Duplicate Quora Questions Detection

Group - 9
K. Rohith Reddy
Puppala Deepika
S. Kumuda Priya

Problem Statement

- ★ Given two questions , we will have to predict whether they are duplicate of each other.
- ★ Example:
  - ” where is the Taj mahal located ?” and “What is the location of Taj mahal ?” Predict = 1
  - "Which one dissolve in water quickly sugar, salt, methane and carbon dioxide?" and "Which fish would survive in salt water?" Predict = 0

Introduction

- ★ Quora being visited by 100 million of people every month contains many similarly worded questions.
- ★ Detecting duplicate questions could help both seekers and writers in the community.
- ★ Quora uses a Random forest model to identify duplicate questions.

Experiments

Reference Paper:

https://pdfs.semanticscholar.org/4c19/2b8f45b1e913ee7da32624cd7559eccb0890.pdf

Abstract

- ★ Main idea is to perform vectorization and feature extractions of all the questions .
- ★ We then train and predict based on question vectors and features previously built.
- ★ We implement that in two different approaches
    - Vectorize words using Google's WORD2VEC model, construct Sentence2Vec model and build feature set for training classifier
    - LSTM(long short term memory) model

Data Acquisition and Dataset Description

- ★ Datasets downloaded from Quora Questions Pairs Competition, Kaggle.
- ★ Training dataset : 404,302 valid question pairs
- ★ Test dataset : 2,345,806 question pairs without qids and is_duplicate label
- ★ Features :
  - Id
  - Qid1
  - Qid2
  - Question1
  - Question2
  - is_duplicate

Sample Dataset:

Train

```
"id","qid1","qid2","question1","question2","is_duplicate"
"0","1","2","What is the step by step guide to invest in share market in india?","What is the step by step guide to invest in share market?","0"
"1","3","4","What is the story of Kohinoor (Koh-i-Noor) Diamond?","What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?","0"
"2","5","6","How can I increase the speed of my internet connection while using a VPN?","How can Internet speed be increased by hacking through DNS?","0"
"3","7","8","Why am I mentally very lonely? How can I solve it?","Find the remainder when [math]23^{24}[/math] is divided by 24,23?","0"
"4","9","10","Which one dissolve in water quikly sugar, salt, methane and carbon di oxide?","Which fish would survive in salt water?","0"
"5","11","12","Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?","I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?","1"
"6","13","14","Should I buy tiago?","What keeps childern active and far from phone and video games?","0"
"7","15","16","How can I be a good geologist?","What should I do to be a great geologist?","1"
"8","17","18","When do you use シ instead of し?","When do you use ""&"" instead of ""and""?","0"
"9","19","20","Motorola (company): Can I hack my Charter Motorolla DCX3400?","How do I hack Motorola DCX3400 for free internet?","0"
"10","21","22","Method to find separation of slits using fresnel biprism?","What are some of the things technicians can tell about the durability and reliability of Laptops and its components?","0"
"11","23","24","How do I read and find my YouTube comments?","How can I see all my Youtube comments?","1"
"12","25","26","What can make Physics easy to learn?","How can you make physics easy to learn?","1"
"13","27","28","What was your first sexual experience like?","What was your first sexual experience?","1"
```

