# Rossmann Store Sales, Winner's Interview: 3rd place, Neokami Inc.

Kaggle Team (http://blog.kaggle.com/author/kaggleteam/) | 01.22.2016


cheng-

gui/#co

Rossmann operates over 3,000 drug stores in 7 European countries. In their first Kaggle competition, Rossmann Store Sales (https://www.kaggle.com/c/rossmann-store-sales), this drug store giant challenged Kagglers to forecast 6 weeks of daily sales for 1,115 stores located across Germany. The competition attracted 3,738 data scientists, making it our second most popular competition by participants ever.

Cheng Guo competed as team Neokami Inc. and took third place using a method, "entity embedding", that he developed during the course of the competition. In this blog, he shares more about entity embedding, why he chose to use neural networks (instead of the popular xgboost), and how a simplified version of his model still manages to perform quite well.

# The Basics

### What was your background prior to entering this challenge?

I work at Neokami (https://www.neokami.com/), a machine learning startup located in Munich. I built the neural network behind some of our computer vision products such as VisualCortex (https://www.visualcortex.io/), which lets you create your own image classifier easily. I hold a PhD in theoretical physics and have developed algorithms to simulate quantum systems while studying at Ludwig Maximilians University (http://www.en.uni-muenchen.de/index.html) in Munich and Chinese Academy of Sciences (http://english.cas.cn/) in Beijing.


(http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2016/01/Screen-Shot-2016-01-21-at-1.59.45-PM.png)

PROFILE (HTTPS://WWW.KAGGLE.COM/ENTRON) FOR CHENG GUO (AKA ENTRON) ON KAGGLE

## What made you decide to enter this competition?

We had a busy year developing our product, but I was not so busy for a few weeks before Christmas. My colleague Felix reminded me that I could join some Kaggle contest. I knew that a very popular German chain store Rossmann had been running a competition on Kaggle for some time, and I thought it would be fun to join it. Our founders Ozel and Andrei liked the idea and supported me, so I joined in the last month of the competition.

## What have you taken away from this competition?

First, I invented a new method "entity embedding" for this competition. It is a general method and can be applied to many other problems.

Second, a special part about this Rossmann competition is that external data is allowed as long as it is shared in the forum. This brought lots of exciting explorations, insights and fun just like in scientific research. There are so many smart and passionate people in Kaggle and there are also many important and difficult questions waiting to be solved. If those problems are carefully divided and well formulated into small and easy to understand subproblems it may be solved collaboratively by the Kaggle community. This is a world changing potential.

# Just For Fun

### If you could run a Kaggle competition, what problem would you want to pose to other Kagglers?

I recently read a news article (http://www.reuters.com/article/us-china-pollution-idUSKBN0UB1KB20151229) about IBM and Microsoft's effort to forecast China's smog. I think it could have a tremendous environmental, social and economical value to run a competition with similar external data policy like Rossmann Sales to forecast smog. This will help to pinpoint what are the most important contributors (factories/sources etc.) to the air pollution.

# Let's Get Technical

### What supervised learning methods did you use?

Deep neural network is very powerful and flexible and it is already the dominant method in many machine learning problems like computer vision and natural language processing. When reading the Rossmann competition forum I was surprised that most top teams used tree based methods like xgboost (https://github.com/dmlc/xgboost) rather than neural network. As a fan of neural networks I decided to use only neural network and see how it compares with xgboost.

To make a neural network work effectively on this type of problem which has many category features, I proposed a new method **Entity Embedding** to represent category features in a multi-dimensional space. It is inspired by semantic embedding in the natural language processing domain. With entity embedding, I found that neural networks generate better results than xgboost when using the same set of features.

