Duplicate Questions Identification: Exploring Effects of Feature Representation on Deep Neural Nets and Random Forest Model

Fan Fan, Hanyang Li, Jiazhi Zhang
Machine Learning Department

Introduction

Motivation

- An important step of identifying duplicate questions is to properly represent sentences as feature vectors.
- Intermediate signals of neural networks (NN) forward pass before the last fully-connected layers can be used as vectors for sentence representation.
- How representation learned by NN affects performance of non-NN models remains an open question.

Objectives

- Find sentence representation strategies that improve the accuracy of NN on duplicate question identification.
- Assess the ability of such strategies to improve the accuracy of random forest models.

Related Work

Quora Random Forest

This approach proposed by Quora involves a random forest model which makes classification based on a set of handcrafted feature, including the number of common words, the number of common topics labeled on the questions, and the part-of-speech tags of the words[1].

Siamese Network

A Siamese network consists of two bidirectional long short-term memory (BLSTM) networks as embedders. The network uses a concatenated embedding of the two sentences for classification[2].

Dataset and Preprocessing

Dataset

The dataset used is the Question Pairs Dataset from Quora. It contains 40,4351 question pairs. Each pair is labeled as either 1 (duplicate) or 0 (not duplicate). In the dataset, 25,5045 (63.1%) examples are duplicate.

Data Preprocessing

- Convert characters to lower case and tokenize them.
- Use GloVe (Wikipedia 2014 + Gigaword 5) as pretrained embedding.
- Switch question 1 and question 2 for data augmentation in selected methods.

Methods: Neural Network

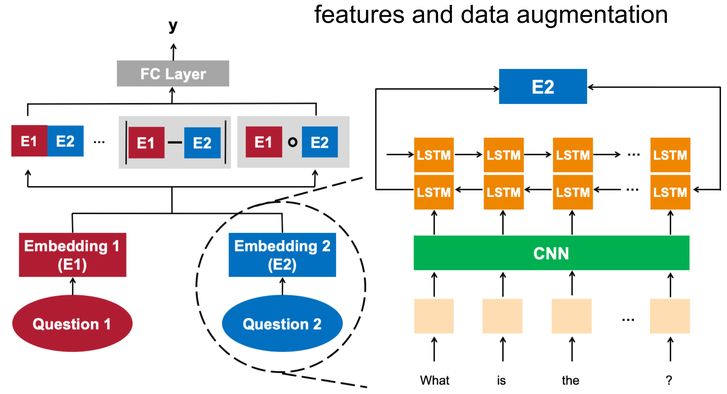
Baseline

- Uses a vanilla BiLSTM to exact sentence embedding.
- Classifies examples using a fully-connected (FC) layer.

Variations of Embedding Extraction

Explores four variations of feature representation techniques:

- BiLSTM + dist_angle
- CNN +dist_angle
- CNN + BiLSTM
- CNN + BiLSTM with more features and data augmentation



CNN + BiLSTM Network for Representation Extraction and Classification

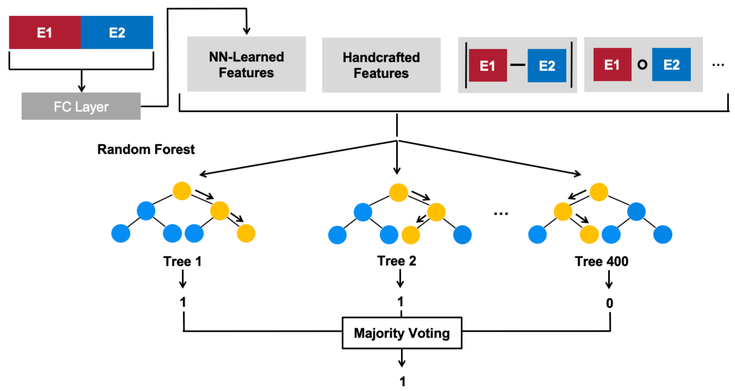
Methods: Random Forest

Baseline

- Uses mean GloVe word embedding and 29 handcrafted features.

Incorporating Learned Features

- Uses sentence embedding learned by CNN+BiLSTM model.
- Uses a FC layer to reduce the dimension of learned embedding.
- Adds features including embedding distance and angle.



Random Forest with Representation Extraced by Neural Networks

Results

	Accuracy	Precision	Recall	F1 Score	Accuracy Improvement
NN Baseline	0.8137	0.7686	0.7161	0.7414	
CNN + dist_angle	0.8524	0.8133	0.7795	0.7960	+4.76%
BiLSTM + dist_angle	0.8560	0.8024	0.8144	0.8084	+5.20%
CNN + BiLSTM	0.8651	0.8335	0.7978	0.8153	+6.32%
CNN + BiLSTM with more features and data augmentation	**0.8692**	0.8265	0.8062	0.8208	+6.82%
Random Forest Baseline	0.8185	0.7750	0.7233	0.7483	
+ NN Embedding	0.8414	0.7892	0.7843	0.7868	+2.80%
+ NN Embedding and dist_angle	**0.8597**	0.8141	0.8083	0.8112	+5.03%

- Both feature engineering and change in model architecture increase the accuracy of NN.
- Compared with structural change, feature engineering contributes more to the improvement of NN.
- Learned sentence representation significantly boosts the performance of random forests.

Conclusion

In this project, we utilize various strategies for sentence feature representation and show that feature engineering can improve both neural network and random forest classifiers. We find that learning an appropriate representation with respect to the task can be very helpful in classification as the network has more expressive power than pre-trained models.

Future Work

- Employ attention mechanism in our language model to better focus on parts of sentence associated with duplication.
- Apply the learned feature representation to other classifiers such as SVM and xgboost and test how their performance changes.

Reference

1. Lili Jiang, Shuo Chang, and Nikhil Dandekar. Semantic question matching with deep learning, Feb 2017. URL https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning.
2. Bromley, Jane, et al. "Signature verification using a" siamese" time delay neural network." *Advances in neural information processing systems*. 1994.