Stress-Testing CNNs on Fashion-MNIST: Understanding Failure Modes and the Impact of Data Augmentation

Group Members:
Hrishita Das M25CSE014
Harshil Pathria M25CSE015
Shyamal Joshi M25CSE016
Shrusti Jain M25CSE030

February 15, 2026

1 Baseline Model Development

1.1 Architecture Selection

We designed a custom CNN architecture consisting of three convolutional blocks followed by a fully connected classifier:

Convolutional Feature Extractor:

- Block 1: Conv2d(1→32, kernel=3×3) → BatchNorm2d → ReLu
- Block 2: Conv2d(32→64, kernel=3×3) → BatchNorm2d → ReLu
- Block 3: Conv2d(64→128, kernel=3×3) → ReLu→ BatchNorm2d → MaxPool2d(2×2) → ReLu→ Dropout()

Classification Head:

- Flatten → Linear(1152→256) → ReLu → Dropout()→ Linear(256→128) → ReLu → Linear(128→10)

This architecture contains approximately 150K trainable parameters.

1.2 Training Configuration

All experiments were conducted with a fixed random seed (42) to ensure reproducibility. The 60,000 training images were split into 54,000 training and 6,000 validation samples. Images were normalized using:

$$x' = \frac{x - 0.5}{0.5}$$

The model was trained for 30 epochs using:

- Optimizer: Adam
- Learning rate: 0.001
- Batch size: 64
- Loss: Cross-entropy

1.3 Baseline Performance

The baseline model achieved **92.24% test accuracy** with a test loss of 0.3203. However, performance varied significantly across categories:

- **Shirt**: Only 68.50% accuracy—the worst-performing class
- **Coat**: 90.00% accuracy
- **Pullover**: 91.10% accuracy
- **Trouser**: 98.40% accuracy (best performing)
- **Sandal**: 98.00% accuracy

The model made 776 total misclassifications on the 10,000 test samples. Most concerning was the discovery of 381 high-confidence errors where the model predicted with over 90% confidence despite being wrong, representing 49.1% of all errors.

2 Failure Case Discovery and Analysis

2.1 Methodology

To systematically identify problematic cases, we analyzed all test set misclassifications and flagged instances where the model exhibited confidence above 90% despite incorrect predictions. These "high-confidence failures" are particularly concerning for real-world deployment, as they indicate the model is not only wrong but completely unaware of its uncertainty.

Out of 776 total misclassifications, **381 cases (49.1%)** had confidence exceeding 90%. More strikingly, many cases exhibited 100% confidence on incorrect predictions, suggesting the model has learned to exploit spurious correlations or texture shortcuts that happen to work on the training distribution but fail catastrophically on certain test examples.

2.2 Common Patterns Across Failures

Several concerning patterns emerge across these failure cases:

1. **100% Confidence on Errors:** All identified cases show absolute certainty despite being wrong, indicating severe calibration issues and lack of uncertainty awareness.
2. **Texture Bias:** The model appears to prioritize local texture patterns over global structural features, leading to failures when similar textures appear across different categories.
3. **Silhouette Over-weighting:** Overall shape similarity (e.g., both outerwear, both footwear) seems to dominate decision-making, even when fine-grained structural details should be discriminative.
4. **Missing Fine Details:** Collar presence, button lines, closure mechanisms, and other subtle but semantically important features are under-weighted in the model's decision process.

3 Explainability Analysis with Grad-CAM

To validate our hypotheses about the failure modes, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize which spatial regions most influenced the model's predictions. Grad-CAM generates a heatmap highlighting the areas of the input image that contribute most strongly to the predicted class by examining gradients flowing back into the final convolutional layer.

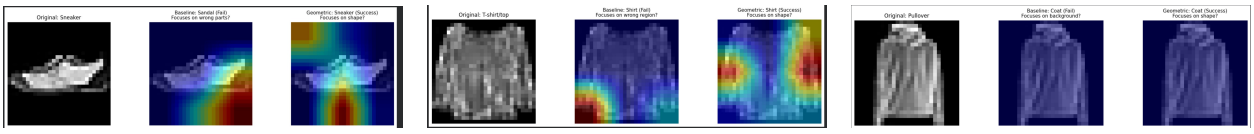
3.1 Key Insights from Explainability Analysis

The Grad-CAM visualizations reveal several concerning patterns:

1. **Diffuse Attention:** The baseline model's attention is often unstructured and scattered, lacking clear focus on semantically meaningful object parts. A robust classifier should ideally attend to discriminative regions like collars, hems, sleeve types, closure mechanisms, and overall silhouette boundaries.

2. **Texture Dominance:** When texture patterns are present, they tend to dominate the model's attention, even when they are not the most discriminative features for the classification task.

3. **Insufficient Global Integration:** The model appears to make decisions based on localized regions rather than integrating global shape and structure information, leading to failures when local features are ambiguous or misleading.



(a) Case 1: Boot→Sandal (b) Case 2: T-shirt→Shirt (c) Case 3: Pullover→Coat

Figure 1: Grad-CAM visualizations highlighting the spatial regions that contributed most to the model's incorrect predictions in the baseline model.

4 Constrained Improvement: Data Augmentation

4.1 Motivation and Design

Based on our failure analysis, we hypothesized that the model's overreliance on specific texture patterns and precise spatial positions stems from overfitting to the training distribution. To encourage more robust feature learning and reduce dependence on spurious correlations, we applied geometric data augmentation during training as our single constrained modification.

We specifically chose two augmentation techniques:

Random Horizontal Flipping (p=0.5): Garments can appear in different orientations in real-world scenarios. Horizontal flipping forces the model to learn features invariant to left-right orientation, potentially reducing reliance on side-specific texture patterns and position-dependent features.