Test

```
"test_id","question1","question2"
0,"How does the Surface Pro himself 4 compare with iPad Pro?","Why did Microsoft choose core m3 and not core i3 home Surface Pro 4?"
1,"Should I have a hair transplant at age 24? How much would it cost?","How much cost does hair transplant require?"
2,"What but is the best way to send money from China to the US?","What you send money to China?"
3,"Which food not emulsifiers?","What foods fibre?"
4,"How ""aberystwyth"" start reading?","How their can I start reading?"
5,"How are the two wheeler insurance from Bharti Axa insurance?","I admire I am considering of buying insurance from them"
6,"How can I reduce my belly fat through a diet?","How can I reduce my lower belly fat in one month?"
7,"By scrapping the 500 and 1000 rupee notes, how is RBI planning to fight against issue black money?","How will the recent move to declare 500 and 1000 denomination lewin illegal will curb black money?"
8,"What are the how best books of all time?","What are some of the military history books of all time?"
9,"After 12th years old boy and I had sex with a 12 years old girl, with her consent. Is there anything wrong?","Can a 14 old guy date a 12 year old girl?"
10,"What is the best slideshow app for Android?","What are the best app for android?"
11,"What services are from Google: Facebook, YouTube betray Twitter?","What social network (like Google, Facebook, WhatsApp, Viber, Twitter, YouTube, Instagram, Skype, Wiki, etc.) made huge impact on people and lifestyles?"
12,"What if a cricket hits a batsman's helmet and then goes to the boundary?","Should carbonated red balls and 8 yellow balls. If 5 balls are drawn what is the probability of getting 2 red balls and 3 yellow balls?"
13,"Just how do you learn fruity loops?","How do Fruity Wrappers work?"
```

Word2Vec Model

- ★ Word2Vec is a neural network that processes text.
- ★ Input: text corpus
- ★ Output: set of vector representations of each word in input
- ★ Available in gensim library
- ★ Performs well when analyzing semantics of single word but can’t be applied to sentence semantic analysis.

Without TF-IDF Weight

- ★ Sentence is represented as mean of 300-dimension word vectors which appeared in the word list along each dimension.
- ★ We compute cosine similarity between each question pairs and call it **“similarity_noidf”**
- ★ Drawback is that two questions with completely different words can collide.

With TF-IDF Weight

- ★ Sentence is represented as mean of word vectors obtained by assigning TF-IDF weights to each word vector before taking the mean.
- ★ We add 0.01 to IDF score, to overcome the problem if a word in two questions, the value becomes zero which would lead to zero similarity between identical sentences!
- ★ We compute cosine similarity between each question pairs and call it **“similarity_idf”**

- ★ We also consider two other features:
- ★ **“Com_perc”:** percent of common words in two questions.
- ★ **“Len_diff_perc” :** absolute difference in length of two questions divided by total length of questions.
- ★ If the Com_perc is very large and Len_diff_perc is rather small, the possibility of duplicate increases.

