Machine Learning in Finance
Course Project

Topic 3 - Daily Stock Return Prediction

Data Pipeline

132 months of daily data, 1.5-2.0 GB per DF.

Used `tf.data.Dataset` to load parquet files piece by piece in parallel.

Custom generator function.

- Read 2 ** 13 records every batch
- Transformed the 2-digit SIC code features using one-hot encoding

Rearranged steps to reduce input time. Interleaving, prefetching for parallel processing but the input pipeline is still a bottleneck.

```
train_gen = \
    tf.data.Dataset\
    .from_tensor_slices(train_files)\
    .interleave(
        lambda filename: tf.data.Dataset.from_generator(
            generator,
            output_signature=(
                tf.TensorSpec(shape=(None, len(COL_X)), dtype=tf.float16),
                tf.TensorSpec(shape=(None, 1), dtype=tf.float32)),
            args=(filename, BATCH_SIZE)),
        num_parallel_calls=tf.data.experimental.AUTOTUNE)\
    .shuffle(len(train_files), reshuffle_each_iteration=True, seed=3141)\
    .prefetch(tf.data.experimental.AUTOTUNE)
```
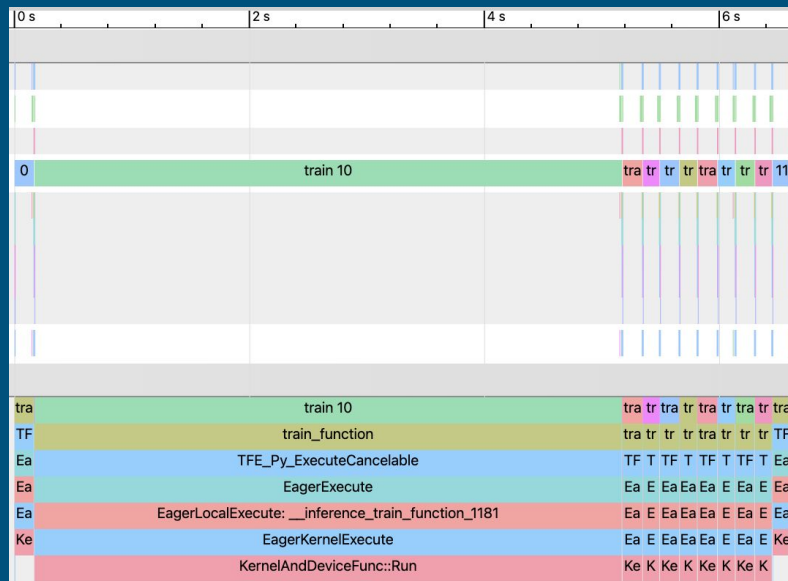
Hyperparameter Tuning

We manually picked (due to time limit)

- Size of rolling windows for tuner
- Batch size → decent GPU utilization
- Epoch
- Kernel initializer
- Kernel Regularizer

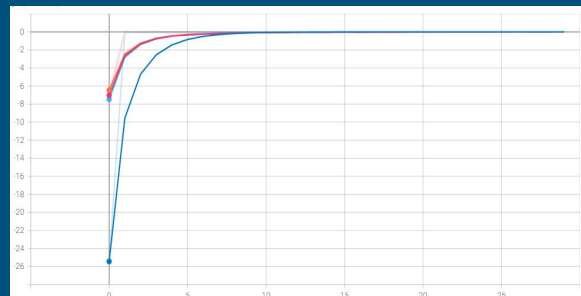`BayesianOptimization` tuner is used to tune

- Number of layers: [4, 6, 8]
- Shape of network: [rectangle, pyramid]
- Activation function: [sigmoid, relu]
- Learning rate: scheduler with exp. Decay [.01,.02]

