Comparative Analysis: CNN vs ViT on Fashion MNIST Dataset

Performance Evaluation and Model Comparison

Introduction

Fashion MNIST is a dataset comprising **28x28 grayscale** images of clothing items.

10 distinct categories/classes of clothing are included in this dataset.

It serves as a standard benchmark dataset for fashion classification tasks in machine learning

Introduction

Dataset was divided into training and test sets for model evaluation.

Data normalization (scaling pixel values to [0, 1]) was performed for model training.

Split	Examples
'test'	10,000
'train'	60,000

```
(x_train, y_train), (x_test, y_test) = tf.keras.datasets.fashion_mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0
```

Methodology in the Project

In this analysis, we've used TensorFlow and Keras to evaluate the performance of Convolutional Neural Networks (CNN) and Vision Transformers (ViT) on the Fashion MNIST dataset. The dataset, and the methods used for comparison :

Dataset : https://www.tensorflow.org/datasets/catalog/fashion_mnist

Methods :

Methods	Sources
Vision transformers	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (https://arxiv.org/abs/2010.11929)
Convolutional Neural Networks	Deep Learning in Neural Networks https://arxiv.org/pdf/1404.7828.pdf)

CNN Model

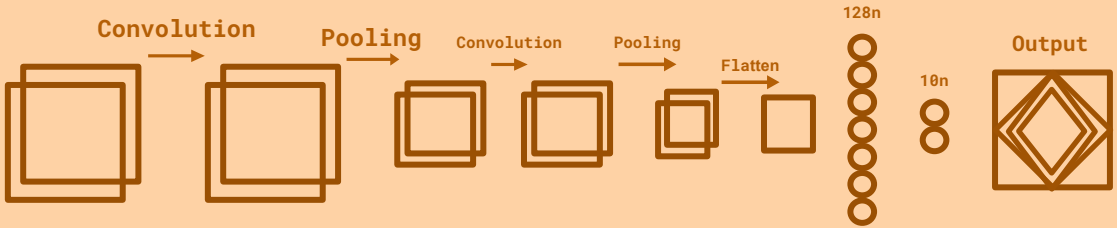
CNN architecture consists of two convolutional layers—initially with 32 filters and subsequently with 64 filters—followed by max pooling layers to downsample the features.

The model employs ReLU activation functions to introduce non-linearity after each convolutional operation.

After flattening the extracted features, the network integrates two dense layers, the first containing 128 neurons with ReLU activation, and the final layer with 10 neurons representing the output classes of Fashion MNIST

CNN Model

```
# CNN Model
cnn_model = models.Sequential([
    layers.Conv2D(32, (3, 3), activation='relu', input_shape=(28, 28, 1)),
    layers.MaxPooling2D((2, 2)),
    layers.Conv2D(64, (3, 3), activation='relu'),
    layers.MaxPooling2D((2, 2)),
    layers.Flatten(),
    layers.Dense(128, activation='relu'),
    layers.Dense(10)
])
```



ViT Model

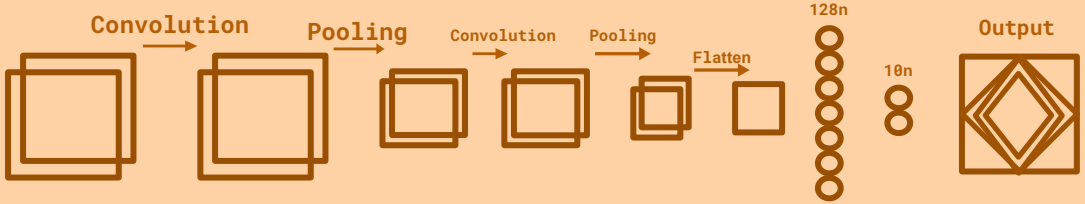
ViT-based model begins with a rescaling layer to normalize pixel values and then employs a single convolutional layer with 32 filters (3x3) and ReLU activation.

Subsequently, max pooling is applied to reduce spatial dimensions, followed by a dropout layer (20% rate) to prevent overfitting.

