Project Title: Clothing Item Classification Using CNN and Spark

By: Feven Tefera
Mihret Kemal

Goal

- Classify clothing items into 10 categories using Fashion MNIST
- Address challenges:
  - Overfitting - with drop out and early stopping
  - Computational efficiency - with Spark

Real World Application

- Online clothing stores like **Amazon**, and **Zara** can use the model to classify product images into categories
- Visual search capabilities where users can find visually similar items based on images **(Google, Pinterest)**
- **Amazon** or **eBay** could use automated captions for product descriptions based on images

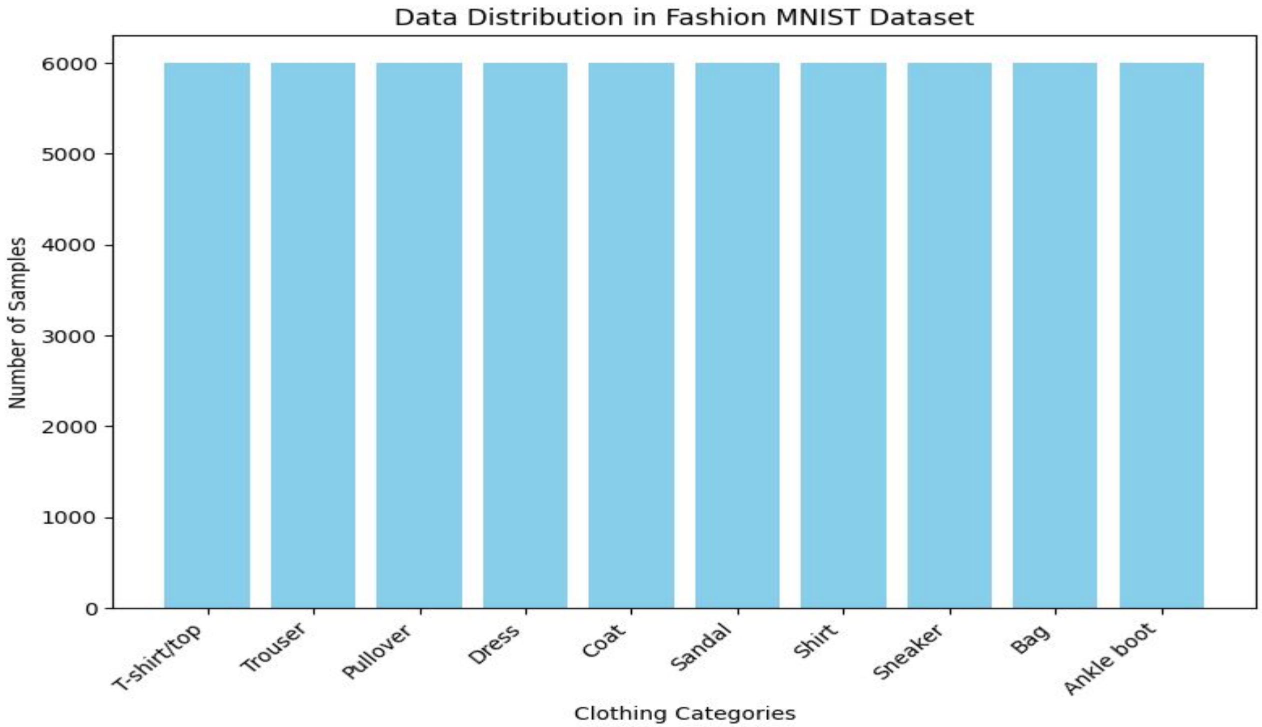
Dataset Overview

- **Dataset**: Fashion MNIST
- **Categories**: T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle Boot
- **Sample Size**:
    - 60,000 training images
    - 10,000 test images

Data split: 90% - training set, and 10% validation set

Data Distribution

- Equal distribution across 10 categories.



