Neural Network Classification on Fashion-MNIST

Introduction and Problem Context

This project examines the use of Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) for image classification of grayscale clothing images taken from the Fashion-MNIST dataset. The dataset contains 70,000 images (60,000 training, 10,000 testing), in 10 categories, including: T-shirt; Trouser; Pullover; Dress; Coat; Sandal; Shirt; Sneaker; Bag; and Ankle Boot. The scope was to develop, train, and evaluate both architectures in PyTorch, compare performance and consider how misclassification occurred between categories that appear visually distinct (e.g., Shirt, Coat) or similar in terms of shape (e.g., Pullover, Dress). The goal of the study is to show how a deep learning model is operationalized for an image classification problem, and how a performance analysis is undertaken thereafter.

Methodology Overview

The Fashion-MNIST dataset was loaded via torchvision.datasets, and data were normalized to [0, 1]. Data were separated into training (80%), validation (10%) and evaluation or testing (10%) sets.
Two models were then created:

- ANN (Multilayer Perceptron) is made of three fully connected layers, each containing: (1) input layer (784 neurons), (2) hidden layer (256 neurons) and (3) output layer (10 neurons); 128 (256 -> 10) neurons were included for added representation in the output layer. ReLU was used for activations, and Dropout (0.2) was used for a regularization method.
- CNN is made of two convolutional layers, each followed by ReLU activations and with MaxPooling layers after respective contractions followed by (2) fully connected layers (128 and 10 neurons). Dropout (0.25) was used to reduce the overfitting.
- Both models were trained in for 10 epochs using Adam optimizer (learning rate = 0.001) coupled with cross-entropy loss, thereby converging with stability. Loss and accuracy were captured at every epoch.

Key Findings and Results

The table below illustrates , the CNN showed improved accuracy across the board, when compared two models, the CNN was able to extract spatial features more efficiently and generalize more robustly then ANN.

Model	Training Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)
ANN	**89.44**	**88.33**	**87.46**
CNN	**94.85**	**91.85**	**91.14**

Table 1. Performance comparison between ANN and CNN models on the Fashion-MNIST dataset

- The ANN seemed to have more confusion between the classes that are visually similar categories as Sneakers, Bags, and similar objects present a blend of categories and represent confusion.
- The CNN captures greater separation across categories and far fewer misclassifications involving visually similar classes (e.g. Shirt vs Coat vs Pullover).
- The CNN gave higher scores of precision, recall, and F1 consistently for most classes (for instance, sneakers and bag).
- In brief, the CNN learned more efficiently and identified visual features more effectively than the ANN, making it better suited for handling spatial image information.

Detailed Per-Class Metric Breakdown (Precision, Recall, F1-Score)

A detailed breakdown of per-class metrics is key to understanding what is a weakness for each model.

- T-Shirt/Top, Pullover, Shirt, Coat: Both models had the lowest scores (Precision, Recall, and F1-score) in these four classes because they had high amounts of confusion for articles that had similar shapes and silhouettes.
- Trousers, Bags, Ankle Boots, Sandals: Both models had high scores (generally >95%) in these classes because they do not have similar shapes and silhouettes to the rest of the clothing articles.
- CNN Strength: The CNN provided higher scores of precision, recall, and F1-score for most of the classes (i.e., the F1-score for Shirt was substantially better on the CNN). This reinforces that CNN had not only fewer mistakes across all and but had a more uniform performance across types of errors. (False positives vs false negatives)

Innovations and Techniques

- The Adam optimizer was applied to speed up the process of learning by adaptively taking into account the learning rate.
- The training dynamics were represented through loss & accuracy curves.
- The confusion matrices were shown for the purpose of class-wise errors analysis.
- The threshold for precision, recall, and F1-score per class reporting was set so as to provide additional insights along with it.

Conclusions

The project successfully demonstrated the better performance of the Convolutional Neural Network architecture compared to a fully connected Artificial Neural Network for the Fashion-MNIST image classification task.

Most of the confusion in the ANN came from classes which were visually similar or had large feature overlap, e.g., Shirt vs Pullover. Its flattening operation discards useful spatial hierarchy.

The CNN learned more efficiently and identified visual features more effectively than the ANN. Its use of convolution and pooling layers allowed it to extract robust, translation-invariant features (like seams, cuffs, or collars), making it better suited for handling spatial image information. This increased depth and complexity of the CNN architecture were justified by a substantial improvement in accuracy and generalization performance.

Future Works

Additional research to improve performance and push the boundaries of the models might be conducted on:

- **Hyperparameter Search:** Systematically evaluating different learning rates, batch sizes and optimiser choices, for example, RMSprop and SGD with momentum.
- **Architectural Depth:** Consider evaluating a deeper CNN, where additional Conv layers are incorporated; it may be that the model can learn at a higher level of abstraction, which leads to improved performance. As a diversion, one could try more advanced blocks, to include things like Inception or Residual connections.
- **Data augmentation:** Randomly rotating, shifting or zooming the training images can be a means of artificially increasing the size and variation of the training set, while potentially also improving generalization and robustness.