De-duplication in Online Discussion Forums

Team 3: Saurabh Aggarwal, Siddharth Sen

May 1, 2022

1 Introduction

Discussion forums like Quora and StackOverflow are experiencing more and more traffic as time passes by. Both increase in the number of users and new things to query about/ discuss are contributing factors. This increase leads to more questions/ answers on these forums, making it difficult to find what you are looking for along with requiring more data and computation resources to manage efficiently. De-duplication of questions will help consolidate information from all similar questions to one thread. This will result in the user having to spend less time to find a solution, less storage and computation on the backend. Enabling fewer questions overall and better discussions on every question.

From our analysis on our test set (containing 80k questions), we were able to identify 12.7k duplicate questions. Removing these questions would lead to a **16%** decrease in the amount of storage space required.

This reduction will aid in speeding up query search, reduce data storage required, and finally, improve user experience.

2 Background

Since the problem involves capturing the inherent "meaning" of two given pieces of text, instead of directly comparing words or tokens and measuring their similarity, we aim to use a recurrent neural network framework in the form of Siamese-Manhattan LSTM networks.

2.1 LSTM Networks

LSTM (Long Short-Term Memory) networks, which are a form of recurrent neural networks (RNNs), are designed specifically to overcome the long-term dependency problem faced by RNNs (due to the vanishing gradient problem). This enables LSTMs to process entire sequences of data without treating each point in the sequence independently, while retaining useful information about previous data in the sequence to help with the processing of new data points. As a result, LSTMs are particularly good at processing sequence data such as text, speech, and time-series data. The output of an LSTM at a particular point in time is dependent on three things:

- Cell State – the current long-term memory of the network

- Hidden State – the output at the previous point in time
- Input Data – at the current time step

LSTMs use a series of gates which control how the information in a sequence of data comes into, is stored in, and leaves the network. There are three gates in a typical LSTM; forget gate, input gate and output gate. These gates can be thought of as filters and are each their own neural network. This combination of gates comprises an LSTM cell, as visualized in the following diagram. Instead of coding the LSTM network by hand, we shall leverage an existing python deep learning library, Keras (built on top of TensorFlow) for the purposes of modelling our data. [1]

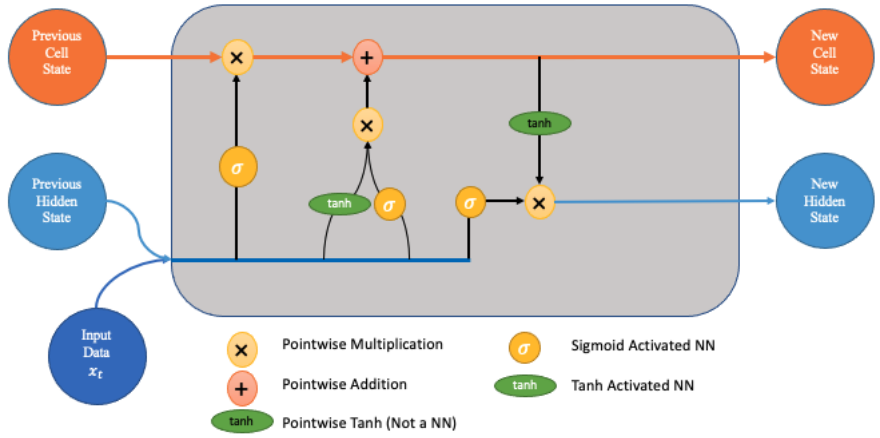

Figure 1: LSTM Cell Diagram

2.2 Siamese Neural Networks

A Siamese Network is a class of neural networks that contain two or more identical sub-networks within them. The networks are identical in the sense that they have the same configuration with the same parameters and weights. Thus, parameter updating is mirrored across both sub-networks and is particularly used to find the similarity of the inputs by comparing its feature vectors, so these networks are used in many applications.

