Neural Nets in NLP Competitions

Two Recent Examples from Kaggle

Carlos Sathler

April 10, 2018

SCHOOL OF INFORMATICS AND COMPUTING

INDIANA UNIVERSITY

Department of Information and Library Science
Bloomington

Agenda

- Kaggle
- Competition 1 - Text Regression
- Lessons learned
- Competition 2 - Toxic Comment Classification
- Lessons learned
- Final thoughts

Kaggle

- Internet platform for data science competitions
- Google company since March 2017
- Large community of practitioners
- Great place to learn

- www.kaggle.com

Competition 1

- Mercari Price Suggestion Challenge by mercari.com
- From item name, description, and a few categorical features, predict item price => Text regression problem
- $100,000 in prizes
- Finished 11%

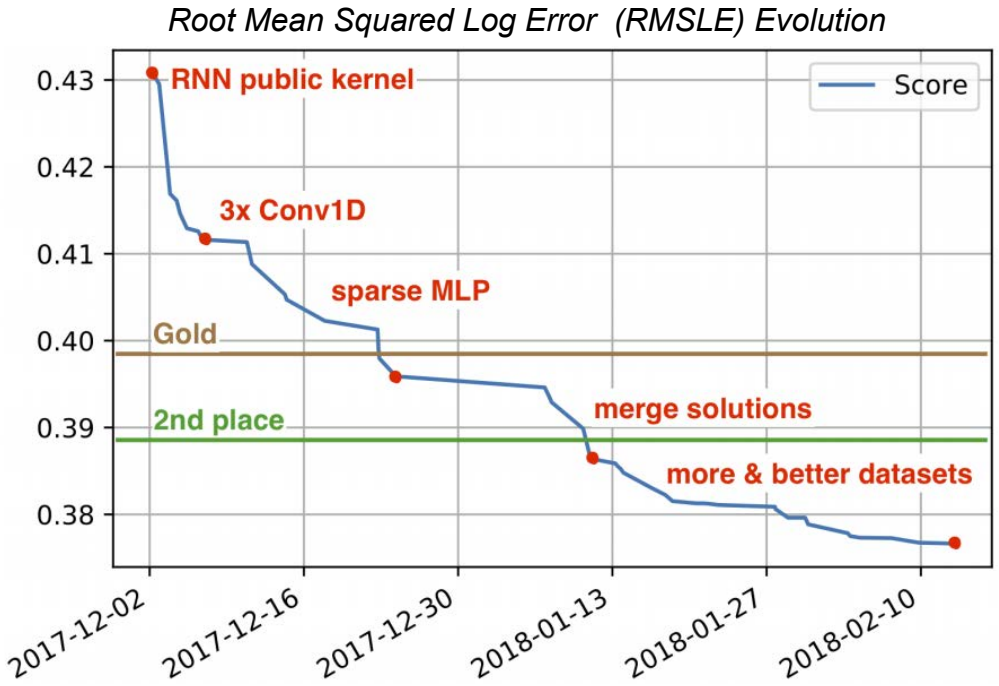
- Competition site: https://www.kaggle.com/c/mercari-price-suggestion-challenge
- Dataset: https://github.com/charrtay/Mercari-Price-Suggestion/blob/master/src/models/sathler_dviz_01.ipynb

Lessons Learned

- Simple works (sparse MLP models)
- Three datasets + combinations
- Ensemble

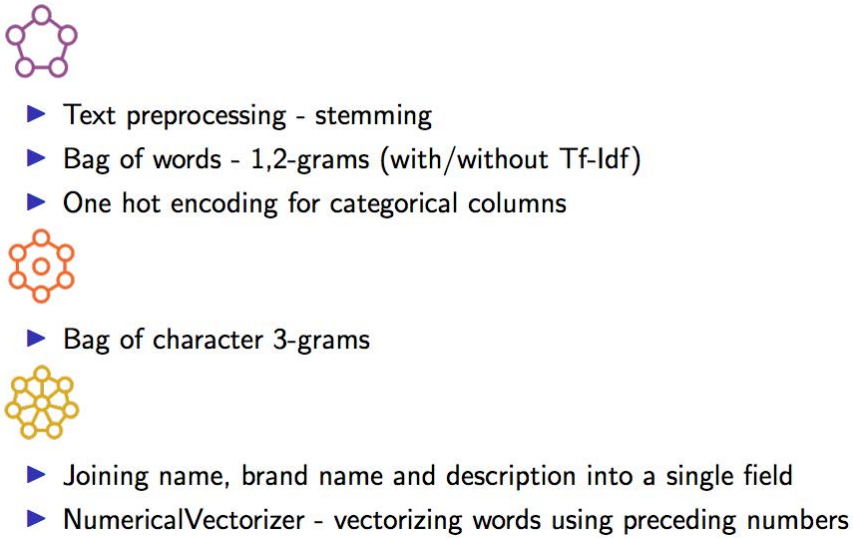
- Winners: @Konstantin_Lopuhin & @Pawel_Jankiewicz
- Winner post: https://www.kaggle.com/c/mercari-price-suggestion-challenge/discussion/50256
- Winner code: https://www.kaggle.com/lopuhin/mercari-golf-0-3875-cv-in-75-loc-1900-s

