This code is used to generate sales prediction for the Kaggle Rossmann Sales competition with deep neural networks.

To effectively incorporate category features, we proposed entity embedding to encode category features using learnt vectors. Like semantic embedding in natural language processing entity embedding enables us to express and learn the complex relations of different categories in a multi-dimensional vector space. This helps to deal with data sparsity and overfitting. The neural network we use has 3 fully connected layers on top of all embeddings and other non-category features. The final submission is an average of the predictions of 10 networks. The best single model I tested is 0.098x on the last 3% training data as validation. Averaging 5 or 10 nns can boost this further, the best score I had is 0.094x.

To run the code one needs first download and unzip the `train.csv` , `test.csv` and `store.csv` files on Kaggle and put them inside this folder. I have already included the extra store states, weather and google trend data shared in the competition forum by dune_dweller, MCFG and Tobias Wolfanger respectively, so you don’t need to download them.

Next run the following scripts to extract and prepare features:

```
python3 extract.py
python3 extract_weather.py
python3 extract_google_trend.py
python3 extract_fb_features.py
python3 prepare_features.py
```

To test the neural network model run (you need to have keras installed first)

```
python3 test_model.py
```

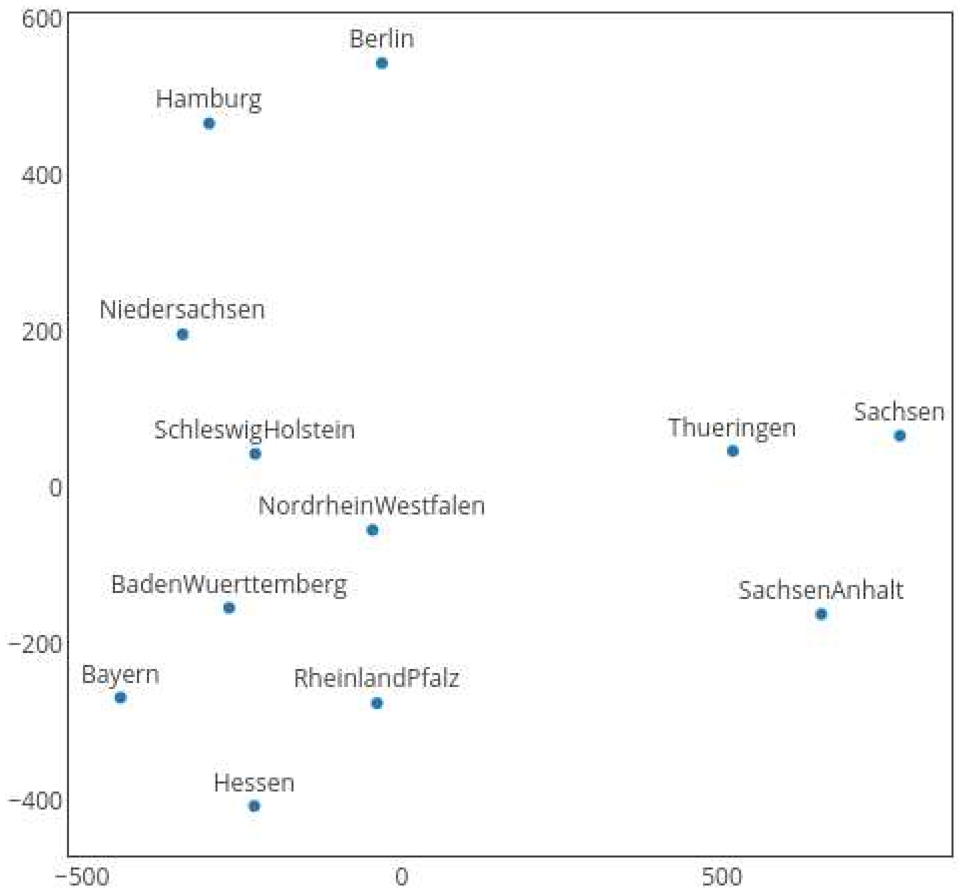
By default it will run one neural net with 0.97 data for training and the rest for test. It takes 20 minutes to run on Nvidia GTX 980 GPU, and it may take a few hours to run on CPU.

