A Clear and Practical Hybrid Deep Learning Model for Fashion Image Classification

Executive Overview

This research aims to investigate whether ensemble learning can enhance image classification accuracy in a deep learning scenario. The study first compares the performance of an Artificial Neural Network (ANN) and a Convolutional Neural Network (CNN), applied to the Fashion-MNIST dataset, and then models an ensemble of both approaches to capitalize on their unique strengths. The ensemble model shows marked improvements in class balance, accuracy and generalisation performance. Overall, the results have both theoretical contributions and practical implications of model-level aggregation by demonstrating how two of the simplest artificial neural network architectures can provide more reliable results for computer vision tasks when capacity is optimally organized in an ensemble model.

1. Introduction

This research examines the development and comparative evaluation of ANN and CNN models for the classification of images from the Fashion-MNIST dataset, which consists of 70,000 greyscale images across ten classes of clothing. The primary aim was to examine whether the predictive power gained of combining ANN and CNN with an ensemble method would be more powerful than either model on its own. This research represents a contribution to the research area of utilizing deep learning for image classification; an area motivated generally by the desire for better robustness, generalisability, and model reliability.

Justification for Model Choices

- The application of ANNs and CNNs was seen to be appropriate because they represent two major learning paradigms in deep learning:
- ANNs learn dense representations of features with fully connected layers.
- CNNs learn local and hierarchical visual patterns suitable for image data.
- The side-by-side evaluation of both models, combined, also offered a useful opportunity to evaluate whether a combination of these learning approaches could offer stronger and more generalisable predictive quality.

2. Methodology and Ensemble Approach

A meticulous research approach was conducted. The Fashion-MNIST data was normalised and underwent one-hot encoding, followed by an 80-10-10 split for the training, validation, and test sets. Controlled data augmentation was used in the first segment of training, but frozen after training for five epochs in order to retain the quality of the image features.

The ANN architecture consisted of fully connected layers using ReLU activation, Batch Normalisation, Dropout, and the AdamW optimiser with L2 regularisation aiding in the generalisation. The CNN used depthwise separable convolutions which were an efficient way to extract spatial and textural features with pooling and regularisation layers. Both models initiated the training stability with EarlyStopping and the LC3 learning-rate schedule (warm-up phase, cosine decay, fine-tuning).

The ensemble model adopted a parallel learning architecture, whereby both models were trained separately and independently on the same dataset. The output probabilities of both models where fused through soft-voting (average fusion). This strategy allowed for low computational performance while improving model confidence, lessening the class biases, and increasing the consistency of the prediction.

3. Performance Summary and Evaluation

The CNN surpassed the ANN in accuracy, with performances of 84.17% and 83.76% respectively, which further asserts the appropriateness of CNN's for visual pattern based learning. The Ensemble model achieved a greater accuracy of 85.66% which indicates a greater improvement over both models. Macro-performance measures also indicated improvement and confirmed a greater class fairness and balanced learning for all ten classes.

Model Performance Overview:

- Accuracy Scores: ANN (83.76%), CNN (84.17%), Ensemble (85.66%)
  Macro Metrics (Ensemble): Precision = 0.858, Recall = 0.857, F1-Score = 0.853
- The ensemble showed enhanced class balance and reduced misclassification inconsistencies across categories.

Model	Accuracy (%)	Macro Precision	Macro Recall	Macro F1
ANN	83.76	0.839	0.838	0.833
CNN	84.17	0.845	0.842	0.840
Ensemble (ANN + CNN)	85.66	0.858	0.857	0.853

The assessment indicated that CNNs provided a better visual feature extraction than ANNs, but this ensemble leveraged its complementary learning strengths to achieve better overall performance. Data augmentation, regularisation, and LC3 scheduling were all influential in reducing overfitting and stabilising model performance. The annotation errors in the test set were largely restricted to visually similar categories (e.g., T-shirt, Shirt, Coat), which indicates that greater feature separation would be needed in some cases.

4. Conclusion

This research provides strong evidence that ensemble learning can boost accuracy, consistency, and balance for individual ANN or CNN models. The research asserts performance advantages of architectural synergy to advocate for increased model generalisation to image classification in deep learning. This research supports the efficacy of hybrid learning strategies and strengthens the case for ensemble learning approaches in large-scale computer vision studies. In summary, these adjustments may encourage model robustness and fairness when deployed in the real world.