Implementation Tools

- **Libraries**: PySpark, TensorFlow, Keras, NumPy, Sklearn, MatplotLib
- **Processing**:

  - Spark for distributed data handling

  - Keras for CNN design and training

CNN Model Architecture

Model Architecture:
Model: "sequential_25"

Layer (type)	Output Shape	Param #
conv2d_50 (Conv2D)	(None, 26, 26, 32)	320
max_pooling2d_50 (MaxPooling2D)	(None, 13, 13, 32)	0
conv2d_51 (Conv2D)	(None, 11, 11, 64)	18,496
max_pooling2d_51 (MaxPooling2D)	(None, 5, 5, 64)	0
dropout_50 (Dropout)	(None, 5, 5, 64)	0
flatten_25 (Flatten)	(None, 1600)	0
dense_50 (Dense)	(None, 128)	204,928
dropout_51 (Dropout)	(None, 128)	0
dense_51 (Dense)	(None, 10)	1,290

Total params: 225,034 (879.04 KB)
Trainable params: 225,034 (879.04 KB)
Non-trainable params: 0 (0.00 B)

Tuning Hyperparameters

- Learning rate: {0.01, 0.001}
- Batch size: {64,32}
- Dropout: {0.3,0.5}
- We used **Grid search** with 3 fold cv for hyperparameter tuning

Best Hyperparameters

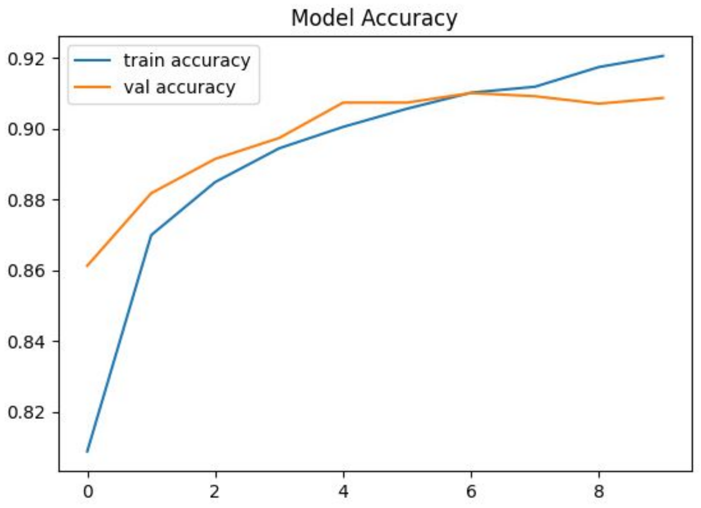
- Learning rate: 0.001
- Batch size: 32
- Dropout: 0.3