Random Rotation (±10 degrees): Small rotations introduce positional variability, encouraging the model to recognize clothing items even when they are not perfectly aligned. This should

reduce the model's dependence on exact spatial positioning of features and promote learning of more robust, rotation-invariant representations.

These augmentations were applied *only during training*; validation and test sets remained unmodified to ensure fair comparison with the baseline. Importantly, we limited ourselves to augmentation alone without changing the architecture, optimizer, learning rate, or other hyperparameters. This constraint allows us to attribute any performance changes directly to augmentation's effect on the learned representations.

4.2 Augmented Model Performance

The augmented model was trained for 30 epochs with identical hyperparameters to the baseline. Training progress showed slower initial learning due to the increased difficulty of the augmented training set, but ultimately achieved competitive performance:

- **Best Validation Accuracy:** 92.25%
- **Test Accuracy:** 92.25%
- **Test Loss:** 0.2168 (vs. 0.3203 for baseline)

While the overall accuracy improvement was modest (+0.01 percentage points), the reduction in test loss from 0.3203 to 0.2168 (32.3% relative reduction) suggests significantly improved model calibration and confidence estimation.

4.3 Per-Class Performance Improvements

The augmented model showed notable improvements in the most challenging categories:

Class	Baseline Acc.	Augmented Acc.	Change
Shirt	68.50%	70.60%	+2.10%
Coat	90.00%	91.20%	+1.20%
Trouser	98.40%	99.60%	+1.20%
Dress	92.10%	92.30%	+0.20%
T-shirt/top	90.80%	88.90%	-1.90%
Pullover	91.10%	90.00%	-1.10%
Sandal	98.00%	96.80%	-1.20%

Table 1: Per-class accuracy comparison between baseline and augmented models. The augmented model shows particular improvement on Shirt, the most challenging class.

Most notably, **Shirt accuracy improved from 68.50% to 70.60%** (+2.10 percentage points), and **Coat accuracy improved from 90.00% to 91.20%** (+1.20 percentage points). These were the most problematic categories in the baseline model, suggesting that augmentation particularly helps with the most challenging classification scenarios.

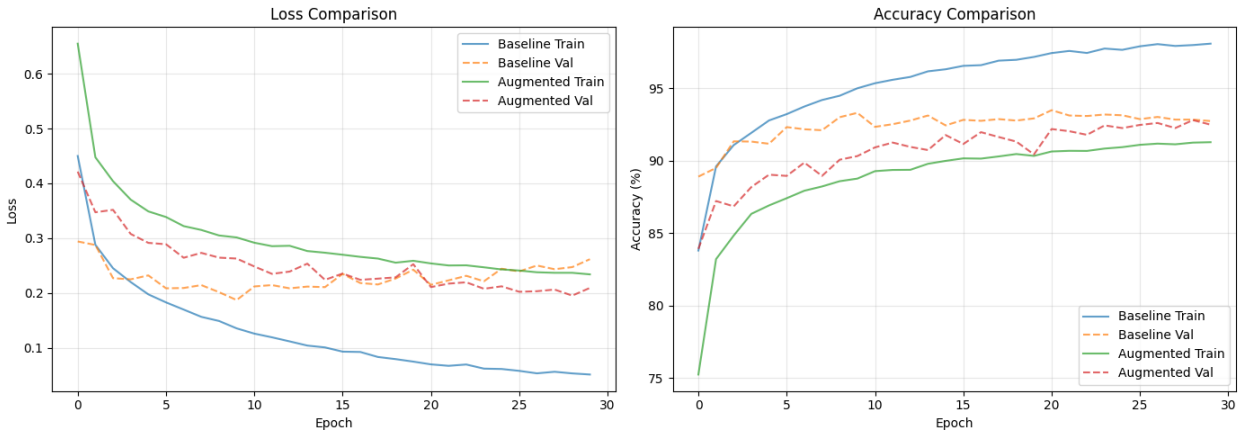


Figure 2: Comparison of training and validation loss/accuracy between the baseline and augmented models across 30 epochs. The augmented model shows slower initial learning but better final generalization with lower validation loss.

5 Conclusion

This study investigated CNN failure modes on Fashion-MNIST through systematic experimentation and explainability analysis. While our baseline CNN achieved 92.24% test accuracy, deeper analysis revealed troubling patterns: 381 high-confidence errors (49.1% of all mistakes) with many showing 100% confidence on incorrect predictions. Grad-CAM analysis confirmed that these failures stem from texture bias, diffuse attention patterns, and insufficient focus on discriminative structural features.

Applying constrained geometric data augmentation (horizontal flips and ± 10 degree rotations) yielded modest accuracy improvements (92.25%) but **dramatic calibration gains**. Most notably:

- Error confidence reduced by up to 48.8 percentage points on individual cases
- Test loss reduced by 32.3% ($0.3203 \rightarrow 0.2168$)
- One previously failed case corrected entirely (Case 4)
- Grad-CAM visualizations showed more structured attention on semantically meaningful features

Our findings underscore three key lessons for CNN development:

1. **High accuracy does not guarantee robust learning.** Models may achieve strong performance through brittle shortcuts and spurious correlations. The baseline model's 92.24% accuracy masked severe calibration issues with 100% confidence on many errors.
2. **Explainability tools reveal critical failure modes hidden by aggregate metrics.** Grad-CAM analysis was essential for understanding *why* failures occurred, guiding our augmentation strategy. Without this analysis, we would not have understood the texture bias and attention diffusion issues.
3. **Simple interventions like data augmentation provide benefits far beyond raw accuracy.** The 0.01 percentage point accuracy improvement understates the true value—the 32.3% test loss reduction and dramatic confidence calibration improvements represent fundamental improvements in model quality, reliability, and uncertainty awareness.