Quora Similar Question Detection
A Convolutional Neural Network Approach

Cheng Zhong
Graduate School of Arts and Sciences
Georgetown University
cz220@georgetown.edu

Abstract

Determine similar text on the internet has been a great focus for natural language processing area. It could be challenging to determine whether two questions are asking the same thing since there might have different words and sentence structure. As the biggest platform asking questions online, Quora provides a dataset of 400,000 labeled question pairs. This project will mainly focus on predicting whether a provided part of questions has the same meaning using deep learning and Nature languages processing techniques such as convolutional neural networks (CNNs), random forest, and multilayer perceptron (MLP). Our results show that CNNs outperform other two methods in terms of accuracy and loss.

1 Introduction

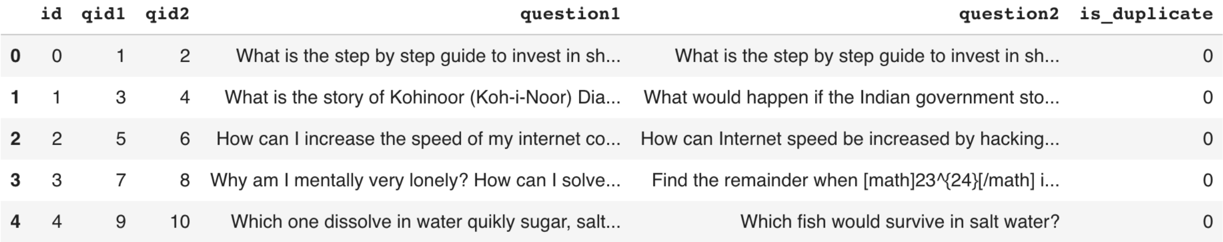
Quora is the biggest platform in the United States to ask questions and connect with people who contribute unique insights and quality answers. According to Quoras description, over 100 million people visit Quora every month, and there exists lots of people ask similarity worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question and make writers feel they need to answer multiple versions of the same question. Thus, by merging duplicate questions together into a single canonical question, there will have several benefits to increase user experience. For example, it saves time for people who want to ask the question if their question has already been answered previously on the site instead of waiting minutes or hours for a response. Also, duplicated questions can frustrate highly engaged users whose feeds become polluted with redundant questions. So, it is necessary to merge these similar questions. The Goal of this project will be predicting which of the provided pairs of questions contain two questions with the same meaning using deep learning and natural language processing techniques in order to help the Quora platform to have a greater user experience. The encoding methods explored include a Convolutional Neural Network (CNNs), a random forest model, and a multilayer perceptron (MLP) model.

2 Dataset

The Data of this project was offered by Quora through Kaggle. The dataset contains two CSV files. One is called train.csv and the other is called test.csv. The train.csv and test.csv dataset will have 6 columns of data including the following information:

- Id : the id of a training set question pair
- qid1, qid2 : unique ids of each question (only available in train.csv)
- question1, question2 : the full text of each question
- is_duplicate : the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

The training dataset has more than 400k records to train the model and we have over



	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

Figure 1: Exploratory Data Analysis

100k record as the test dataset to test the result. The ground truth is the set of labels that have been supplied by human experts. Usually, we define two questions are duplicates if the question expressed the same meaning or intent. In that case, a valid answer for one question will also be a valid answer to the other question. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty.

2.1 Exploratory Analysis (EDA)

The exploratory analysis shows the result that there is 36.92% of questions in the training set are duplicated. The total useable question pairs for training is 404290. In preprossessing data for this project, we removed question pairs that is exactly the same. More specifically, we removed question pairs whose combined word count across both questions was under 10 characters, as these questions might be nonsensical.

3 Related Work

It has been a long-standing problem in natural language processing to detect semantically equivalent sentences. According to Dey et al., traditional machine learning algorithms such as support vector machine (SVM) using hand-picked features and extensively preprocessed data perform well on the SemEval-2015 dataset.

What is more, deep learning techniques made several impressive signs of progress in recent years. More specifically, for the Quora dataset, as Nikhil demonstrated, by using an in-house deep architecture, the Long Short-Term Memory network (LSTM), which is a variant of Recurrent Neural Networks (RNNs). They achieved an accuracy near 87%.