The network then flattens the output before transitioning to a dense layer comprising 256 neurons with ReLU activation, followed by another dropout layer. Finally, a 10-unit dense layer with softmax activation serves as the output

ViT Model

```
# ViT Model
vit_model = tf.keras.Sequential([
    layers.Rescaling(1./255, input_shape=(28, 28, 1)),
    layers.Conv2D(32, 3, activation='relu'),
    layers.MaxPooling2D(),
    layers.Dropout(0.2),
    layers.Flatten(),
    layers.Dense(256, activation='relu'),
    layers.Dropout(0.2),
    layers.Dense(10, activation='softmax')
])
```

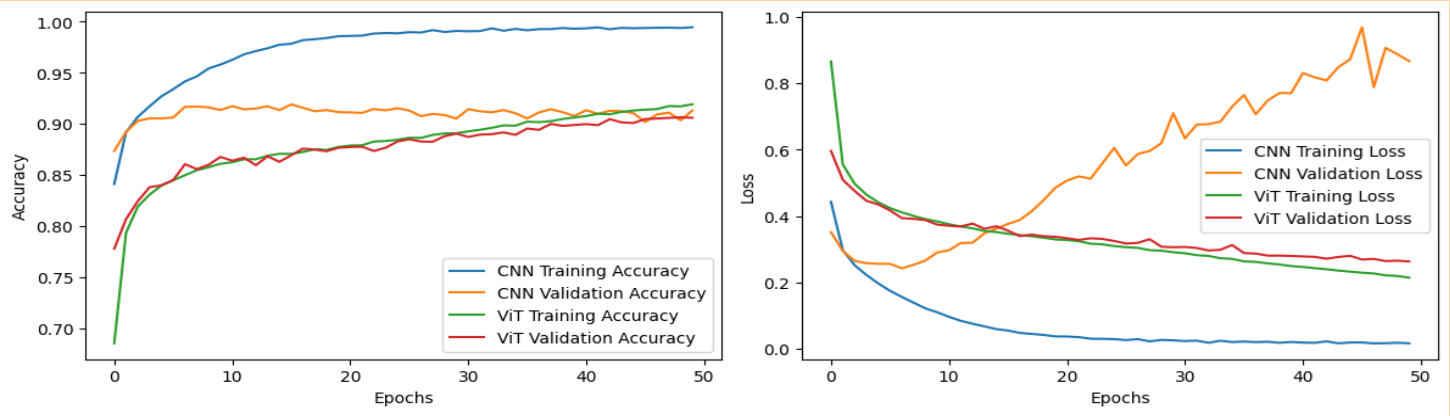


Results and Conclusion

Comparing the performance of Convolutional Neural Networks (CNN) and Vision Transformers (ViT) on the Fashion MNIST dataset :

```
CNN Test Accuracy: 0.9102000207901001
ViT Test Accuracy: 0.906000018119812
```

Both models exhibited remarkable accuracy in classifying Fashion MNIST items. The CNN model slightly outperformed the ViT model by 0.42%.



Results : CNN Model

CNN Classification Report:

	precision	recall	f1-score	support
0	0.84	0.89	0.87	1000
1	0.99	0.98	0.98	1000
2	0.83	0.88	0.86	1000
3	0.93	0.92	0.92	1000
4	0.84	0.87	0.85	1000
5	0.98	0.98	0.98	1000
6	0.80	0.69	0.74	1000
7	0.96	0.97	0.97	1000
8	0.98	0.98	0.98	1000
9	0.98	0.96	0.97	1000
accuracy			0.91	10000
macro avg	0.91	0.91	0.91	10000
weighted avg	0.91	0.91	0.91	10000

