch 8/15, Train Loss: 8.84843415406067e-05
Epoch 9/15, Train Loss: 6.60626067656267e-05
Epoch 10/15, Train Loss: 5.6612152911839074e-05
Epoch 11/15, Train Loss: 4.2699315781646877e-05
Epoch 12/15, Train Loss: 5.2386614834176724e-05
Epoch 13/15, Train Loss: 3.592259099605144e-05
Epoch 14/15, Train Loss: 3.7741099140475855e-05
Epoch 15/15, Train Loss: 3.626326015364612e-05
Test Accuracy: 98.67%

Confusion Matrix:
 [[100   0   0]
 [  0  98   2]
```

Figure 2.1:

- **Transformer Blocks:** A sequence of layers comprising multi-head self-attention and feedforward networks.

- **Decoder:** A linear layer for reconstructing masked patches in MAE.
- **Classifier:** A linear layer for fine-tuning and classification.

2.8.2 Loss Functions

The combined loss function consists of:

- **MAE Loss:** Measures the reconstruction error of masked patches.
- **InfoNCE Loss:** Encourages the model to distinguish between similar and dissimilar embeddings using the CLS token.
- **Combined Loss:** A weighted sum of the MAE and InfoNCE losses, allowing for balanced training.

2.9 Training Strategy

2.9.1 Pretraining with MAE and InfoNCE Loss

During pretraining:

- The model learns to reconstruct masked patches while also maximizing similarity among CLS token embeddings.
- Training runs for 15 epochs using the Adam optimizer.

2.9.2 Fine-tuning for Classification

The model is fine-tuned using cross-entropy loss:

- The cross-entropy loss is calculated using the outputs from the classification layer.
- The fine-tuning process also lasts for 15 epochs.

2.10 Results

2.10.1 Training Loss

The combined loss of MAE and InfoNCE decreased over epochs, indicating effective learning during pretraining.

2.10.2 Test Accuracy

Upon evaluation, the model achieved a test accuracy of **98%** (to be filled in with actual results).

2.10.3 Confusion Matrix

The confusion matrix illustrates the model's performance across the three classes, highlighting areas of misclassification.

```
Epoch 1/15, Combined Loss (MAE + InfoNCE): 0.4298112213611603
Epoch 2/15, Combined Loss (MAE + InfoNCE): 0.38169167190790176
Epoch 3/15, Combined Loss (MAE + InfoNCE): 0.36468955874443054
Epoch 4/15, Combined Loss (MAE + InfoNCE): 0.35817169547080996
Epoch 5/15, Combined Loss (MAE + InfoNCE): 0.351166270673275
Epoch 6/15, Combined Loss (MAE + InfoNCE): 0.35008817464113234
Epoch 7/15, Combined Loss (MAE + InfoNCE): 0.34758263528347016
Epoch 8/15, Combined Loss (MAE + InfoNCE): 0.34628394097089765
Epoch 9/15, Combined Loss (MAE + InfoNCE): 0.34064235985279084
Epoch 10/15, Combined Loss (MAE + InfoNCE): 0.33858578503131864
Epoch 11/15, Combined Loss (MAE + InfoNCE): 0.3400235965847969
Epoch 12/15, Combined Loss (MAE + InfoNCE): 0.3398679971694946
Epoch 13/15, Combined Loss (MAE + InfoNCE): 0.34004348069429396
Epoch 14/15, Combined Loss (MAE + InfoNCE): 0.33851189613342286
Epoch 15/15, Combined Loss (MAE + InfoNCE): 0.33613596111536026



Epoch 1/15, Train Loss: 0.44645991930738094
Epoch 2/15, Train Loss: 0.03944973384932382
Epoch 3/15, Train Loss: 0.019647303910460323
Epoch 4/15, Train Loss: 0.05055731045540597
Epoch 5/15, Train Loss: 0.004878741052743862
Epoch 6/15, Train Loss: 0.002473948434453632
Epoch 7/15, Train Loss: 0.003054968750075204
Epoch 8/15, Train Loss: 0.00047998256231949197
Epoch 9/15, Train Loss: 0.0008290031980322965
Epoch 10/15, Train Loss: 0.00016779703455540584
Epoch 11/15, Train Loss: 0.00010879741287226353
Epoch 12/15, Train Loss: 4.938644860885688e-05
Epoch 13/15, Train Loss: 3.04537356896617e-05
Epoch 14/15, Train Loss: 3.393970503111632e-05
Epoch 15/15, Train Loss: 3.327038602947141e-05
Test Accuracy: 98.00%