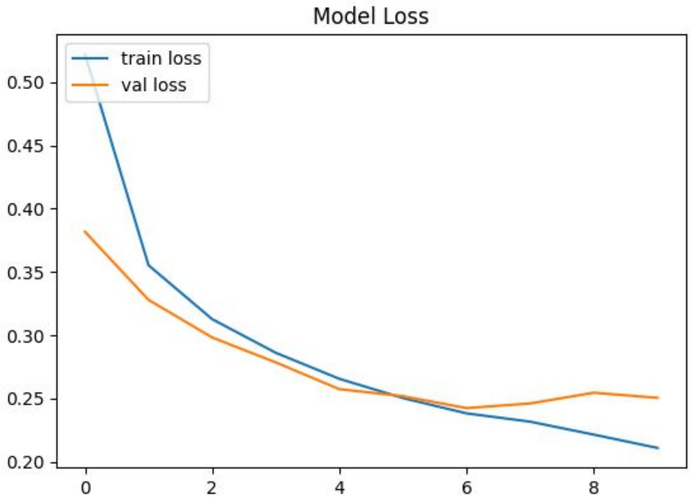
```
Best parameters: {'batch_size': 32, 'model__dropout_rate': 0.3, 'model__learning_rate': 0.001}
Best accuracy: 0.9096000000000001
```

```
Epoch 1/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 156s 82ms/step - accuracy: 0.7414 - loss: 0.7043 - val_accuracy: 0.8612 - val_loss: 0.3818
Epoch 2/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 202s 82ms/step - accuracy: 0.8646 - loss: 0.3694 - val_accuracy: 0.8817 - val_loss: 0.3280
Epoch 3/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 207s 85ms/step - accuracy: 0.8862 - loss: 0.3123 - val_accuracy: 0.8914 - val_loss: 0.2983
Epoch 4/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 171s 68ms/step - accuracy: 0.8938 - loss: 0.2844 - val_accuracy: 0.8973 - val_loss: 0.2786
Epoch 5/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 144s 70ms/step - accuracy: 0.9005 - loss: 0.2667 - val_accuracy: 0.9073 - val_loss: 0.2575
Epoch 6/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 141s 69ms/step - accuracy: 0.9059 - loss: 0.2501 - val_accuracy: 0.9073 - val_loss: 0.2519
Epoch 7/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 142s 69ms/step - accuracy: 0.9084 - loss: 0.2409 - val_accuracy: 0.9100 - val_loss: 0.2425
Epoch 8/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 141s 69ms/step - accuracy: 0.9122 - loss: 0.2280 - val_accuracy: 0.9091 - val_loss: 0.2462
Epoch 9/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 141s 68ms/step - accuracy: 0.9185 - loss: 0.2196 - val_accuracy: 0.9070 - val_loss: 0.2547
Epoch 10/10
1875/1875 ━━━━━━━━━━━━━━━━━━━━ 142s 68ms/step - accuracy: 0.9214 - loss: 0.2065 - val_accuracy: 0.9086 - val_loss: 0.2506
```

Performance Metrics

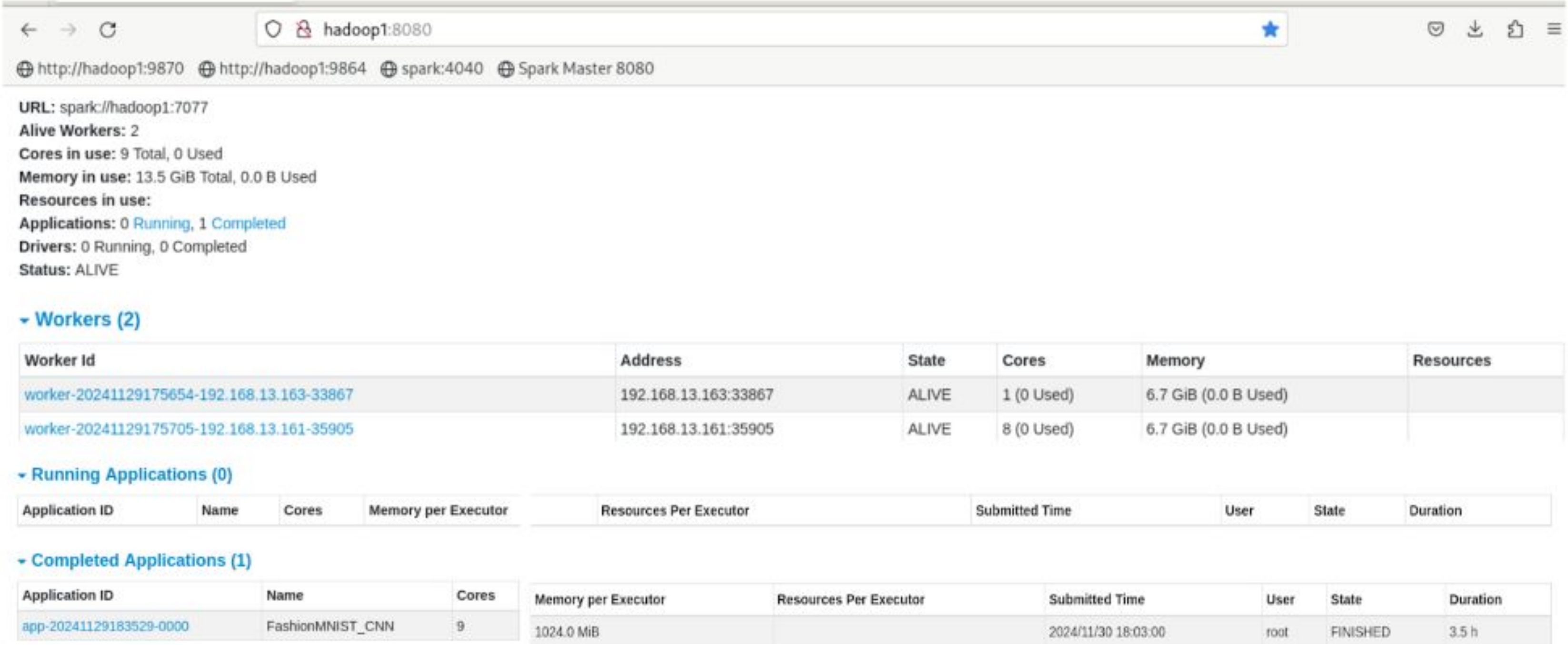
- **Test Accuracy**: 0.908599
- **Test Loss**: 0.2506





