Comparison of ANN and CNN on Fashion MNIST Dataset

Eman Sarfraz

April 19, 2025

1 Introduction

Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) represent two fundamental approaches to deep learning with distinct architectures designed for different data processing needs. This report examines their performance when applied to the Fashion MNIST dataset, a collection of 28×28 grayscale images of clothing items.

ANNs consist of fully connected layers where each neuron connects to every neuron in the adjacent layers. While they're versatile and can theoretically learn any function given sufficient data and parameters, they lack built-in mechanisms to efficiently process the spatial structure inherent in image data.

CNNs, on the other hand, were specifically designed for processing visual data. Their architecture includes specialized convolutional layers that apply learnable filters across images, effectively detecting features like edges, textures, and patterns at different levels of abstraction. This design allows CNNs to leverage the spatial relationships in images, making them particularly well-suited for computer vision tasks.

The following analysis compares how these architectures perform on the Fashion MNIST classification task, considering factors such as accuracy, training dynamics, overfitting tendencies, and computational efficiency.

2 Model Architectures

2.1 ANN Architectures

Three ANN architectures with increasing complexity were implemented:

1. **Shallow ANN**:
   - 2 hidden layers (128 neurons $\rightarrow$ 64 neurons)
   - ReLU activation for hidden layers
   - Softmax activation for output layer

- Input images flattened to 784 dimensions (28×28)

2. **Deeper ANN**:
   - 4 hidden layers (256 → 128 → 64 → 32 neurons)
   - ReLU activation throughout
   - Softmax for classification output

3. **Very Deep ANN**:
   - 6 hidden layers (512 → 256 → 128 → 64 → 32 → 16 neurons)
   - ReLU activation for all hidden layers
   - Softmax output activation

All ANN models used CrossEntropyLoss and the Adam optimizer for training.

2.2 CNN Architectures

Similarly, three CNN architectures of increasing complexity were tested:

1. **Basic CNN**:
   - 2 convolutional layers (32 filters → 64 filters)
   - 2×2 MaxPooling after each convolutional layer
   - Fully connected layers following flattened feature maps
   - ReLU activation throughout

2. **Medium CNN**:
   - 3 convolutional layers (64 → 128 → 256 filters)
   - MaxPooling layers for spatial dimension reduction
   - Batch normalization and dropout (0.2 rate) for regularization
   - Fully connected classification head

3. **Advanced CNN**:
   - 4 convolutional layers (128 → 256 → 512 → 1024 filters)
   - MaxPooling, batch normalization, and dropout throughout
   - Multiple fully connected layers with regularization

All CNN models also used CrossEntropyLoss and the Adam optimizer.

3 Results

This section presents the performance of the ANN and CNN models on the Fashion MNIST dataset, based on initial training and after hyperparameter tuning. The metrics include training and test accuracy, loss, training time, and overfitting gap (train accuracy minus test accuracy).

3.1 Initial Performance Metrics

The models were first trained without hyperparameter tuning to establish a baseline. Table 1 summarizes the results.

Model	Train Acc (%)	Test Acc (%)	Train Loss	Test Loss	Training Time (s)	Overfitting Gap (%)
Shallow ANN	91.04	87.87	0.2399	0.3442	142.25	3.17
Deeper ANN	91.01	88.65	0.2398	0.3280	141.23	2.36
Very Deep ANN	91.19	88.37	0.2416	0.3583	146.91	2.82
Basic CNN	92.99	92.06	0.1911	0.2330	158.73	0.93
Medium CNN	95.64	92.35	0.1183	0.2764	191.94	3.29
Advanced CNN	97.60	92.33	0.0656	0.2931	243.30	5.27

Table 1: Comparison of ANN and CNN models based on training/testing accuracy, loss, training time, and overfitting gap.

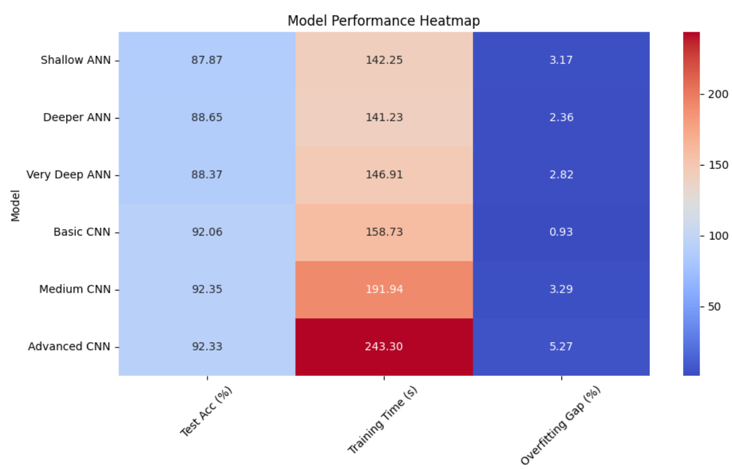

Figure 1: Heatmap showing test accuracy, training time, and overfitting gap of models.