Confusion Matrix:
 [[100   0   0]
 [  0  98   2]
```

Figure 2.2:

2.11 Introduction

This project explores the use of a Video Masked Autoencoder (MAE) for classifying specific digits from the MNIST dataset. The model leverages temporal information by treating the input images as sequences of frames, enhancing feature representation for improved classification performance.

2.12 Dataset

The MNIST dataset comprises 70,000 grayscale images of handwritten digits (0-9), with each image sized at 28x28 pixels. For this study, three classes (0, 1, and 2) are selected, resulting in a balanced dataset of 600 training samples and 600 testing samples.

2.12.1 Data Preparation

- **Transformations:** Images are transformed into tensor format.
- **Sampling:** A custom sampling function is implemented to select a specified number of samples per class.

2.13 Model Architecture

The architecture consists of the following components:

2.13.1 Video MAE Model

The model is designed to process video-like inputs where each image is treated as a frame:

- **Patch Embedding:** A convolutional layer extracts patches from each frame.
- **CLS Token:** A learnable token representing the entire sequence.
- **Positional Embedding:** Provides positional context to the token sequence.
- **Transformer Layers:** A series of transformer blocks for processing the tokens.
- **Decoder:** Reconstructs masked patches for the MAE task.
- **Classifier:** A linear layer for fine-tuning and classification.

2.13.2 Loss Functions

The model employs L1 loss for the reconstruction of masked tokens during pretraining. The classification loss (cross-entropy loss) is utilized during fine-tuning.

2.14 Training Strategy

2.14.1 Pretraining with Video MAE

During the pretraining phase:

- Each input image is replicated to simulate three frames.
- Random tokens are masked to train the model on reconstruction tasks.
- The training runs for 20 epochs using the Adam optimizer.

2.14.2 Fine-tuning for Classification

In the fine-tuning phase:

- The model is trained on the classification task using the cross-entropy loss.
- Fine-tuning also lasts for 20 epochs.

2.15 Results

2.15.1 Training Loss

The pretraining loss decreased steadily over epochs, indicating effective learning.

2.15.2 Test Accuracy

The model achieved a test accuracy of **97%** (replace with actual results).

2.15.3 Confusion Matrix

The confusion matrix illustrates model performance across the three classes, highlighting areas for improvement.

```
Epoch 1/15, Pretraining Loss: 0.1785506833540766
Epoch 2/15, Pretraining Loss: 0.12507327881298566
Epoch 3/15, Pretraining Loss: 0.0990088311465163
Epoch 4/15, Pretraining Loss: 0.08159578199449338
Epoch 5/15, Pretraining Loss: 0.06819076581220877
Epoch 6/15, Pretraining Loss: 0.05648602917790413
Epoch 7/15, Pretraining Loss: 0.047559505034434166
Epoch 8/15, Pretraining Loss: 0.03970790281891823
Epoch 9/15, Pretraining Loss: 0.033669185677641315
Epoch 10/15, Pretraining Loss: 0.028426068001671842
Epoch 11/15, Pretraining Loss: 0.024199749294080232
Epoch 12/15, Pretraining Loss: 0.020845749856610047
Epoch 13/15, Pretraining Loss: 0.018305804580450058
Epoch 14/15, Pretraining Loss: 0.01576484288824232
Epoch 15/15, Pretraining Loss: 0.013948314921244195



Epoch 1/15, Train Loss: 0.9132978194638303
Epoch 2/15, Train Loss: 0.139540547701089
Epoch 3/15, Train Loss: 0.07354439404106845
Epoch 4/15, Train Loss: 0.06267643017130659
Epoch 5/15, Train Loss: 0.06070152276017899
Epoch 6/15, Train Loss: 0.07813960807872813
Epoch 7/15, Train Loss: 0.0849825587999811
Epoch 8/15, Train Loss: 0.008329189428446912
Epoch 9/15, Train Loss: 0.00108689245596332
Epoch 10/15, Train Loss: 0.0007009258651554486
Epoch 11/15, Train Loss: 0.00037358876235023334
Epoch 12/15, Train Loss: 0.0002901060613926108
Epoch 13/15, Train Loss: 0.00023939999702729677
Epoch 14/15, Train Loss: 0.00020825333118218144
Epoch 15/15, Train Loss: 0.0001940380195599956
Test Accuracy: 98.00%

Confusion Matrix:
 [[99  0  1]
 [ 0 99  1]
```

Figure 2.3: