**FASHION\_MNIST**

* **MLP:** Starting performance yielded a test accuracy of 0.8899000287055969.

|  |  |
| --- | --- |
| **Parameter name** | **Value** |
| Layers (exluding output) | 2 |
| Units | 512 |
| Dropout | 0.2 each |
| Activation | SiLu, SiLu, Softmax |
| Optimiser | RMSPROP |
| Epochs | 20 |

* Increased dropout for regularisation
* Increased depth to 4 layers
* Added batch normalisation
* Changed optimiser to ADAM

This improved the accuracy to 0.9010000228881836

* **CNN:** Starting performance yielded a test accuracy of 0.72589999437332150.7

|  |  |
| --- | --- |
| **Parameter name** | **Value** |
| Layers (exluding output) | 4 |
| Units | 32,64,128 |
| Dropout | 0.25,0.5 |
| Activation | ReLu |
| Optimiser | Adedelta |
| Epochs | 20 |

* Adding more Conv2D layers (3 per block) and changing optimiser to Adam elevates accuracy all the way to 0.9373000264167786

**CIFAR-10**

* **MLP:** Proposed MLP architecture achieved a test accuracy of 0.5302000045776367
* **CNN:** Proposed CNN architecture achieved a test accuracy of 0.8593999743461609

This goes on to show that MLP is severely limited with regard to image classification. Therefore, our CNN implementation is able to generalise much better (even though on CIFAR-10 it still falls significantly short of its Fashion MNIST performance)

**TELL THE TIME NETWORK – 75x75 dataset**

* The first approach we tried is arguably the most intuitive one with 720-way classification.

1. Data processing

First, we transformed the original (hour,minute) format labels into combined hour\*60 + minute label space. Afterwards we split the dataset 0.8/0.1/0.1 (just like before) and prefetched the data, which – as per official TensorFlow guide – reduces training time

*Prefetching overlaps the preprocessing and model execution of a training step. While the model is executing training step s, the input pipeline is reading the data for step s+1. Doing so reduces the step time to the maximum (as opposed to the sum) of the training and the time it takes to extract the data. (*[*https://www.tensorflow.org/guide/data\_performance*](https://www.tensorflow.org/guide/data_performance)*)*

1. *Network Architecture*

As for the neural network architecture, we decided to make it sufficiently deep in order for its capacity to be suitable for coping with this many different classes. This resulted in 4 separate blocks being formed:

* Block 1 for low level feature extraction:

2x convolution layer of size 64 with (3,3) kernel and ReLu activation

Followed by (2,2) max pooling and a 0.2 dropout layer

* Block 2 for mid-level features:

The same but with convolutions of size 128

* Block 3 for high-level features:

One convolutional layer of size 256 with (3,3) kernel

Followed by (2,2) max pooling and a 0.2 dropout

* Block 4 of dense layers to combine the aforementioned features into meaningful information

Input flattening followed by a 1024-size dense layer and 512-size dense layer, with 0.4 dropout after each

* Output layer with softmax activation appropriate for multi-class classfication

1. Training and compilation

The loss function of choice is sparse categorical crossentropy as our true labels are provided in the form of integers (as opposed to one-hot encoding).

Further, we used 3 callback functions to better steer the learning process:

* **ReduceLROnPlateau** - Adaptive Learning Rate: we want to make good progress at the beginning (fast learning), but as we approach the theoretical optimum slower LR becomes more desirable for escaping plateaux. If validation loss does not improve for 3 consecutive epochs, we multiply it by a factor of 0.5
* **EarlyStopping** - Prevents Overfitting: In case of 7 consecutive epochs without improvement on validation loss, we stop training to avoid overfitting
* **ModelCheckpoint** - Saves Best Model: As soon as a validation loss improvement is registered, the incumbent models get saved as a back-up against potential crashes.

1. Evaluation

We evaluate our model on the test set by setting its predictions (converted back actual hours/minute values) against the true labels.

As for the common-sense metric, we compute it via *calculate\_time\_difference\_minutes(true\_hours, true\_minutes, pred\_hours, pred\_minutes)*

function that handles clock wrap-around by taking a minimum of error and 720-error (12 hours \* 60 minutes).

