Coin Classification Report

Tien Thang Dinh
Touch Sensing and Processing (WiSe24/25)
TU Dresden

January 6, 2025

Introduction

This report presents the results of a coin classification task using the given tactile images dataset. The goal of this project was to classify coins into six categories using a pre-trained Vision Transformer (ViT) from Hugging Face as a feature extractor and a custom classifier head. The dataset was split into training (80%), validation (10%), and test (10%) subsets.

Methodology

The model used in this project was the Vision Transformer (ViT), pre-trained on ImageNet ("google/vit-base-patch16-224"), link to model . The feature extraction layers were frozen, and only the classifier head was trained. The following steps were performed:

- **Dataset Preparation:** Scikit-learn was used for splitting the dataset with the ratio 80/10/10. After that, PyTorch's `ImageFolder` and `DataLoader` were used to prepare the dataset, where all images were resized to 224 × 224, as required by the ViT model. No augmentation was used.

- **Define the Model:** The Vision Transformer (ViT) was adapted with a custom classifier head as shown in the following Python snippet:

```python
model_name = "google/vit-base-patch16-224"
model = ViTForImageClassification.from_pretrained(model_name)
model.classifier = nn.Linear(model.config.hidden_size, len(train_loader.dataset.classes))
```

- **Training Loop Setup:** The model was trained for 10 epochs using a learning rate of 5×10^{-5} and a batch size of 32. At each training iteration, metrics such as training loss, training accuracy, validation loss, and validation accuracy were calculated and saved for the report.
- **Test Model on Test Dataset:** After the training finished, a classification test on the test dataset was run, producing a confusion matrix.

Results

Figure 1 shows the training and validation loss over 10 epochs. It is evident that the loss decreases steadily during training, indicating that the model is learning effectively. Interestingly, the validation loss (orange) is slightly lower than the training loss (blue) throughout the training. This might indicate that the validation set is simpler than the training set.

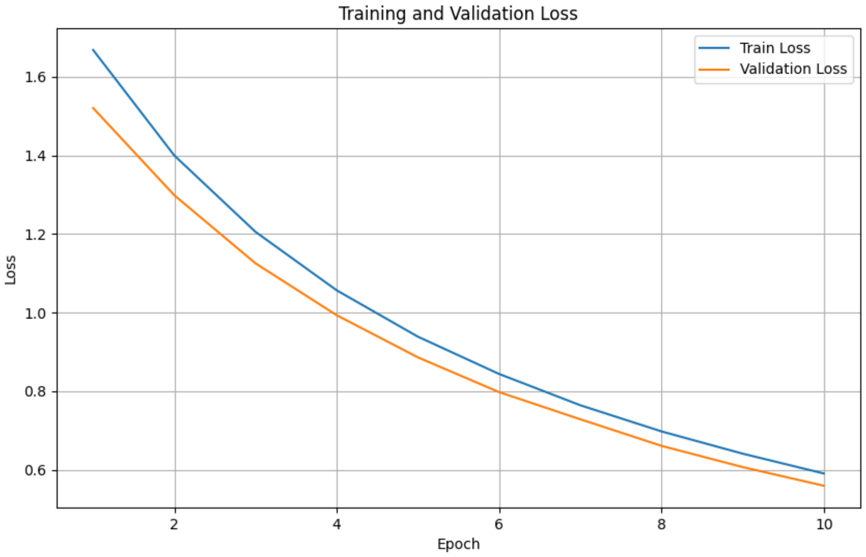

Figure 1: Training and Validation Loss

Figure 2 illustrates the training and validation accuracy, which improves significantly with each epoch. The final validation accuracy reached 94.05%, demonstrating strong model performance. Similar to the loss trends, the validation accuracy (orange) is higher than the training accuracy (blue). This may really due to my assumption above that the validation set is actually simpler than training set. However the overal result indicates that the model is learning effectively.