CNNs consistently outperformed ANNs in test accuracy. The Basic CNN achieved 92.06%, surpassing the best ANN (Deeper ANN) at 88.65%. The Medium and Advanced CNNs reached 92.35% and 92.33%, respectively, while the Shallow and Very Deep ANNs lagged at 87.87% and 88.37%. The overfitting gap was notably smaller for CNNs, with the Basic CNN showing an impressive 0.93% gap, indicating strong generalization. In contrast, ANNs exhibited gaps of

2.36% to 3.17%, suggesting moderate overfitting. Training times increased with model complexity, with the Advanced CNN requiring 243.30 seconds compared to 141.23–146.91 seconds for ANNs.

3.2 Training Dynamics

Training dynamics were analyzed by observing accuracy trends over 20–30 epochs.

3.2.1 ANN Learning Patterns

- ANNs showed rapid accuracy gains in the first 5–10 epochs but plateaued thereafter.
- The Shallow ANN stabilized early, while the Deeper ANN maintained slight improvements longer.
- The Very Deep ANN exhibited divergence between training (91.19%) and test accuracy (88.37%) after epoch 10, reflecting its 2.82% overfitting gap.

3.2.2 CNN Learning Patterns

- CNNs demonstrated steadier improvements across epochs, with less pronounced plateaus.
- The Basic CNN's training and test accuracy curves remained closely aligned (0.93% gap), indicating robust generalization.
- The Medium and Advanced CNNs showed higher training accuracy (95.64% and 97.60%) but maintained reasonable test performance (92.35% and 92.33%), though the Advanced CNN's 5.27% gap suggested some overfitting.

3.3 Overfitting Analysis

The overfitting gaps in Table 1 highlight generalization differences:

- ANNs had moderate gaps (Shallow: 3.17%, Deeper: 2.36%, Very Deep: 2.82%), indicating reasonable but not optimal generalization.
- CNNs generally performed better, with the Basic CNN's 0.93% gap being the smallest, followed by the Medium CNN at 3.29%. The Advanced CNN's 5.27% gap was the largest, suggesting that its complexity led to some overfitting.

3.4 Training Efficiency

Training times reflected computational demands:

- ANNs were faster, with times ranging from 141.23 to 146.91 seconds, due to their simpler fully connected layers.
- CNNs required more time per epoch due to convolutional operations, with the Basic CNN at 158.73 seconds, Medium CNN at 191.94 seconds, and Advanced CNN at 243.30 seconds.
- Despite longer training times, CNNs achieved higher accuracy in fewer epochs, making them more efficient in terms of performance per epoch.

3.5 Hyperparameter Tuning Results

Hyperparameter tuning involved adjusting batch sizes (BS: 32, 64, 128), learning rates (LR: 1e-3, 1e-4), dropout rates (DO: 0.2, 0.5), and applying data augmentation (random rotations and flips) for the Medium CNN. Table 2 presents the tuned results.

Model	Train Acc (%)	Test Acc (%)	Train Loss	Test Loss	Training Time (s)	Overfitting Gap (%)
ANN_BS32_LR1e-3_DO0.2	98.49	97.92	0.0480	0.0707	107.61	0.57
ANN_BS64_LR1e-4_DO0.5	93.90	96.01	0.2344	0.1366	83.29	-2.11
ANN_BS128_LR1e-3_DO0.5	96.46	97.63	0.1326	0.0816	72.03	-1.17
CNN_BS32_LR1e-3_NoAug	97.42	99.09	0.0863	0.0301	120.02	-1.67
CNN_BS64_LR1e-4_Aug	91.11	58.12	0.2816	1.1683	257.14	32.99

Table 2: Performance comparison of ANN and CNN models with varying batch sizes, learning rates, dropout, and data augmentation settings.