Lessons Learned



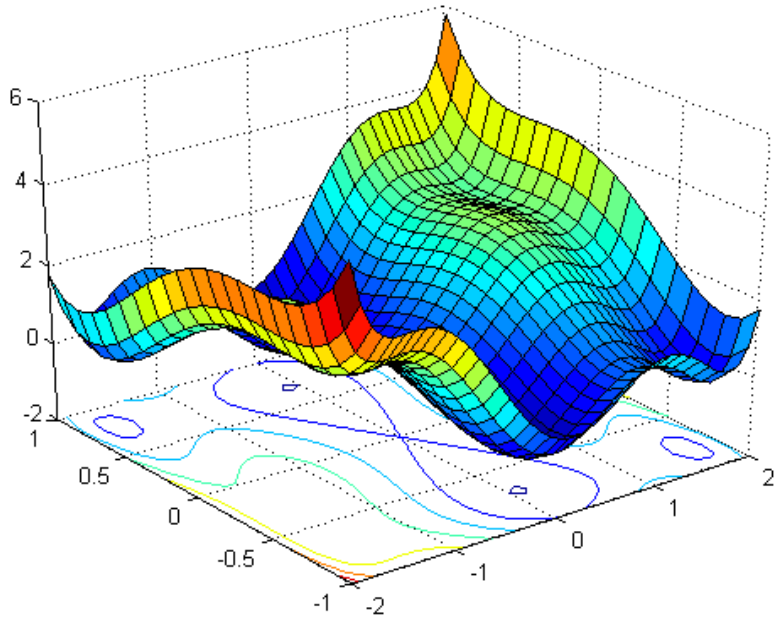
- Figure copied from winner's presentation: https://github.com/pjankiewicz/mercari-solution/raw/master/presentation/build/yandex.pdf

Lessons Learned



- Figure copied from winner's presentation: https://github.com/pjankiewicz/mercari-solution/raw/master/presentation/build/yandex.pdf

Lessons Learned



- Multiple Datasets + Ensembles
- Universal approximation theorem (p. 192 of our book)

Lessons Learned

```
1111100000  Ground truth
1011110100  Weak learner (70%) Good at predicting 1s
1101000010  Weak learner (70%) Good at predicting 0s
0110101001  Weak learner (60%) Not good at predicting anything
1111100000  Vote average of weak learners (100%)
```

- From @tilli post: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/51058

Competition 2

- Toxic Comment Classification Challenge by Jigsaw
- From comment predict: toxic? threat? etc.
- Text classification problem
- $35,000 in prizes
- Finished 55% (not enough time…)

- Competition site: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
- Dataset: https://www.kaggle.com/jagangupta/stop-the-s-toxic-comments-eda (by @jagan)

My Solution (one of them)

```
X_tag_voc_size = nlp_features['X_tag']['voc_size']
X_tag_input = Input(shape=(MAX_SEQ,), dtype='int32')
X_tag_embed = Embedding(X_tag_voc_size, EMBEDDING_DIM, input_length=MAX_SEQ)(X_tag_input)
X_tag_embed = Reshape((MAX_SEQ,EMBEDDING_DIM,1,))(X_tag_embed)

X_dep_voc_size = nlp_features['X_dep']['voc_size']
X_dep_input = Input(shape=(MAX_SEQ,), dtype='int32')
X_dep_embed = Embedding(X_dep_voc_size, EMBEDDING_DIM, input_length=MAX_SEQ)(X_dep_input)
X_dep_embed = Reshape((MAX_SEQ,EMBEDDING_DIM,1,))(X_dep_embed)

X_upper_voc_size = nlp_features['X_upper']['voc_size']
X_upper_input = Input(shape=(MAX_SEQ,), dtype='int32')
X_upper_embed = Embedding(X_upper_voc_size, EMBEDDING_DIM, input_length=MAX_SEQ)(X_upper_input)
X_upper_embed = Reshape((MAX_SEQ,EMBEDDING_DIM,1,))(X_upper_embed)

X_lemma_voc_size = nlp_features['X_lemma']['voc_size']
X_lemma_input = Input(shape=(MAX_SEQ,), dtype='int32')
X_lemma_embed = Embedding(X_lemma_voc_size, EMBEDDING_DIM, input_length=MAX_SEQ)(X_lemma_input)
X_lemma_embed = Reshape((MAX_SEQ,EMBEDDING_DIM,1,))(X_lemma_embed)

# create text window with 4 channels
L1 = concatenate([X_tag_embed, X_dep_embed, X_upper_embed, X_lemma_embed], axis=3)

# first convolutions for a window of 5 words
C1 = Conv2D(64, kernel_size=(50,1), padding='same', strides=(1,1), activation='relu')(L1)
C1 = Dropout(0.5)(C1)
```

- Based on paper “Natural Language Processing (almost) from Scratch.
- Used Spacy to extract NLP features: https://spacy.io/usage/linguistic-features

Lessons Learned (more of the same)

- RNNs (GRU) performed the best
- Diverse word-embeddings
- Dataset augmentation
  - Translated text to French, German, Spanish
  - Then back to English
- Ensamble, ensamble
- Other ML tricks (check link below)

- Winners @Chun_Ming_Lee and @To_train_them_is_my_cause
- Details of the solution: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557

Final thoughts

- Neural nets rule
- More datasets, better results
- Ensemble, better results
- More linguistic features, better results?
- No free lunch

- Team up!

Questions?

- Email csathler@iu.edu
- Please cc Dr. Cavar dcavar@indiana.edu