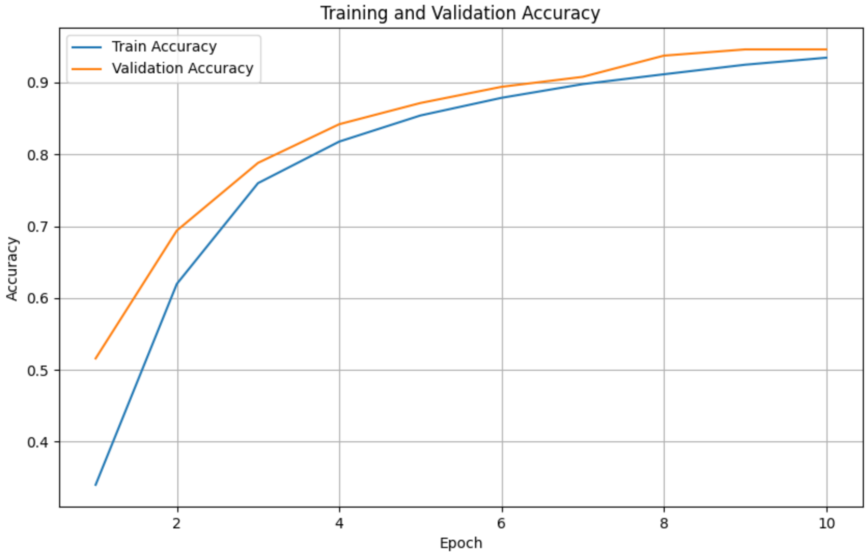

Figure 2: Training and Validation Accuracy

The confusion matrix for the test set is presented in Figure 3. It highlights the classification performance across all coin categories. The following things can be observed:

- **High Accuracy:** Most classes have high diagonal values, indicating strong classification performance.
- **Common missclassification:**
  - The most confusion between `20cent` and `50cent`.
  - This could be due to overlapping features between these classes. For example both `20cent` and `50cent` have `0` digit, as well as `2` and `5` both have a round path in its features.
- **Best Performing Class:** The `1cent` class has the highest number of correctly classified examples (109) and is far better than the second best classes (87).
- **Challenging Classes:** The `20cent` performs the worst for the reason mention aboved.

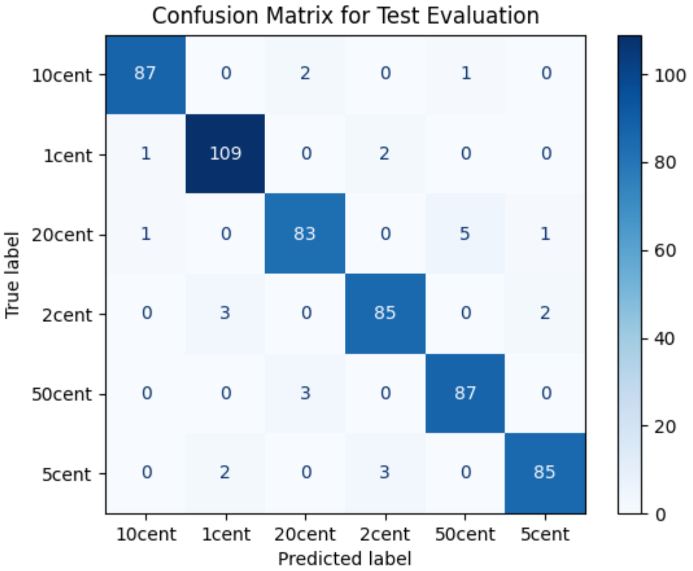

Figure 3: Confusion Matrix for Test Set

Conclusion

The Vision Transformer (ViT) model, fine-tuned with a custom classifier head, demonstrated effective coin classification with an accuracy of 95.37% on the test set. The learning curves indicate steady improvement throughout training, and the confusion matrix highlights strong performance across all categories.

This result suggests that the model is highly effective for tactile-based coin classification. Compared to a human using a single finger for similar tasks, this model not only achieves superior accuracy but also provides rapid inference that we as human might not be able to achieve. These findings open up exciting opportunities for practical applications in areas requiring accurate tactile-based classification where vision is not available.