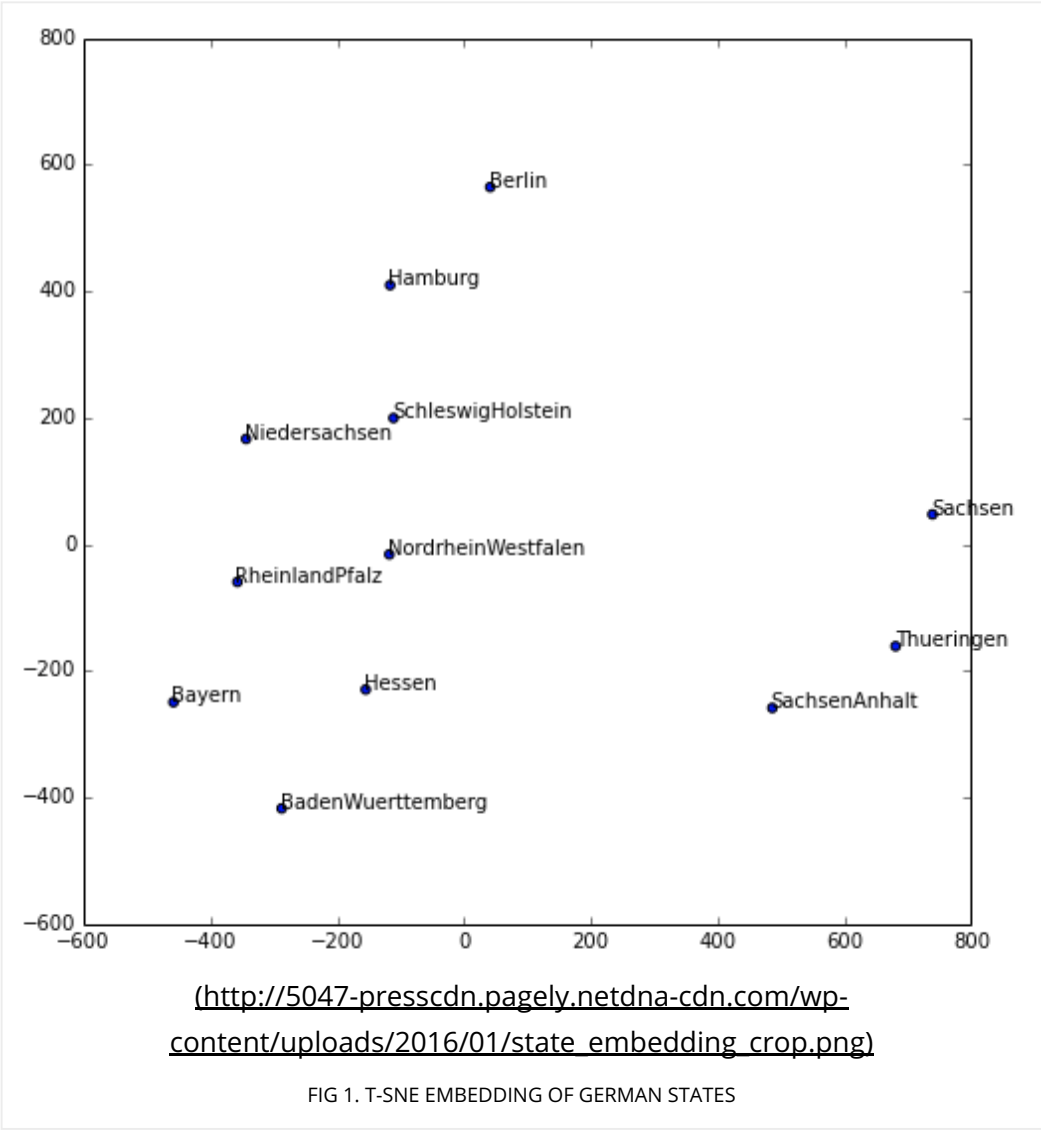
I also used an unusual small dropout 0.02 after the input layer to improve the generalization. The reasoning is that the 0.02 dropout will randomly remove one small feature or a few dimensions of large features and the model should still be able to get similar result based on the remaining features.

I have shared our code in the Kaggle forum here (https://www.kaggle.com/c/rossmann-store-sales/forums/t/17974/code-sharing-3rd-place-category-embedding-with-deep-neural-network).

## What was your most important insight into the data?

My second favorite approach to the Rossmann Sales problem is to use the historical median of the 4 features (*store_index, day_of_week, promo, year*) as the prediction. Its score on my validation set, which is close to the final score on the leader board, is 0.133. The result is amazingly good considering how simple this approach is. I wanted to include more features to improve it, unfortunately then comes the data sparsity problem. More specifically, after I added another feature "*month*", some combinations of the 5 features don't have any historical record in the dataset leave aside the median. To overcome the data sparsity problem I got the idea to represent the discrete category features in a continuous space in which the distance between different "category points" reflects the similarity of the categories. This is the idea behind the entity embedding method. In this way one can interpolate or use nearby data points to approximate missing data points.

To get an intuitive idea about entity embedding, I used t-SNE to map the high dimensional embeddings into 2D figures. First, let's see the German states (Fig. 1).