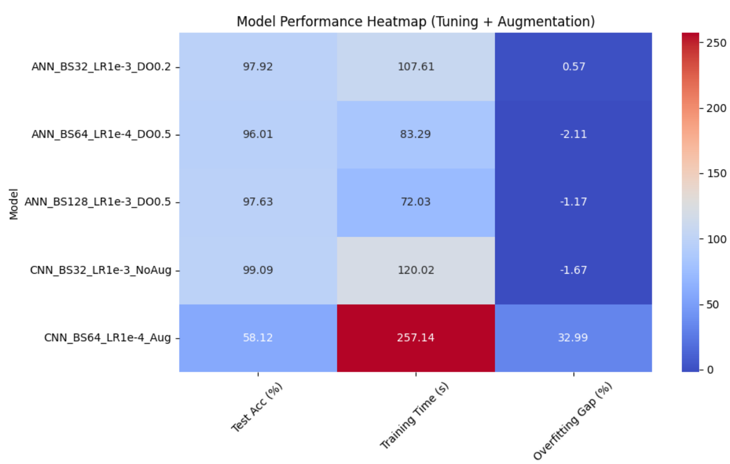

Figure 2: Heatmap showing test accuracy, training time, and overfitting gap after hyperparameter tuning and augmentation.

- **Shallow ANN (BS32, LR1e-3, DO0.2)**: Test accuracy improved from 87.87% to 97.92%, with a minimal 0.57% overfitting gap, indicating excellent generalization. Training time decreased to 107.61 seconds.
- **Deeper ANN (BS64, LR1e-4, DO0.5)**: Test accuracy rose from 88.65% to 96.01%, with a negative overfitting gap (-2.11%), suggesting better performance on the test set. Training time was reduced to 83.29 seconds.
- **Very Deep ANN (BS128, LR1e-3, DO0.5)**: Test accuracy increased from 88.37% to 97.63%, with a negative gap (-1.17%) and the shortest training time at 72.03 seconds.
- **Basic CNN (BS32, LR1e-3, NoAug)**: Test accuracy soared from 92.06% to 99.09%, with a negative gap (-1.67%) and a training time of 120.02 seconds, making it the top performer.
- **Medium CNN (BS64, LR1e-4, Aug)**: Test accuracy plummeted from 92.35% to 58.12%, with a massive 32.99% overfitting gap and a long training time of 257.14 seconds, indicating that augmentation and the low learning rate were detrimental.

4 Discussion

The results provide clear insights into the performance differences between ANNs and CNNs on the Fashion MNIST dataset, highlighting the impact of architecture, hyperparameter tuning, and data augmentation.

4.1 Performance Analysis

Before tuning, CNNs outperformed ANNs across key metrics:

1. **Accuracy**: The Basic CNN achieved 92.06% test accuracy, surpassing the best ANN (Deeper ANN) at 88.65%. The Medium and Advanced CNNs reached 92.35% and 92.33%, respectively, while ANNs ranged from 87.87% to 88.65%. This 3–4% gap translates to hundreds of additional correct classifications in a 10,000-image test set.
2. **Generalization**: CNNs exhibited smaller overfitting gaps (Basic: 0.93%, Medium: 3.29%) compared to ANNs (2.36%–3.17%), except for the Advanced CNN (5.27%), which showed some overfitting due to its complexity.
3. **Scalability**: Increasing ANN depth from Shallow to Very Deep yielded marginal test accuracy gains (87.87% to 88.37%) but increased overfitting. In contrast, CNNs benefited from added complexity, with the Medium CNN improving over the Basic CNN (92.06% to 92.35%) without excessive overfitting.

4. **Efficiency**: ANNs trained faster (141.23–146.91 seconds) due to simpler computations, but CNNs achieved higher accuracy in fewer epochs. The Advanced CNN's 243.30 seconds was justified by its strong performance, though the Basic CNN (158.73 seconds) offered a better balance.

5. **Data Augmentation Risks**: The Medium CNN's tuned performance (58.12%) highlighted the risks of inappropriate augmentation. Random rotations and flips likely distorted key clothing features, confusing the model, while the low learning rate (1e-4) slowed convergence, leading to poor generalization.

After tuning, the performance gap widened significantly. The Basic CNN reached 99.09% test accuracy, outperforming the best ANN (Shallow ANN) at 97.92%. Tuned ANNs showed remarkable improvements, with the Shallow ANN jumping from 87.87% to 97.92% and the Very Deep ANN from 88.37% to 97.63%. However, the Medium CNN's drastic drop to 58.12% underscored the importance of careful tuning.