Siamese neural networks are used to tackle one-shot learning problems, where the goal is to identify whether two inputs are similar or not, as compared to traditional classification tasks where you assign a class to each input. Therefore, Siamese networks seem to perform well on similarity and comparison use-cases and are readily used for tasks like sentence semantic similarity, recognizing forged signatures and many more. [2]

2.3 Manhattan Distance

Manhattan distance is a distance metric between two points in a N-dimensional vector space. It is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes. It is a specific case of Minkowski distance (p=1) and is also referred to as L1 norm.

Since our task is that of measuring sentence (question) similarity, a Siamese network is well-suited for our application. We shall combine the use of Siamese networks with LSTMs and use Manhattan distance in conjunction with the neural network architecture in order measure similarity.

3 Data Collection

We used the openly available data from a 2017 Kaggle Competition on Sentence Similarity. The dataset given had $\tilde{4}$00k Quora Questions Pairs (800k questions). The dataset contains rows with pairs of questions (question header visible on Quora) along with an IsDuplicate boolean attribute for the training data, which indicates whether or not the two questions are similar (duplicates) or not. [3]

question1	question2	is_duplicate
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
What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-Noo	0
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
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
Which one dissolve in water quikly sugar, salt, methane and carbon di oxide	Which fish would survive in salt water?	0
Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this	1
Should I buy tiago?	What keeps childern active and far from phone and video games?	0
How can I be a good geologist?	What should I do to be a great geologist?	1

Figure 2: Quora Question Pairs Dataset

4 Analysis

Given we had textual data we approached the problem starting with typical textual prepossessing techniques like:

- Stopword Removal
- Stemming
- Lemmatization
- Tokenization

Next step is to convert text token to mathematical vectors, suitable to be given as input to a Neural Network model. For this purpose we have used the Word2vec embeddings (by Google) [4]. These embedding help us capture contextual meaning of the words in addition to just there presence in the dataset (which is captured by bag-of-words and tfIdf methods).

For modelling purposes we have used a Siamese Neural Network with LSTM and Manhattan Distance for similarity. The architecture of the model is shown below:

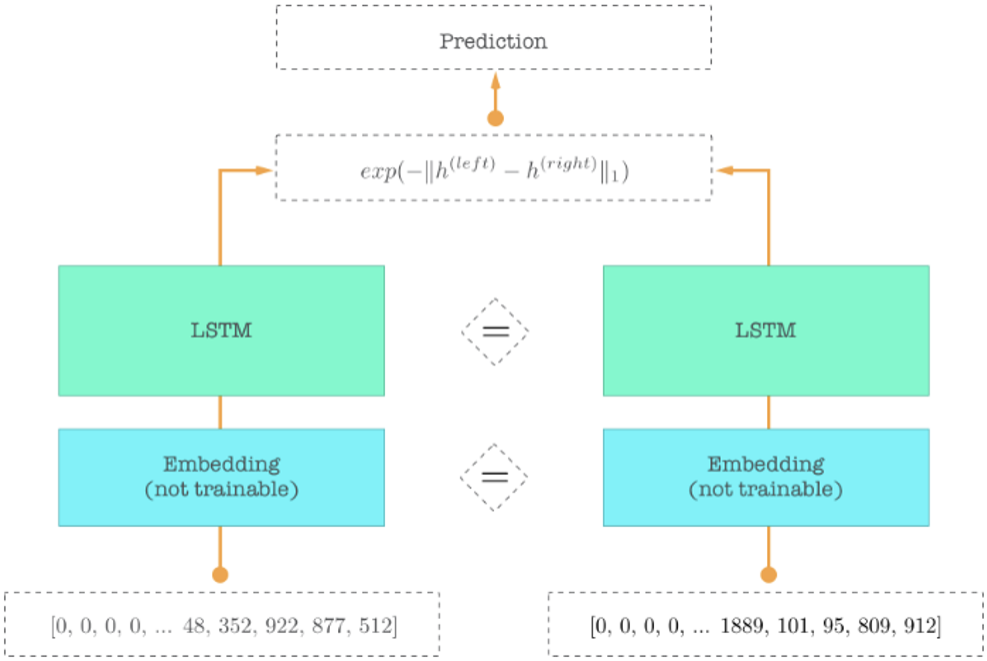

Figure 3: Siamese LSTM Network

We used cross validation for Hyperparameter tuning over a small set of hyperparameter combinations (limited compute power). We ran our models for few epochs only, again because of compute limitations. The best performing model had the following Hyperparameters:

Model Details

Optimizer: Adam (cross-entropy loss)

batch_size: 64

units in LSTM: 50

epochs: 10 (Hardware limitations)

learning rate: 0.001

Figure 4: Model Hyperparameters

We achieved an accuracy of 83.48% on our test set. The confusion matrix of our predictions is shown below:

	Pred 0	Pred 1
True 0	22914	2238
True 1	4370	10478

Figure 5: Confusion Matrix

From the confusion matrix we can observe that our model had a precision of 0.824 and a recall of 0.7, which means that whenever our classifier claims that the questions are similar it is correct 82.4% of the time and it correctly captures 70% of all the similar questions in out dataset.

The loss and accuracy plots of our model are shown below:

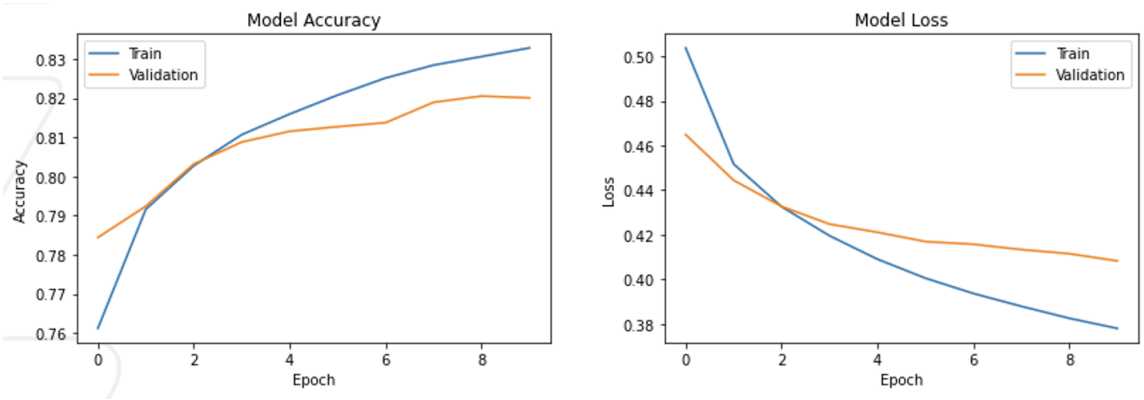

Figure 6: Model Accuracy and Loss Curves

From these plots we can see that both accuracy and loss have not reached an asymptotic state, which means that the model can still get better at predicting pairs if it is run for more epochs.

4.1 Future Work

We have identified a few tasks for future work:

- Running the model for longer till the loss stabilizes.
- Identify chains of similar questions (in addition to just pairs) – by permuting dataset.
- Real-Time identification of similar question (If a user is typing a new question, we can search if any similar question has already been asked (and answered) and refer the user to that post)

5 Findings

We ran our models on a subset of our data (40K questions pairs), we were able to identify 12.7k pairs out of the 40k. This is not considering similarity in questions across pairs (which can be captured by permuting the dataset, and would result in increasing the similar pairs even further).

Removing one question in each pair would lead to a **16%** (12.7k out of 80k) decrease in the amount of storage space required.

6 Conclusion

We have observed a **16%** compression just within our dataset. The actual databases used by Quora contain questions data in the order of tens of terabytes. If we observe a similar rate of compression for Quora's databases, it would imply a drastic compression in their storage space requirements further reducing their cost for storage, speeding up search, and improving user experience for users on their website.

References

[1] A Detailed Explanation of LSTM Networks, https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9

[2] Siamese Neural Networks, https://towardsdatascience.com/a-friendly-introduction-to-siamese-networks-85ab17522942

[3] Quora Question Pairs Dataset, https://www.kaggle.com/competitions/quora-question-pairs/data

[4] Word2Vec: Google, https://code.google.com/archive/p/word2vec/