(http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2016/01/state_embedding_crop.png)

FIG 1. T-SNE EMBEDDING OF GERMAN STATES

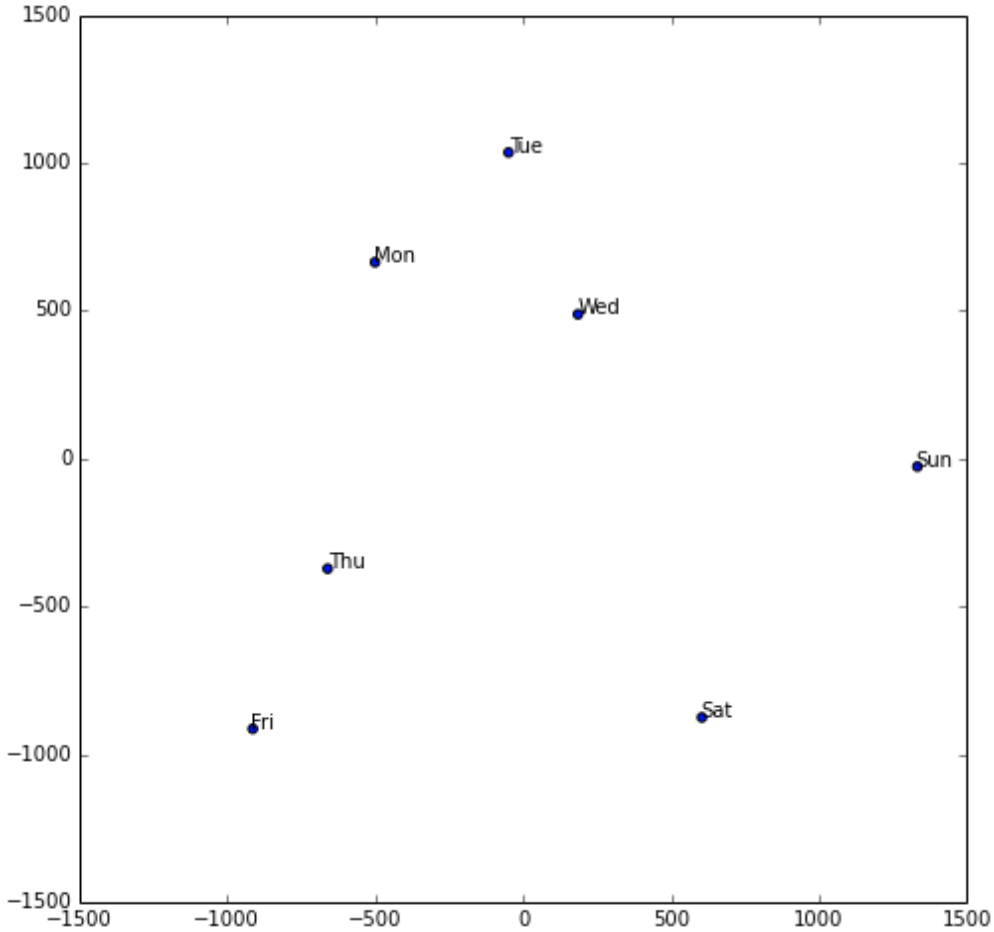
Though the algorithm does not know anything about German geography and society, the relative positions on the learnt embedding of German states resemble that on the below map (Fig. 2) surprisingly well!


(http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2016/01/German_States.png)