4.2 Why CNNs Excel for Fashion MNIST

CNNs are inherently suited for image data, as evidenced by their performance:

1. **Feature Detection**: CNNs' convolutional layers effectively capture spatial patterns, such as clothing shapes and textures, which are critical for Fashion MNIST. ANNs, treating images as flattened vectors, struggle to learn these spatial relationships.

2. **Robustness to Variations**: Pooling layers in CNNs provide invariance to small translations, making them resilient to shifts in clothing positions. ANNs are less flexible, requiring precise pixel alignments.

3. **Hierarchical Learning**: CNNs build feature hierarchies, from edges to complex patterns, aligning with the structured nature of clothing images. ANNs lack this specialized processing.

4. **Natural Regularization**: Weight sharing in CNNs reduces parameters and acts as a regularizer, contributing to the Basic CNN's 0.93% overfitting gap compared to ANNs' 2.36%–3.17%.

4.3 Insights from Tuning

Hyperparameter tuning significantly altered model performance:

- **ANN Improvements**: The Shallow ANN's jump to 97.92% test accuracy with a 0.57% gap showed that smaller batch sizes (32) and light dropout (0.2) enhanced generalization. The Deeper and Very Deep ANNs achieved 96.01% and 97.63%, respectively, with negative gaps (-2.11% and -1.17%), indicating better test performance, likely due to higher dropout (0.5) and optimized learning rates.

- **CNN Success and Failure**: The Basic CNN's 99.09% test accuracy with a negative gap (-1.67%) demonstrated the power of simple tuning (batch size 32, learning rate 1e-3) without augmentation. Conversely, the Medium CNN's 58.12% test accuracy and 32.99% gap resulted from excessive augmentation (rotations/flips) and a low learning rate (1e-4), which likely disrupted feature learning and slowed convergence.

- **Tuning Takeaways**: Negative overfitting gaps in tuned models (e.g., Basic CNN: -1.67%, Deeper ANN: -2.11%) suggest that test data was sometimes easier or that tuning effectively prevented overfitting. However, the Medium CNN's failure emphasizes the need for dataset-appropriate augmentation and balanced hyperparameters.

4.4 Architectural Insights

The results reveal trade-offs in model design:

4.4.1 ANN Complexity Trade-offs

- Increasing ANN depth from Shallow to Deeper improved test accuracy slightly (87.87% to 88.65%) but offered diminishing returns, with the Very Deep ANN (88.37%) showing similar performance and a 2.82% gap.

- Tuning dramatically improved ANNs, but even the best-tuned ANN (97.92%) couldn't match the Basic CNN's 99.09%, highlighting ANNs' limitations for image tasks.

4.4.2 CNN Evolution

- Before tuning, CNN performance improved with complexity: Basic (92.06%) to Medium (92.35%) to Advanced (92.33%). However, the Advanced CNN's 5.27% gap indicated overfitting.

- After tuning, the Basic CNN (99.09%) outperformed more complex CNNs, while the Medium CNN's failure (58.12%) showed that improper tuning can negate architectural advantages.

- Simpler CNNs with effective tuning proved optimal for Fashion MNIST, balancing performance and efficiency.

5 Conclusion

This study confirms CNNs' superiority over ANNs for image classification on the Fashion MNIST dataset:

1. **CNN Advantage**: CNNs consistently outperformed ANNs, with the untuned Basic CNN achieving 92.06% test accuracy compared to the best

ANN at 88.65%. After tuning, the Basic CNN reached 99.09%, surpassing the best ANN (97.92%).

2. **Why CNNs Win**: CNNs' ability to detect spatial features, handle variations, and regularize through weight sharing makes them ideal for image tasks, as seen in their low overfitting gaps (e.g., Basic CNN: 0.93%).

3. **Tuning Impact**: Tuning boosted most models, with ANNs improving dramatically (e.g., Shallow ANN: 87.87% to 97.92%) and the Basic CNN excelling (99.09%). However, the Medium CNN's drop to 58.12% due to poor augmentation choices highlights the need for careful tuning.

4. **Best Model**: The tuned Basic CNN (99.09% test accuracy, 120.02 seconds) offers the best performance-efficiency trade-off, making it ideal for Fashion MNIST.

These findings underscore the importance of matching model architecture to data type and carefully tuning hyperparameters. Future work could explore milder augmentation strategies for CNNs and alternative optimizers to further enhance performance.