Classification models

K nearest neighbors

SVM

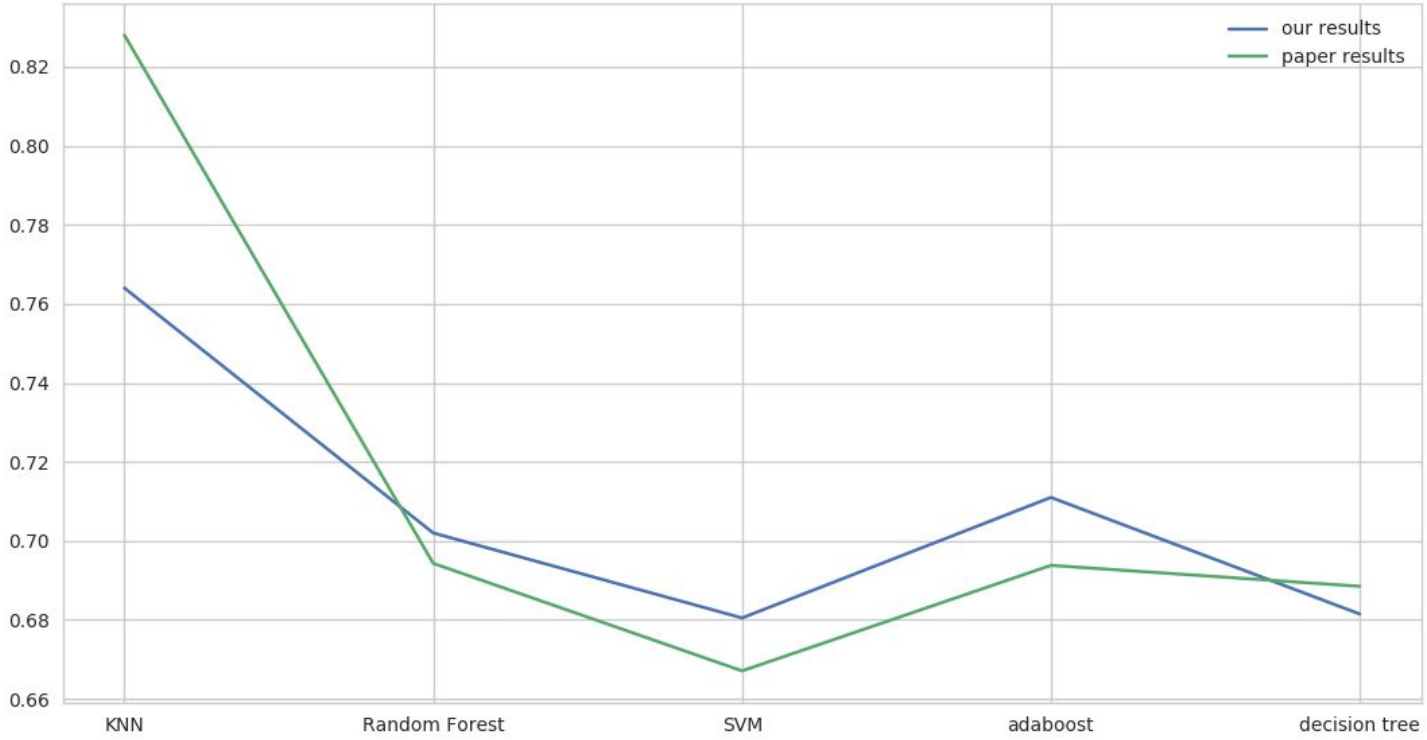
Decision Tree

Random Forest

AdaBoost

Choosing the Best model : Validation Accuracy.

Results


our results
paper results
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
0.66
KNN
Random Forest
SVM
adaboost
decision tree

References to LSTM(discussed in next slides):

Understanding LSTM : http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Using Keras LSTM inbuilt model :
https://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras/

Second Approach : LSTM Model

- ★ Well known model to work with sequential input
- ★ Model exhibits persistence.
- ★ Widely used in Nlp tasks like sentence completion.

Important Preprocessing Steps

- Each word is identified by unique id which reflects the order of frequency in dataset.
- Pad or truncate to make the length of sentences equal.

EG:- [[1,2,5,4,3] , [1,2,6,7,4,3]]

| Padding with 0 to make the len equal.

[[0,1,2,5,4,3] , [1,2,6,7,4,3]]

- Padding with 0 => Null word => mapped to all zeros in Word2Vec representation.

Model

Embedding layer :- It has word2vector representations of words matrix.