You can change these two parameters in `test_model.py` if you want to use more models or a different train-test ratio, and the following is what I used for finial submission:

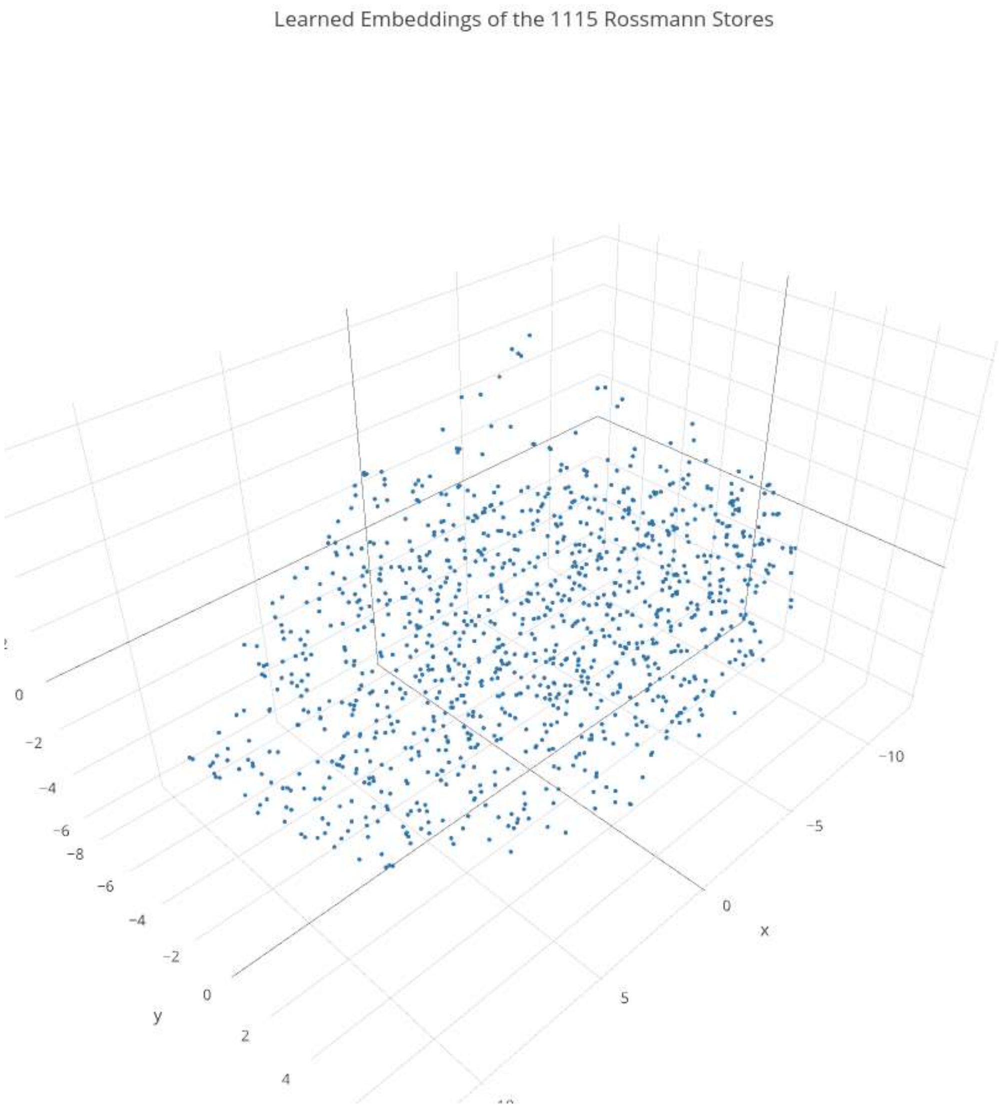
```
num_networks = 10
train_ratio = 1
```

After the script finishes it will generate a file `predictions.csv` which is used for submission to Kaggle.

You can anaylize the embeddings with the ipython notebook included. This is the learned embeeding of German States printed in 2D:


Learned Embedding of German States from Kaggle Rossmann Competition

and this is the learned embeddings of 1115 Rossmann stores printed in 3D:



Acknowledge:

Thank our founder Ozel Christo, Andrei Ciobotar and all colleagues at Neokami for supporting and encouragement!