# Executive Summary: Neural Networks Classification of Fashion-MNIST

It systematically addressed image classification on the Fashion-MNIST dataset through rigorous design, implementation, and critical comparison of two radically different network architectures: one was an ANN, and the other was CNN. The investigation was executed at a high level of experimental rigor, resulting in the selection of a demonstrably superior model and clear, data-driven strategies for its further performance improvement.

## Methodology and Experimental Rigor

To enable controlled, fair comparison, the entire dataset of 70,000 images was prepared. An important step in prepossessing was normalization, scaling raw pixel values from a range of (0-255) down to a stable (0.0) – (1.0) floating point range. This step is important for training stability and speeding up convergence of the optimizer. Lastly, the data was split into separate training, validation, and test sets in fractions of 50,000 / 10,000 / 10,000 images.

Both models were trained with the same settings: using the adaptive Adam optimizer, considered efficient for maintaining individual learning rates, was combined with the standard Categorical Cross-Entropy loss function in multi-class tasks. Most importantly, Early Stopping (patience=5) represents a powerful regularization method, which avoids overfitting by stopping the training process if it detects no further reduction in the validation loss, selecting the best-generalized weights for the final evaluation.

The results for the performance metrics are also subject to intentional hyperparameter tuning and selection. Although the adaptive Adam optimizer was selected as a starting optimization for performance, the initial work conducted intended to change the learning rate and batch size. The fixed learning rate of 0.001 was selected after seeing the instability or overshooting of convergence at higher rates, while convergence was excessively slow at lower rates. The empirical batch size of 64 was selected to optimize GPU memory usage with balanced desire to improve stability of the update of the gradient, which makes both the training process reasonable and effective.

## Key Results and Results Comparison

When evaluated on the previously held-out test set of 10,000 samples, there was a definitive result that supported the architectural choice regarding image data. The CNN achieved nearly a 4% greater accuracy, signifying that it was better suited to visual data tasks than ANNs.

| Model Architecture | Test Accuracy | Train Accuracy | Validation Accuracy |
|---|---|---|---|
| **ANN** | 87.74% | 89.66% | 89.34% |
| **CNN** | 91.59% | 93.47% | 92.25% |

The root behind such superior performance of CNN lies in its structural efficiency. The convergence of its network was faster, at around Epoch 7, while ANN achieved its best performance after Epoch 9, thus confirming that a CNN learns salient features more effectively with fewer iterations.

Moreover, the model's overall discriminative capacity was assessed using the Area Under the Curve (AUC) measure from a one-vs-all ROC, with the CNN achieving a Micro-average AUC of 0.9967 and a Macro-average AUC of 0.9948. These near perfect values confirm the CNN's remarkable strength and capability to very confidently discriminate between all ten classes,

validating the efficacy of the convolutional feature extraction mechanism across the entire dataset.

**Architectural Insights and Critical Analysis**

The large discrepancy in performance emerges from crucial differences in the way each model represents spatial information. The ANN architecture begins with a mandatory Flatten layer that is essentially disruptive, treating the image as 784 independent inputs, and compelling the network to ineffectively reconstruct spatial relationships from a non-local vector.

In contrast, CNN is engineered based on the principle of locality. It depends on convolutional layers and max-pooling to implement two essential properties: translational invariance and parameter sharing. This structure then allows the CNN to build a hierarchical understanding of the image first detecting simple edges, then combining them into complex forms like sleeves or necklines and recognize such features regardless of their exact position within the frame of the image.

Critical analysis from the confusion matrix showed that the residual error from the CNN of 8.41% is highly concentrated within those classes that are visually ambiguous. The highest confusion rates were between Pullover vs. Coat and between Shirt vs. T-shirt/Top, where salient subtle features that distinguish them, such as collars or lapels, are largely lost or blurred in the low 28 x 28 resolution.

**Conclusion and Innovation**

The CNN model provides an extremely robust foundation for classification. To achieve the accuracy benchmark of more than 92% and to resolve the residual ambiguities within the classes, two highly effective and innovative strategies are recommended for future development.

**Data Augmentation:** Upon reviewing carefully, the residual error generated from the CNN, it was found that the misclassification errors were primarily confined to visually indeterminate classes, such as Shirt/T-shirt and Pullover/Coat. This is expected due to a low 28 x 28 image resolution, which restricts the feature space to identify subtle features of the garment like collars or button lines. Hence, the most important area of improvement is not to change the architecture, but to implement Data Augmentation. Through the synthetic enlargement of the training set through minor transformations, the model learns to classify the ambiguous garments positioned at different angles and locations to directly address the positional quality which was resulting in the highest error.

**Batch Normalization:** Batch Normalization layers would stabilize the distribution of layer activations during training if integrated right after convolutional blocks. This crucial stabilization allows for the use of higher learning rates and speeds up the training process to clearly open avenues for marginal performance gains.

**Transformer:** A distinct future direction for the work would be to completely shift the architecture by exploring a Vision Transformer model. In contrast to the CNN which has some inductive biases (i.e., locality and translation invariant), the ViT breaks an image down into sequence of patches and learns "global" relationships between the patches using a self-attention module. This method changes the prior size limitations of a convolutional kernel and can learn much longer range dependencies across the entire garment (i.e., the relationship between the sleeve of a coat and its collar). While potentially more expensive in terms of computation, this method could eliminate ambiguity in visually similar classes by developing a more nuanced, high-level context.