CS-233 Machine learning project : Milestone 2

Darius Foodeei
darius.foodeei@epfl.ch
EPFL

Theo Le Fur
theo.lefur@epfl.ch
EPFL

Rayane Charif Chefchaouni
rayane.charifchefchaouni@epfl.ch
EPFL

June 2, 2024

Abstract

This report is a summary of our machine learning project for CS-233 milestone 2. We implement and compare various deep learning methods on the Zalando Fashion-MNIST dataset for classification.

1 Introduction

In this report, we present and analyze the performance of four deep learning models on a subset of the Fashion-MNIST dataset, which contains images of Zalando's article types. We benchmark the models using a Multi-Layer Perceptron (MLP), an AlexNet and a ResNet-50, and a Transformer implemented with PyTorch. Additionally, we apply Principal Component Analysis (PCA) to the MLP model to enhance its performance. The goal is to recognize each article type in the images accurately. We discuss our results and provide insights for future improvements.

2 Methodology

Multiple neural network architectures were implemented to benchmark their classification capabilities on Fashion MNIST.

The MLP was configured with three hidden layers, each containing 512 neurons, to ensure sufficient capacity for learning complex patterns in the data. Principal Component Analysis (PCA) was applied to the input data for the MLP to reduce dimensionality and improve training efficiency.

We have implemented two state of the art convolutional neural network architectures. Firstly, we train an AlexNet, that uses max-pooling and dropout regularization to prevent overfitting. Secondly, we train a ResNet50, involving 50 neural network layers along with batch normalization, to attempt to capture complex, hierarchical relationships in the data. Such a deep network is trained efficiently using skip connections, that make the gradient flow more effective. It leverages residual connections, that prevent the gradients from vanishing, globally improving the training procedure.

The Vision Transformer (ViT) model was implemented to classify images by dividing each image into 49 non-overlapping patches, embedding these patches using a linear layer. Sinusoidal positional encodings are added to retain spatial information. The model comprises several transformer blocks, each featuring multi-head self-attention, layer normalization, and a feed-forward network with GELU activation. Configured with specific parameters such as embedding dimension, number of blocks, and attention heads, the model was trained using cross-entropy loss and the Adam optimizer.

The dataset was pre-processed by normalizing the images and creating a validation set from the training data to evaluate model performance after training. Performance was evaluated on the test set using accuracy as the primary metric.

3 Experiments and Results

For the classification of the dataset, the models are trained on the normalized training data then evaluated on a validation set consisting of 10k images out of the 60k total (50k used for training).

We utilize the Adam optimizer for every model. We train the MLP on a batch size of 32 with and without PCA. We chose to take 530 components (cf. **Appendix.A1**) We found that both the AlexNet and the ResNet50 achieve higher accuracy on a smaller batch size of 64. Finally, we found that for the ViT, it is better to use a larger batch size of 256. We also use an exponential learning rate scheduler with weight decay of .99, which has been found to help the model train when the loss stagnates.

	MLP	**MLP with PCA**	**Res-Net**	**Alex-Net**	**Transformer**
Hyper-parameters	**lr** : 3e-4 **epochs** : 50	**lr** : 3e-4 **epochs** : 50	**lr** : 3e-4 **epochs** : 20	**lr** : 3e-4 **epochs** : 40	**lr** : 3e-4 **epochs** : 20
Training set Accuracy	98.8%	99.5%	98.9%	95.4%	93.377%
Training set F1 score	0.988	0.995	0.989	0.954	0.933838
Validation set Accuracy	85.0%	84.6%	87.5%	**90.7%**	85.033%
Validation set F1 score	0.851	0.847	0.875	**0.907**	0.850649

Table 1: Experiment results in Fashion-MNIST classification

4 Discussion and Conclusion

Firstly, our results for the MLP indicate significant overfitting, with high accuracy (99...%) on the training set but only around 85% accuracy on the validation set. Using PCA did not improve the accuracy nor the F1 score, suggesting that dimensionality reduction was not beneficial in this context. We expect an MLP to overfit since it treats an image as a one dimensional input, not leveraging the spacial context nor the hierarchy inside the image.

As expected, both CNNs achieve better scores than the other models. The reason is the intrinsic inductive bias of convolutional layers, given by translation invariance, local receptive fields, hierarchical features and weight sharing. The dropout and max-pooling layers were found to be key to the performance of the AlexNet, allowing it to generalize better than the other models. We notice that the AlexNet outperforms the ResNet in terms of accuracy and F1 score, despite the ResNet having a larger number of layers. This discrepancy is explained by the low numbers of epochs used for training, due to limited resources, as well as the small dataset size, on which the ResNet architecture tends to perform worse.

Finally the ViT experiment can be analyzed from a couple different perspectives. First of all the dataset size consisted of a relatively small amount of images compared to the datasets typically used to train transformers such as ImageNet. Transformers, including ViTs, generally excel in scenarios where they can learn from large amount of complex, high dimensional data. The limited resolution and simpler features of Fashion MNIST might not be sufficient to leverage the capabilities of ViTs effectively which translates to lower accuracy averaging around 85% consistently.

The conclusion we came to for this study is that higher parameter count for both the ViT (9,492,490 params) and CNN (23,548,746 params for ResNet vs 11,696,202 params for AlexNet) did not necessarily correlate to drastically better accuracy which can be an overkill for simple datasets where local features and small-scale patterns are more pronounced and sufficient for good performance.