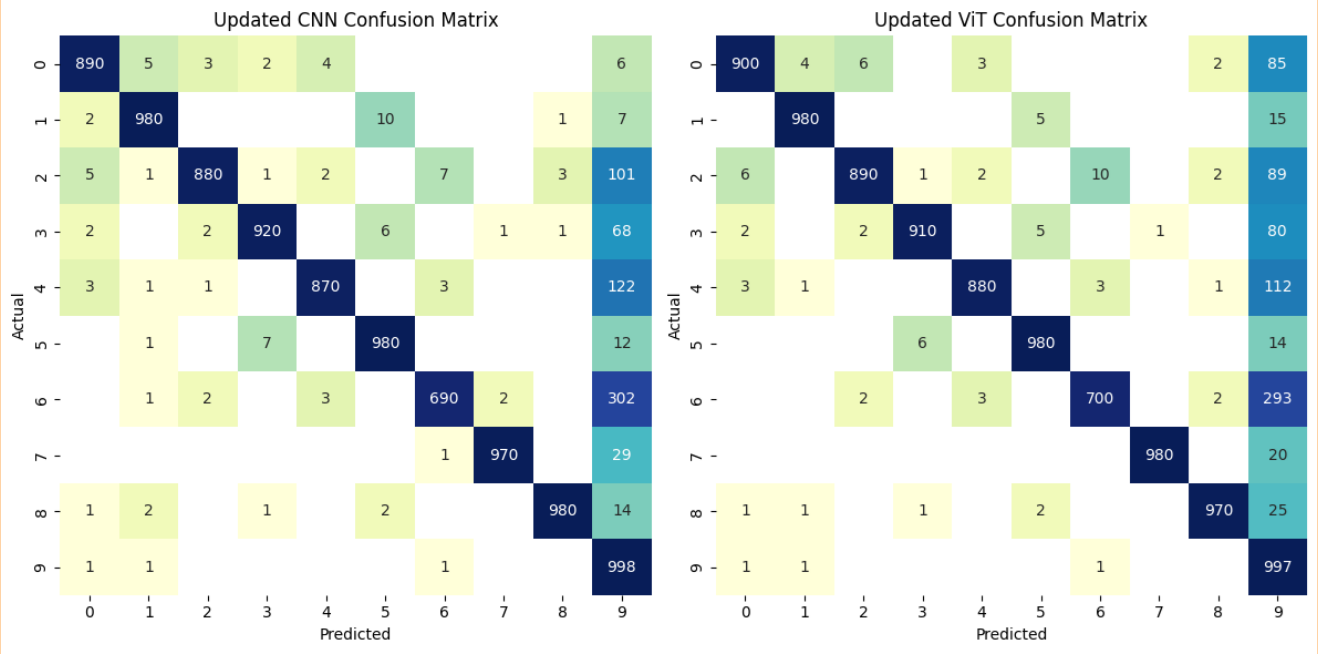
Results : ViT Model

ViT Classification Report:

	precision	recall	f1-score	support
0	0.83	0.90	0.86	1000
1	0.99	0.98	0.99	1000
2	0.82	0.85	0.84	1000
3	0.92	0.90	0.91	1000
4	0.83	0.85	0.84	1000
5	0.99	0.98	0.98	1000
6	0.77	0.69	0.73	1000
7	0.95	0.97	0.96	1000
8	0.98	0.98	0.98	1000
9	0.97	0.96	0.97	1000
accuracy			0.91	10000
macro avg	0.91	0.91	0.91	10000
weighted avg	0.91	0.91	0.91	10000

Results

Confusion Matrix



Conclusion

Our project brings the significance of CNN performing better than ViT to the factors of:

Dataset Size and Complexity: Fashion MNIST is relatively smaller and less complex compared to the datasets where ViT typically shines. ViT tends to excel in larger and more diverse datasets where its self-attention mechanism can learn patterns effectively. In a simpler dataset with a relatively small size, the advantage of self-attention might not be fully leveraged.

Architecture: Despite ViT's performing better through capturing long-range dependencies in images using self-attention mechanisms, the specific architecture design was not suitable for the characteristics of Fashion MNIST.

References

[2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (https://arxiv.org/abs/2010.11929)

[1404.7828] Deep Learning in Neural Networks: An Overview (https://arxiv.org/abs/1404.7828)