LSTM -- Sentence to Vector Conversion

- Input Nx(len of sentence = 25)X300 matrix
- output is Nx300

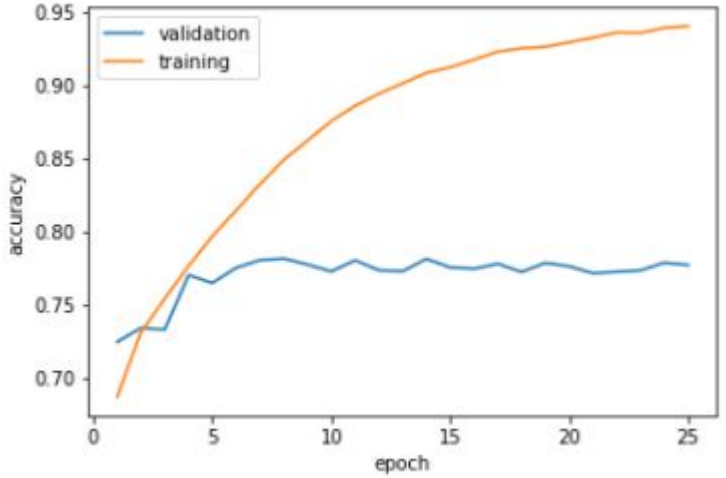
Concatenate layer

Output layer

- Sigmoid Activation function
- Binary cross entropy as Loss function

Results

```
Building model costs: 196.7135670185089
Training...
Train on 291088 samples, validate on 32344 samples
Epoch 1/1
Epoch 00000: val_loss improved from inf to 0.48071, saving model to ../da
9240s - loss: 0.5052 - acc: 0.7552 - val_loss: 0.4807 - val_acc: 0.7719
Training neural network costs: 9440.463582992554
predict...
80858/80858 [==============================] - 485s
Evaluation...
80858/80858 [==============================] - 473s
Test loss/accuracy final model = 0.4821, 0.7714
80858/80858 [==============================] - 468s
Test loss/accuracy best model = 0.4821, 0.7714
```



Train : Test Split = 0.8 : 0.2
Train : Validation Split = 0.9 : 0.1

	Paper	Experiment
Validation Accuracy	77.19	78.17
Test Accuracy	77.14	78.14

Improvement

References

Model Reference

https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning

Improved Model

Stack earlier LSTM model on top of Neural Netv

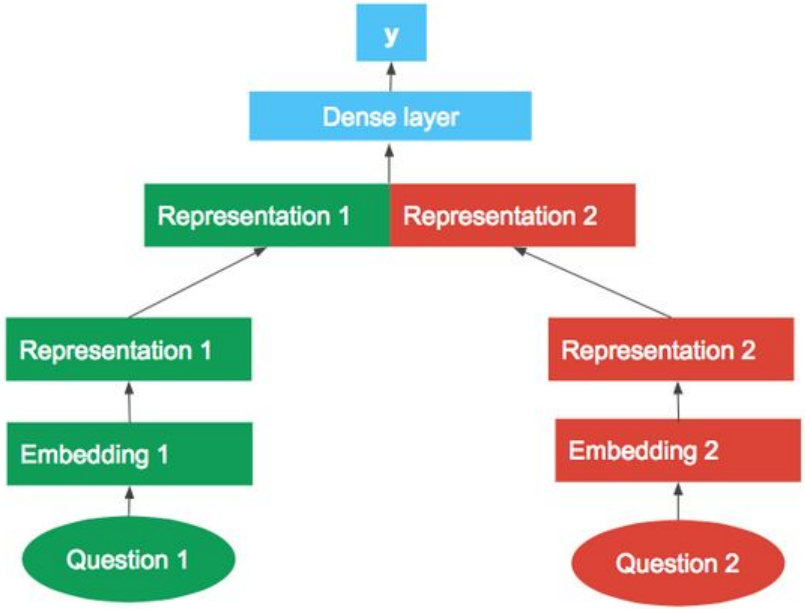
Added the following layer in between concaten
and output layer

