Experiment Summary

This experiment explored the performance of different neural network architectures for image classification using the Fashion MNIST dataset. The goal was to compare a simple Multilayer Perceptron (MLP) with Convolutional Neural Networks (CNNs), including an improved version with data augmentation and batch normalization.

Methodology

1. **Data Loading and Preprocessing:** The Fashion MNIST dataset, consisting of 60,000 training images and 10,000 test images of 28x28 grayscale fashion items, was loaded. The pixel values were normalized to the range [0, 1].

2. **Baseline MLP Model:** A sequential MLP model was built with a flatten layer to process the 28x28 images as vectors (784 features), followed by a dense layer with 128 units and ReLU activation, and an output dense layer with 10 units and SoftMax activation for classification. The model was compiled with the Adam optimizer and sparse categorical crossentropy loss.

3. **Initial CNN Model:** A sequential CNN model was constructed, starting with a Conv2D layer (32 filters, 3x3 kernel, ReLU activation), followed by MaxPooling2D (2x2 pool size). Another Conv2D layer (64 filters) and MaxPooling2D layer were added. The output was then flattened and fed into a dense layer with 128 units and ReLU activation, a dropout layer (0.5), and a final dense layer with 10 units and SoftMax activation. This model was also compiled with Adam optimizer and sparse categorical crossentropy loss.

4. **Improved CNN Model:** This model extended the initial CNN by adding a data augmentation layer at the beginning (random rotation, translation, flip, and zoom) and Batch Normalization layers after the MaxPooling2D layers. The rest of the architecture and compilation were the same as the initial CNN.

5. **Training and Evaluation:** All models were trained for 10 epochs using the training data and evaluated on the test data. Performance was assessed using test accuracy, confusion matrices, and classification reports. Misclassified images were also visualized to gain qualitative insights.

Results

The test accuracies for the three models were as follows:

* Baseline MLP Test Accuracy: 0.8757
* Initial CNN Test Accuracy: 0.9107
* Improved CNN Test Accuracy: 0.8590

Comparing the models, we observed the following accuracy improvements:

* **MLP to Initial CNN:** A significant improvement in test accuracy from approximately 0.876 to 0.908, demonstrating the effectiveness of convolutional layers in capturing spatial features for image classification.
* **Initial CNN to Improved CNN:** A slight decrease in test accuracy from approximately 0.908 to 0.861. While data augmentation and batch normalization are often beneficial, in this specific implementation and with the chosen parameters, they did not lead to improved performance on the test set.

Further analysis through confusion matrices and classification reports revealed that:

* The Baseline MLP struggled significantly with visually similar classes like 'Shirt' and 'T-shirt/top', and 'Coat' and 'Pullover'.
* The Initial CNN showed reduced confusion across most classes and generally higher precision and recall.
* The Improved CNN, despite the additions, still exhibited similar confusions as the initial CNN, particularly within the footwear categories and between 'Shirt' and 'T-shirt/top'.

Insights

The results highlight the power of CNNs over traditional MLPs for image classification tasks like Fashion MNIST. The convolutional and pooling layers effectively extract hierarchical features from the images, leading to better performance.

Interestingly, the improved CNN with data augmentation and batch normalization did not outperform the initial CNN in this experiment. This could be due to several factors:

* **Data Augmentation Parameters:** The chosen augmentation parameters might have been too aggressive or not optimally tuned for this dataset and architecture, potentially introducing variations that made classification more difficult for certain classes (e.g., 'Shirt').

* **Batch Normalization Placement:** While generally helpful for stabilizing training, the placement and interaction of batch normalization with other layers and the data augmentation strategy could require further fine-tuning.

* **Dataset Characteristics:** Fashion MNIST images are relatively low-resolution and have limited variability compared to more complex datasets. The added complexity of the improved model and augmentation might not provide significant benefits or could even be detrimental without careful tuning.