After experimenting on several training windows, we observe [4, rectangle, sigmoid, 0.02] is the best hyperparameter combination.
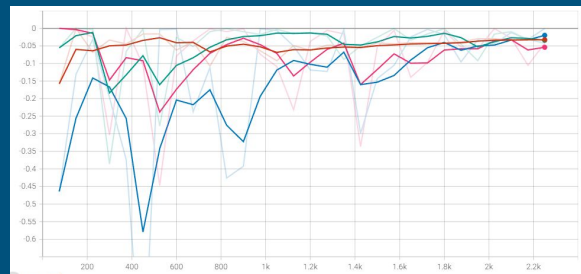
MSE
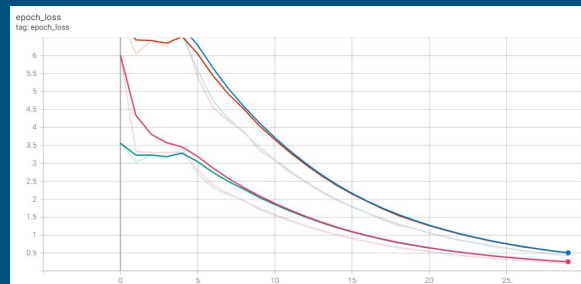
Train R^2

Validation R^2
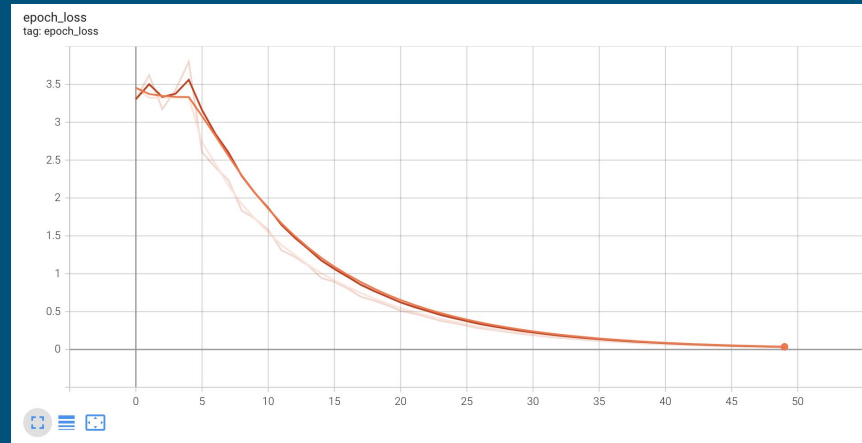
Model Performance

Hardware Setup:

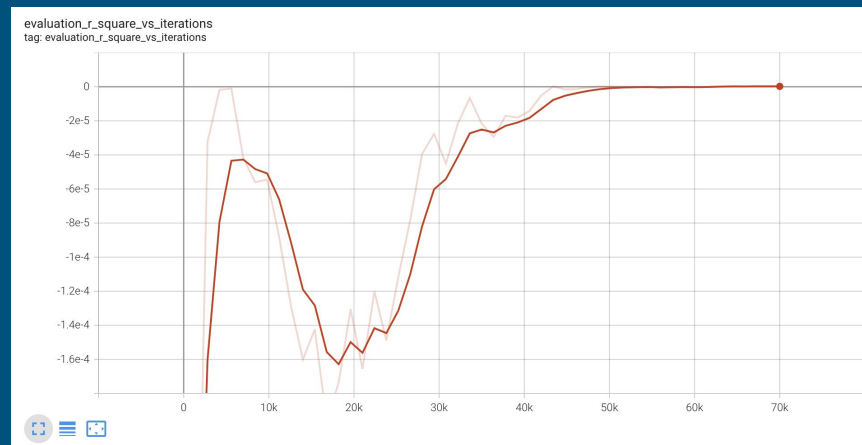Tuning: AMD Ryzen 3800X + NVIDIA 1080Ti (11GB VRAM) + 32GB RAM

Training: (Colab pro) Tesla T4 (16GB VRAM) + 25 GB RAM

Finally, we choose the best hyperparameters to train the model with full data (epoch = 50):
108 months of training
12 months of validation
12 months of test

MSE

Validation R^2

Thank!

Q&A

Chuang Yao
Jianqi Zhu
Zhaoning Qi