Model Performance on Two VMs

Time: 3.5 hours



URL: spark://hadoop1:7077
Alive Workers: 2
Cores in use: 9 Total, 0 Used
Memory in use: 13.5 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241129175654-192.168.13.163-33867	192.168.13.163:33867	ALIVE	1 (0 Used)	6.7 GiB (0.0 B Used)	
worker-20241129175705-192.168.13.161-35905	192.168.13.161:35905	ALIVE	8 (0 Used)	6.7 GiB (0.0 B Used)	

Running Applications (0)

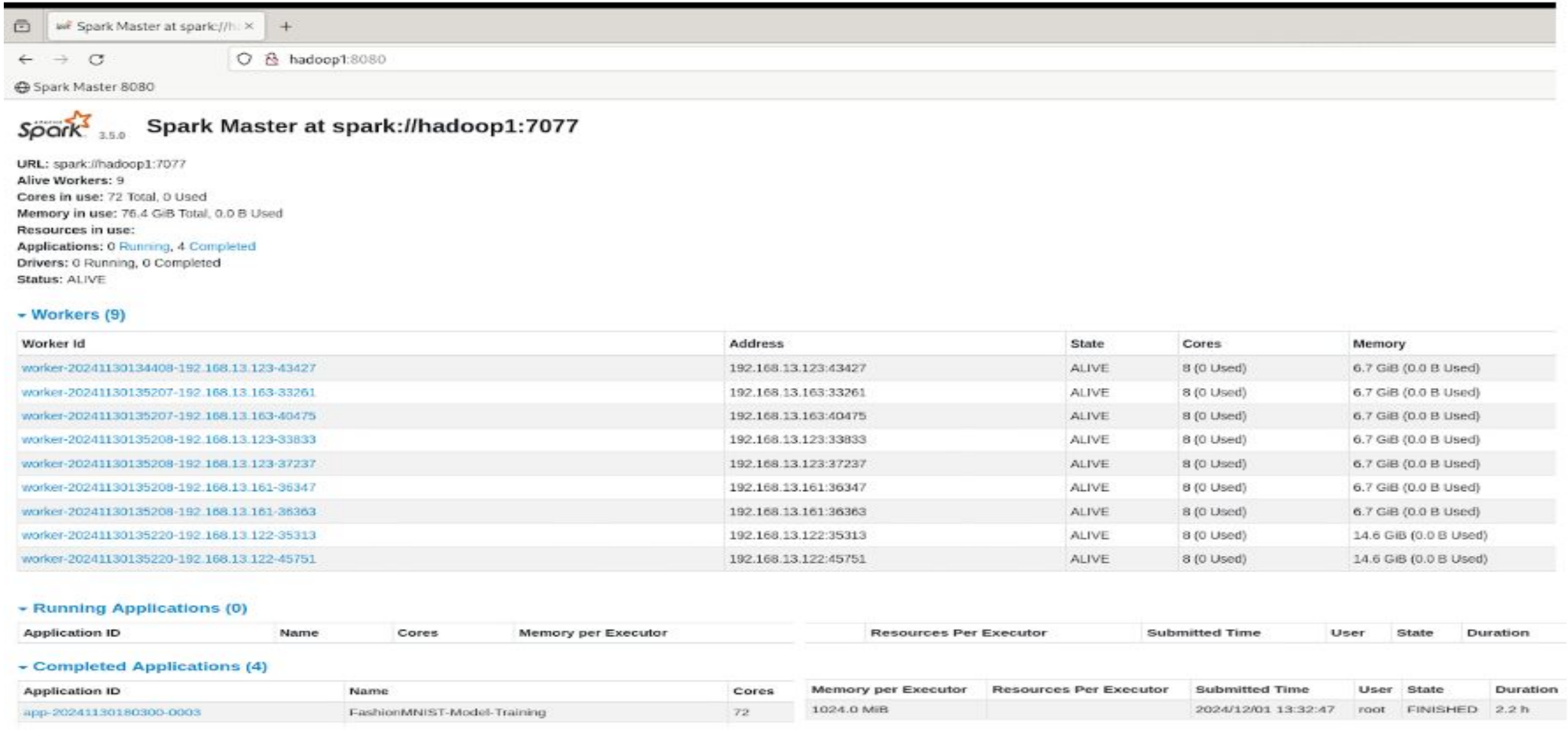
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241129183529-0000	FashionMNIST_CNN	9	1024.0 MiB		2024/11/30 18:03:00	root	FINISHED	3.5 h