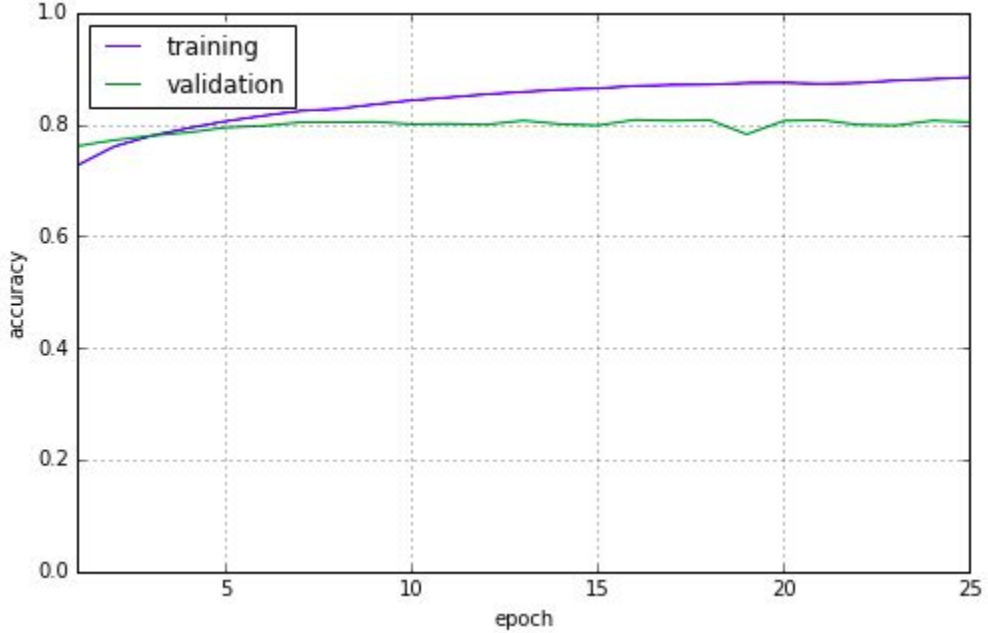
DENSE LAYER

DROP OUT LAYER

BATCH NORMALISATION

Above set of layers is added multiple times bef
the output layer.





Validation Accuracy	80.8
Test Accuracy	81.1

References

Code reference

https://github.com/aerdem4/kaggle-quora-dup

Keras: The Python Deep Learning library

https://keras.io

Model

- Preprocessing is done to remove different forms of writing .
  - .replace("won't", "will not").replace("cannot", "can not").replace("can't", "can not")
- We have 25 different features to train the model .
  - 4 features from question pairs
    - How many of them are numeric (union and intersection)
    - How many of the rare words are common (union and intersection)
  - 15 NLP features
    - len(q1_words.intersection(q2_words)) / (min(len(q1_words), len(q2_words)))
  - 6 NON-NLP features
    - Common neighbour
- We use keras.layers to perform Dropout , Dense , lstm , embedding
- We use these features to train using keras library(keras.model) and predict the output .
  - model.predict([test_data_1, test_data_2, features_test], batch_size=BATCH_SIZE, verbose=1)

Post processing

"Our original sampling method returned an imbalanced dataset with many more true examples of duplicate pairs than non-duplicates. Therefore, we supplemented the dataset with negative examples. One source of negative examples were pairs of "related questions" which, although pertaining to similar topics, are not truly semantically equivalent."

Reference - https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

- Having a non-duplicate common neighbor does not mean that these questions are not duplicates . so, we update the index predicted
  - DUP_UPPER_BOUND = 0.98 # do not update dup probabilities above this threshold
    - if dup_neighbor_count > 0 and test_label[index] < DUP_UPPER_BOUND:

      update = min(MAX_UPDATE, (DUP_UPPER_BOUND - test_label[index])/2)

      test_label[index] += update
  - NOT_DUP_LOWER_BOUND = 0.01 # do not update dup probabilities below this threshold
    - if dup_neighbor_count > 0 and test_label[index] > NOT_DUP_LOWER_BOUND:

      update = min(MAX_UPDATE, (test_label[index] - NOT_DUP_LOWER_BOUND)/2)

      test_label[index] -= update

Model Properties

- All the features are question order independent.
  - For example, instead of using question1_frequency and question2_frequency, we use min_frequency and max_frequency.
- Feature values are bounded when necessary.
  - df["common_neighbor_count"] = common_nc.apply(lambda x: min(x, NEIGHBOR_UPPER_BOUND)) , number of neighbors are set to 5 for everything above 5.
- Good preprocessing is one of the factor for good performance of the model
- Replacing the rare words with a placeholder .

Results

Validation Accuracy : 84%

Test Accuracy : 83.5%

Thank you