FIG. 2 MAP OF GERMAN STATES

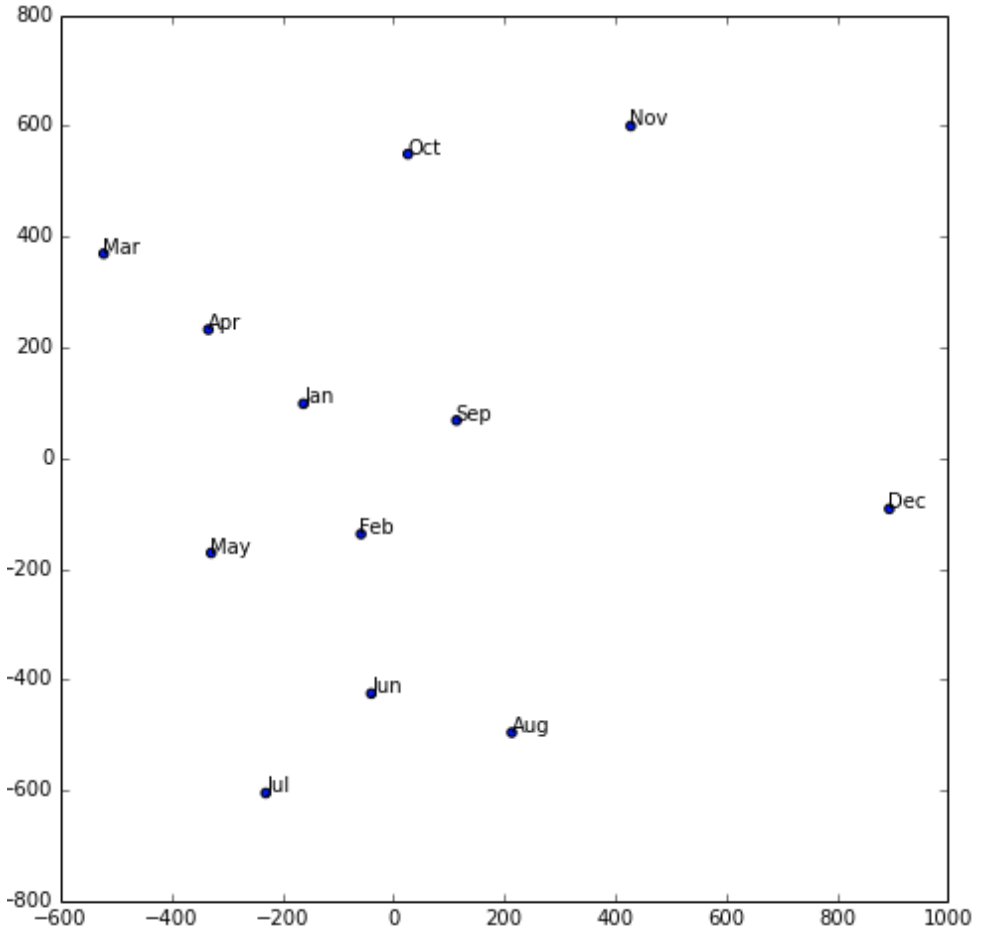
The reason is that the embedding maps states with similar distribution of features, i.e. similar economical and cultural environments, close to each other, while at the same time two geographically neighboring states are likely sharing similar economy and culture. Especially, the three states on the right cluster, namely "Sachsen", "Thueringen" and "SachsenAnhalt" are all from eastern Germany while states in the left cluster are from western Germany. This shows the effectiveness of entity embedding for abductive reasoning.

Similarly, the following are the learnt embeddings (after converted to 2D) of day of week (Fig. 3), month (Fig. 4) and Rossmann stores (Fig. 5).