Table 1: Accuracy Table

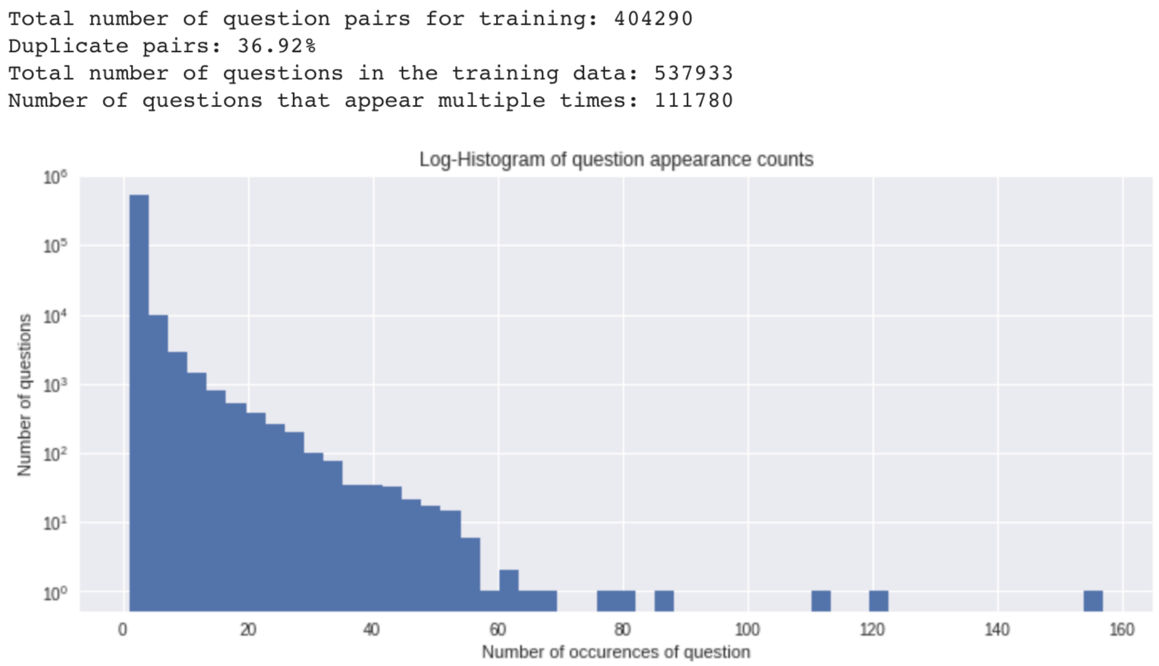
Metric	LSTM with concatenation	LSTM with distance and angle
Accuracy	0.87	0.87
Precision	0.88	0.83
Recall	0.86	0.94
F1 Score	0.87	0.88

4 Methods

This project applied several machine learning and deep learning method to classify the data and will use the log loss to evaluate the result. Since log-loss function heavily penalized classifiers that are confident about an incorrect classification. The loss score from Kaggle will evaluate the final result. The three main models used in this project is Random Forest, Multilayer Perceptrons (MLP), and Convolutional Neural Network (CNN).

4.1 Random Forest

The random forest model was set as the baseline model for this project. The reason that choosing random forest as the base model is random forest does not have a high requirement on the dataset. What is more, Random forest is the algorithm that Quora currently using to detect the same questions.



```
Total number of question pairs for training: 404290
Duplicate pairs: 36.92%
Total number of questions in the training data: 537933
Number of questions that appear multiple times: 111780
```

Figure 2: Sample Questions from the Dataset

4.2 Multiplayer Perceptrons (MLP)

The second model we used here is a simple feed forward neural network, a multilayer perceptron whose input to the concatenation layer are two vectors. Since the model is built in in the sklearn package. It would be simple to see how simple ANN works on the dataset.

4.3 Convolutional Neural Network (CNN)

This project will be using a CNN model trained from scratch. The CNN model contains four convolutional layers and four global average layers. With a final dense layer has size 512. Also, a dropout layer with dropout rate 0.2 was added to the model.

4.4 Assessment Metric

The evaluation metric in this project is log loss, which is the binary cross-entropy. The formula (1) is shown as below:

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j} log(p_{i,j}) \qquad (1)$$

5 Result

5.1 Random Forest

The log-loss shown on the Kaggle public scoreboard is 0.46317 on test data.

5.2 Multilayer Perceptrons (MLP)

The log-loss shown on the Kaggle public scoreboard is 0.47682 on test data. The loss is similar to the random forest which means the performance of MLP and random forest are similar.

5.3 Convolutional Neural Network (CNN)

The log-loss shown on the training set is 0.1886 after 5 epochs. For test data, the log-loss shown on the Kaggle public scoreboard is 0.38594 which is better than the other two models.