We kept increasing the number of epochs until the Early Stopping callback criterion (for validation accuracy) kicked in and found out that point is at Epoch with the following results:

The evaluation results are as follows:

Hour accuracy: 0.6944 (69.44%)

Minute accuracy: 0.2394 (23.94%)

Exact match (both correct): 0.2317 (23.17%)

"Common Sense" Time Difference Accuracy:

Average error: 0 hour(s) and 44 minute(s)

Median error: 0 hour(s) and 1 minute(s)

A graph with red and blue lines

AI-generated content may be incorrect.Std deviation: 84.85 minutes

* Our second approach was to frame it as a multi-head output problem, i.e. to treat hours and minutes separately.

1. Data processing:

We treat labels as tuples:

dataset = tf.data.Dataset.from\_tensor\_slices((data, (hour\_labels, minute\_labels)))

Data split and prefetching as above

1. Network architecture:

CNN backbone the same as with multi-class classification, the difference manifests itself in the dual nature of dense layers:

**Hour head**: Specialized branch with 128 units → 12 classes

**Minute head**: Specialized branch with 256 units → 60 classes (larger because it's a harder task)

Output is naturally also twofold: [hour\_output, minute\_output]

1. Training and compilation:

Both output heads use sparse categorical crossentropy and the same callback functions as the model above.

1. Evaluation:

We kept increasing the number of epochs until the Early Stopping callback criterion kicked in and found out that point is at Epoch 92 with the following results:

Hour accuracy: 96.78%

Minute accuracy: 49.28%

Exact match (both correct): 48.33%

"Common Sense" Time Difference Accuracy:

Average error: 0 hour(s) and 2 minute(s)

Median error: 0 hour(s) and 1 minute(s)

Std deviation: 10.30 minutes

Average hour error: 0.03 hours

Average minute error: 0.68 minutes

A graph of a graph

AI-generated content may be incorrect.

* Final approach: regression with angle conversion

The most sophisticated way of addressing this problem is to treat it as multi-output regression

1. Data processing:

We converted all the hour and minute labels into angles and then converted those angles into sin/cos function (to handle circularity issues).

Then we stacked them as 4 outputs via np.stack([hour\_sin, hour\_cos, minute\_sin, minute\_cos], axis=1)

1. Network architecture

Here we found out that fewer layers are sufficient to elicit good test and validation performance by preserving all the important details (mainly hand positions) and consequently settled for the following architecture:

* Block 1 comprising two convolutional 2D layers of size 64, each with kernel size (3,3), standard ReLu activation and batch normalisation in place. Then we added a MaxPooling2D of size (2,2) and 0.2 dropout for regularisation purposes.
* Block 2 is almost identical with the only difference being Convolution layer size (double the previous amount, so 128 each)
* Block 3 consists of one Flatten layer and two dense layers of sizes 512 and 256, each followed by a 0.3 dropout
* Finally, the network yield 4 outputs (sin and cos for hours and minutes respectively). The activation function of choice is tanh – our transformed sin and cos values are between -1 and 1 and so is the range of tanh (ReLu does not output negative values and thus is inferior in this case).

1. Training and compilation

For this regression problem the most natural loss function appeared to be MSE.

1. Evaluation

Importantly, we assess hour and minute accuracy separately by employing our *convert\_predictions\_to\_time* function.

We kept increasing the number of epochs until the Early Stopping callback criterion kicked in and found out that point is at Epoch 65 (best result being from epoch 58) with the following results:

Test Results:

Hour Accuracy: 53.00%

Minute Accuracy: 28.39%

Exact Match (both correct): 14.33%

"Common Sense" Time Difference Accuracy:

Average error: 0 hour(s) and 28 minute(s)

Median error: 0 hour(s) and 2 minute(s)

Std deviation: 29.85 minutes

A graph with red and blue lines

AI-generated content may be incorrect.

Quite unexpectedly, the most complex approach did not end up bringing the best results as corroborated by the following combined table of results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric/Model** | 720 class classification | Multi-head Classification | Regression |
| Hour acc | 69.44% | 96.78% | 53.00% |
| Minute acc | 23.94% | 49.28% | 28.39% |
| Exact match | 23.17% | 48.33% | 14.33% |
| Avg error | 0h44min | 0h2min | 0h28min |
| Median error | 0h1min | 0h1min | 0h2min |