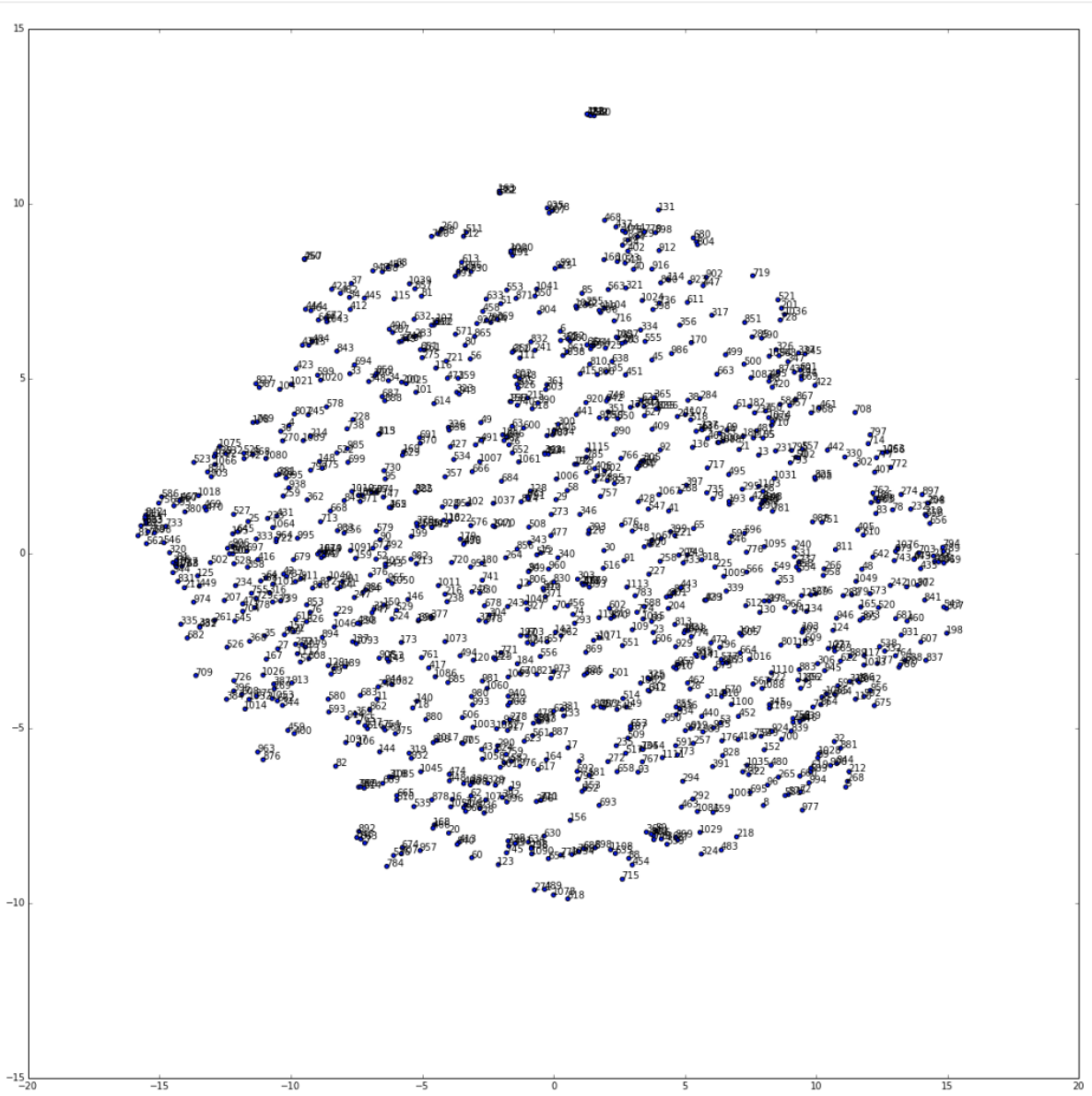

(http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2016/01/dow_embedding_crop.png)

FIG. 3 LEARNT EMBEDDING OF DAYS OF THE WEEK




(http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2016/01/month_embedding_crop.png)

FIG. 4 LEARNT EMBEDDING OF MONTHS



(http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2016/01/store_embedding_crop.png)

FIG. 5 LEARNT EMBEDDED OF ROSSMANN STORES

Entity embedding may be applied to many other problems to find the hidden relations between entities based on their interaction with the external environment. For example, based on huge databases about the relations of genes, mutations, proteins, medicines and diseases, one may map those entities into multi-dimensional spaces which can guide the understanding of biological process or drug discovery etc. This an exciting direction for further exploration.

## Which tools did you use?

I used a new python neural network frame work Keras (https://github.com/fchollet/keras). It is simple, flexible and powerful. It can use Theano (https://github.com/Theano/Theano) or TensorFlow (https://www.tensorflow.org/) as the backend. I also used many common python packages like sklearn (http://scikit-learn.org/stable/), numpy (http://www.numpy.org/) and pandas (http://pandas.pydata.org/). I used Nvidia GTX 980 GPU to run the neural network as it is more than one order of magnitude faster than a CPU.

## What was the run time for both training and prediction of your winning solution

It takes 20 minutes to train one network on GPU. For our finial submission we averaged the result of 10 networks, so altogether it takes about 3.5 hours. The time spend on the prediction is little.

---


DEEP NEURAL NETWORKS (HTTP://BLOG.KAGGLE.COM/TAG/DEEP-NEURAL-NETWORKS/)

REGRESSION PROBLEM (HTTP://BLOG.KAGGLE.COM/TAG/REGRESSION-PROBLEM/)

REVENUE FORECAST (HTTP://BLOG.KAGGLE.COM/TAG/REVENUE-FORECAST/)

ROSSMANN STORE SALES (HTTP://BLOG.KAGGLE.COM/TAG/ROSSMANN-STORE-SALES/)

1 Comment No Free Hunch 1 Login

Recommend Share Sort by Best

Join the discussion…

LOG IN WITH OR SIGN UP WITH DISQUS ?

Name

Dan Ofer • 2 years ago

Where in your Git is the entity embedding code/description exactly? I'd like to dig into it. Great idea!

Reply • Share ›

Subscribe Add Disqus to your siteAdd DisqusAdd Privacy

(https://www.facebook.com/kaggle) (https://twitter.com/kaggle)