6 Discussion

From the above result, we can see that the CNN model definitely outperforms the random forest model and the MLP model. However, as we know that for text data, the normal approach would be Recurrent Neural Network (RNN) like LSTM. The reason we are

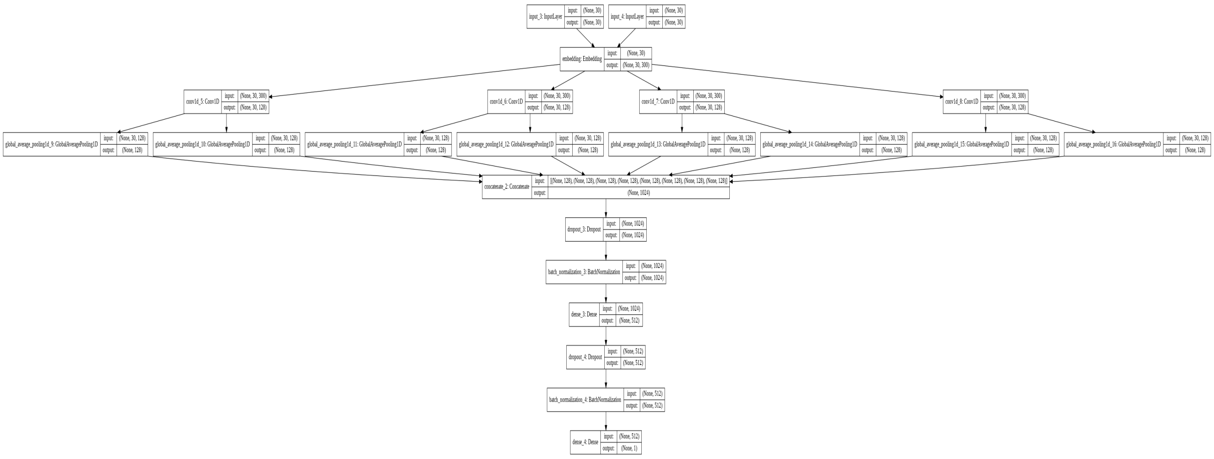


Figure 3: CNN Structure

```
Epoch 1/5
404290/404290 [==============================] - 235s 581us/step - loss: 0.4836 - acc: 0.7632
Epoch 2/5
404290/404290 [==============================] - 228s 563us/step - loss: 0.3656 - acc: 0.8320
Epoch 3/5
404290/404290 [==============================] - 227s 562us/step - loss: 0.2909 - acc: 0.8713
Epoch 4/5
404290/404290 [==============================] - 226s 560us/step - loss: 0.2350 - acc: 0.8992
Epoch 5/5
404290/404290 [==============================] - 226s 559us/step - loss: 0.1940 - acc: 0.9182
```

Figure 4: Train Result for CNN

Table 2: Result Comparison

Model	Private Test Loss	Public Test Loss
Random Forest	0.47697	0.47682
MLP	0.46331	0.46317
CNN	0.38988	0.38594

not choosing LSTM rather than CNN is that the CNN model will not overfit as much as LSTM. Although from the result, we can definitely find out there already exists some overfitting issue in the CNN model because the loss in the test set is three times larger than the training set. For LSTM, the result is even worse as the overfitting issue. What is more, computational power and time is another significant aspect we take into consideration. As we are using Keras as our framework and run it on Google Colab with the TESLA K80 GPU, the running time is around 220s/epoch for our CNN model as opposed to 400s/epoch for a 250-unit LSTM. Especially, in our CNN model, we choose to use the Global Average Pooling layer rather than Global Max Pooling layer is because it will make the model take less computation and run faster. Hence, we choose the CNN model in this project and it outperforms the random forest model Quora is currently using.

7 Future Work

There are three main issue exists for my current CNN model. The first problem is that the CNN model has no concept of word importance (such as TFIDF). Hence the model could not get the key point of a sentence. The second problem of the current CNN model is that It cannot distinguish sentence have same words but have different word order. For example, Quora Similar Question Detection is same as Question Similar Detection Quora. The last issue is obviously about the overfitting.

In order to solve these problems, future work including add dense layer to do the TFIDF and a more aggressive dropout layer can be done to solve the issue and might able to get a better result.

References

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2016. A paraphrase and semantic similarity detection system for user generated short-text content on microblogs. In COLING, volume 16, pages 2880 - 2890

Dandekar, Nikhil, et al. Semantic Question Matching with Deep Learning. Semantic Question Matching with Deep Learning, Quora, 14 Feb. 2017, engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning.