SSL Project with FashionMNIST and pretraining

Erik Myhre
Nicolai Molstad

November 13, 2023

SSL Project

This is the short report for our project in the module SSL (TDT05). Code can be found here

The ViT model was heavily influenced by this github repository [2]

What is the project about?

In the project we trained a vision transformer (ViT) [2] model on the FashionMNIST dataset. The goal was to perform self-supervised pretraining on the model before fine-tuning it on the FashionMNIST dataset. To achieve this we wanted to do a pretext task such as rotation prediction.

Which part did we use self-supervised learning (SSL)?

We pretrained the encoder with the rotationprediction [1] and then fine-tuned the classification head to the FashionMNIST dataset.

In the first experiment only the decoder was fine tuned and the encoder was frozen. This means that the encoder is random for the model without pretraining. In the second experiment both the encoder and decoder was fine-tuned. This goes for both the pretrained and non-pretrained encoder.

How is the comparison against the solution without SSL?

As we can see in the figures below; at the first experiment the model with pretaining performed at 79% while the one without performed at 64%. A difference of 15% points.

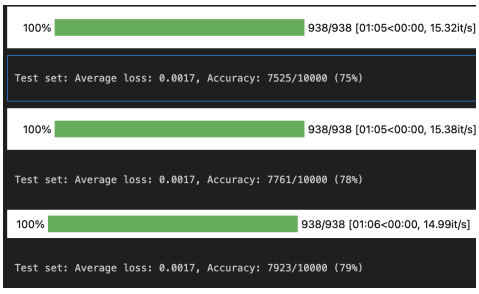

Figure 1: With pretrained encoder

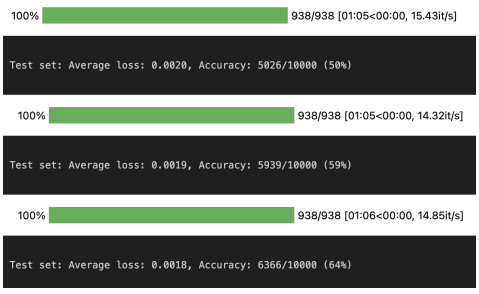

Figure 2: Without pretrained encoder

Here is the loss and accuracy for the pretrained model.

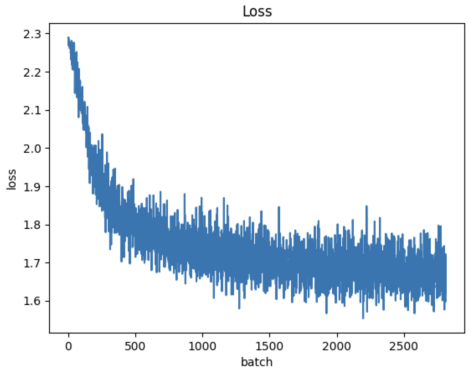

Figure 3: Loss

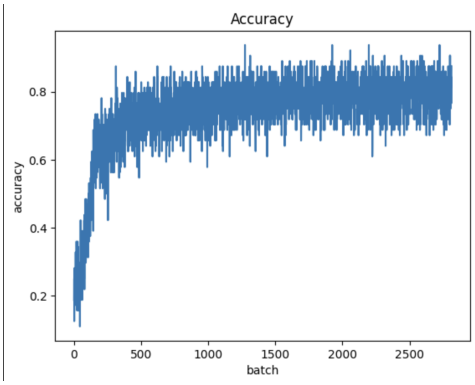

Figure 4: Accuracy

The second experiment did not show as much af an improvement. As we can see in the figure 5

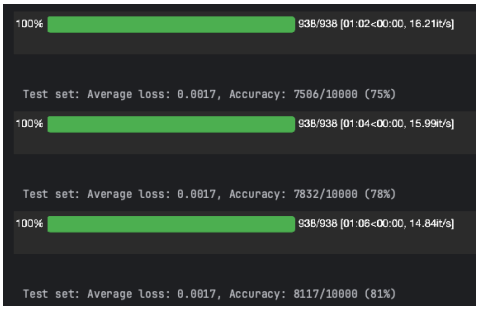

Figure 5: With pretrained encoder

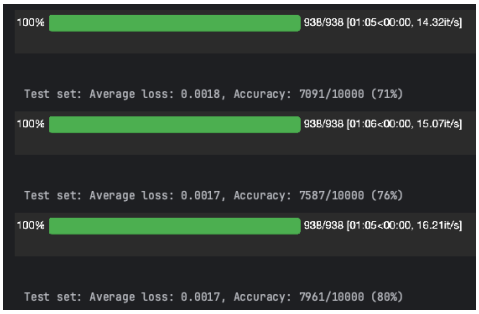

Figure 6: Without pretrained encoder

This is only an improvement of 1% point, however it is an improvement. It is also a little facinating to see that the model with only the decoder fine-tuned (Figure 1) almost performs as well as the model where both the encoder and decoder (Figure 6) is fine-tuned.

Normally a model would be pretrained on a much larger dataset than the fine-tuning dataset, which is why we probably could have gotten better results if we had used such a dataset.

References

[1] Shin'ya Yamaguchi et al. *Image Enhanced Rotation Prediction for Self-Supervised Learning*. June 4, 2021. DOI: 10.48550/arXiv.1912.11603. arXiv: 1912.11603[cs,stat]. URL: http://arxiv.org/abs/1912.11603 (visited on 11/09/2023).

[2] Francesco Saverio Zuppichini. *Implementing Vi(sual)T(transformer) in PyTorch*. original-date: 2021-01-01T15:14:05Z. Nov. 9, 2023. URL: https://github.com/FrancescoSaverioZuppichini/ViT (visited on 11/09/2023).