Combined Performance of Four VMs:

Time: 2.2 hours



Spark Master at spark://h: ×
hadoop1:8080
Spark Master 8080

Spark 3.5.0 **Spark Master at spark://hadoop1:7077**

URL: spark://hadoop1:7077
Alive Workers: 9
Cores in use: 72 Total, 0 Used
Memory in use: 76.4 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 4 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (9)

Worker Id	Address	State	Cores	Memory
worker-20241130134408-192.168.13.123-43427	192.168.13.123:43427	ALIVE	8 (0 Used)	6.7 GiB (0.0 B Used)
worker-20241130135207-192.168.13.163-33261	192.168.13.163:33261	ALIVE	8 (0 Used)	6.7 GiB (0.0 B Used)
worker-20241130135207-192.168.13.163-40475	192.168.13.163:40475	ALIVE	8 (0 Used)	6.7 GiB (0.0 B Used)
worker-20241130135208-192.168.13.123-33833	192.168.13.123:33833	ALIVE	8 (0 Used)	6.7 GiB (0.0 B Used)
worker-20241130135208-192.168.13.123-37237	192.168.13.123:37237	ALIVE	8 (0 Used)	6.7 GiB (0.0 B Used)
worker-20241130135208-192.168.13.161-36347	192.168.13.161:36347	ALIVE	8 (0 Used)	6.7 GiB (0.0 B Used)
worker-20241130135208-192.168.13.161-36363	192.168.13.161:36363	ALIVE	8 (0 Used)	6.7 GiB (0.0 B Used)
worker-20241130135220-192.168.13.122-35313	192.168.13.122:35313	ALIVE	8 (0 Used)	14.6 GiB (0.0 B Used)
worker-20241130135220-192.168.13.122-45751	192.168.13.122:45751	ALIVE	8 (0 Used)	14.6 GiB (0.0 B Used)

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (4)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241130180300-0003	FashionMNIST-Model-Training	72	1024.0 MiB		2024/12/01 13:32:47	root	FINISHED	2.2 h

Challenges and Solutions

Challenges:

- Preventing overfitting
- Computational resource constraints

Solutions:

- Dropout layers and early stopping(patience = 3)
- Efficient training using Spark

Thank you!

Any Questions?