Quora Question Similarity Using BERT Embeddings with Siamese Networks as Feature Extractors

Ahmad Nayyar and Ganesh Rajasekar

Custom Project- MSCI641 Electrical and Computer Engineering,University of Waterloo

Problem Description

The main task of this project is to identify semantically similar pair of questions as multiple questions with same context and intent can cause major confusion in the Quora discussion forum. In doing so, the seekers would not have to go through the hassle of searching the best answer among many and writers would not have to write variant versions of the similar content. Quora currently uses Random Forest to identify similar question but with the success of deep networks it is highly probable that they can surpass the previously set benchmarks.

Dataset and Preprocessing

The data set contains 404351 question pair along with the label identifying each example as either positive - duplicates or negative – non duplicates. There are 255045 negative (non-duplicate) and 149306 positive (duplicate) instances. For our analysis we overcome this class imbalance by analyzing the model performance in a balanced distribution

Model Design and Architecture

a) Input Layer : Sentence Embeddings with different Word Embeddings

Why Required?

- Input required as a fixed length feature
- Bag of Words does not capture semantics
- Convert Pair of Questions to Vectors using Embeddings:
  - **Word Embeddings used:**
    - **Word2Vec** (Predictive - using n-grams)
    - **Glove** (Count based – using co-occurrence matrix)
    - **FastText** (Faster – using Huffman algorithm while storing categories in the form of trees)
    - **BERT** (Bi-directional -using Transformer's Encoder) {used for feature engineering only}
  - **Sentence Embeddings used:**
    - **Sent2Vec** (Unsupervised embeddings) with Word2Vec (E1)
    - **Sent2Vec** (Unsupervised embeddings) with Glove (E2)
    - **InferSent** (Supervised embeddings) with FastText (E3)

b) Feature Extraction Layer : Siamese Architectures with Sentence Features

What are Siamese Networks?

- Two or more identical subnetworks
- Find similarity or relationship between two comparable things(Question 1 & 2)
- **Siamese Architectures used:**
  - 3 channel CNN Siamese (channel 1 with E1 embeddings , channel 2 with E2 embeddings , channel 3 with E3 embeddings)
  - DNN Siamese 1 (with E1 embeddings)
  - DNN Siamese 2 (with E2 embeddings)
  - DNN Siamese 3 (with E3 embeddings)

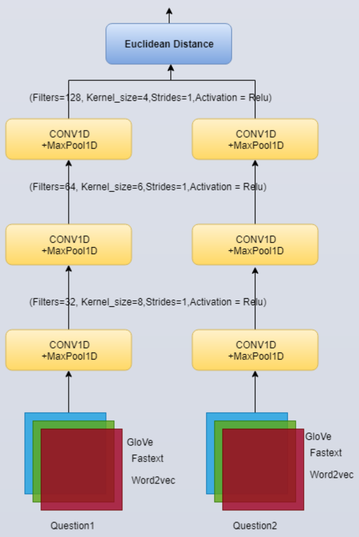


Figure1(a). CNN Siamese Architecture with embeddings as three Channels

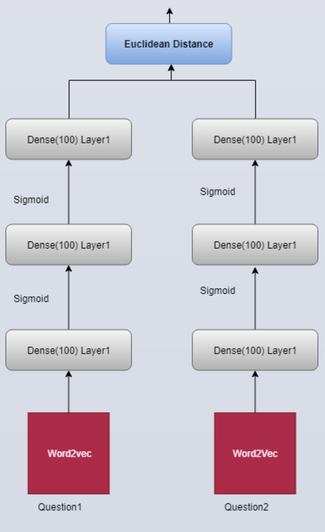


Figure1(b). Dense Layer Siamese Architecture for Individual Embeddings

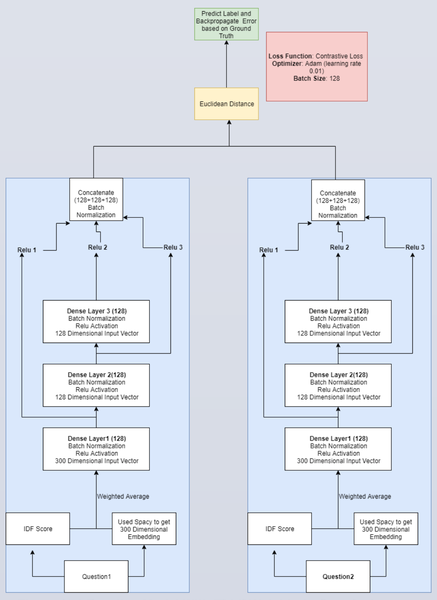


Figure 1(c) Baseline Model With Tf-idf Weighted w2vec features

c) Sentence Features:

- BERT [CLS] token (Stores the entire representation of question pair)
- Fuzzy, similarity(distance, count based) features using BERT embeddings of question1 & question 2 separately

d) Concatenation Layer:

- Combined output of Siamese Architecture and Sentence Features

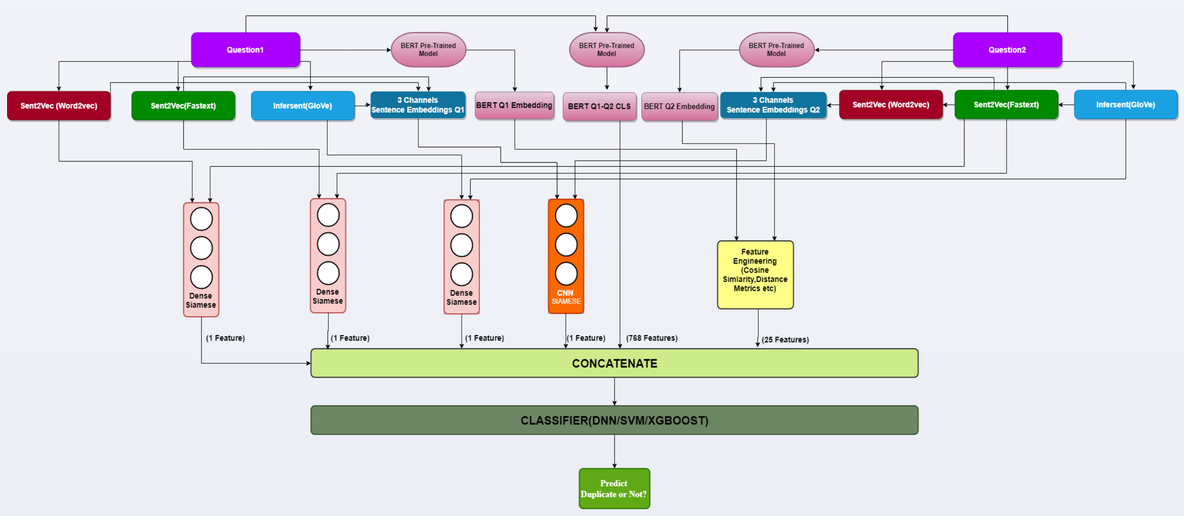
e) Classification Layer:

- Predict whether the question pairs are duplicate or not
- DNN Layer / Classical ML{SVM,XgBoost} used

Base Line Architecture :

- DNN Siamese Architecture with simple tf-idf weighted w2vec sentence representation used without Feature extraction
- **Results in an accuracy of only 71.4%.**

Complete Overview



Results

Model	Training Accuracy	Training Loss	Testing Accuracy	Test Loss
3-CNN+ 3SiaM+BF+DNN	0.8517	0.502	0.7749	0.6113
3-CNN+ 3SiaM+BF+ CLS+DNN	**0.8185**	**0.544**	**0.8018**	**0.5678**
4-CNN+ 4SiaM+BF+DNN	0.8005	0.5948	0.7738	0.6423
4-CNN+ 4SiaM+BF+ CLS+DNN	0.8041	0.4834	0.7961	0.4996
3-CNN+ 3SiaM+BF+ CLS (PCA)+DNN	0.8238	0.4999	0.7876	0.5385
4-CNN+ 4SiaM+BF+ CLS (PCA)+DNN	0.8279	0.547	0.8004	0.6137
Baseline Model	0.7904	0.1717	0.7174	0.206

Table1. Accuracy Results with DNN as Classifcation Layer

Model	Training Accuracy	Training Loss	Testing Accuracy	Test Loss
Baseline Model	0.7904	0.1717	0.7174	0.206
3-CNN+ 3SiaM+BF+ CLS+DNN	**0.8185**	**0.5440**	**0.8018**	**0.5678**
3-CNN+ 3SiaM+BF+ CLS+SVM	0.9276	0.6541	0.7042	0.7121
1-CNN+ 3SiaM+BF+ CLS+XGBOOST	0.8386	0.5622	0.7742	0.6342

Table2. Accuracy Results by Varying Classifcation Layer

- Compared baseline model with different combinations of features ,model architecture variations and classification algorithms {DNN,SVM, Xgboost}
- Out of the plethora of combinations, configuration with 3 channel CNN Siamese + 3 DNN Siamese + BERT Features + BERT [CLS] + DNN classifier results in best accuracy of 80.81 - a significant increase of about 10% from the baseline

Ablation Study

Ablation Studies on Best Model	Training Accuracy	Training Loss	Testing Accuracy	Test Loss
Varying Siamese Layers with DNN:				
3-CNN+ 3SiaM+BF+ CLS	**0.8185**	**0.544**	**0.8018**	**0.5678**
3-CNN+ 2SiaM+BF+ CLS	0.8531	0.4892	0.7911	0.5617
3-CNN+ 1SiaM+BF+ CLS	0.8552	0.4741	0.7877	0.5594
3-CNN+BF+ CLS	0.8245	0.5239	0.7812	0.5687
Varying CNN Layers with DNN:				
2-CNN+ 3SiaM+BF+ CLS	0.8120	0.4927	0.7761	0.5097
1-CNN+ 3SiaM+BF+ CLS	0.8071	0.4936	0.7670	0.5110
3SiaM+BF+ CLS	0.7991	0.4471	0.7482	0.4598
Using LSTM Siamese with DNN:				
3-CNN+ 3 LSTM SiaM+BF+ CLS	0.8123	0.586	0.7815	0.6012

Table3. Ablation Study Results on the Best Model

- Assess the importance of every component in the optimal model by removing each part individually and testing the results
- Each component in the optimal model has significance as removing each part results in reduction of the overall accuracy

CONCLUSION

Deep Learning architectures indeed improve the performance as compared to Random Forests used by Quora previously to tackle the problem

REFERENCES

[1] Ameya Godbole, Aman Dalmia, and Sunil Kumar Sahu. Siamese neural networks with random forest for detecting duplicate question pairs. arXiv preprint arXiv:1801.07288, 2018.

[2] Yushi Homma, Stuart Sy, and Christopher Yeh. Detecting duplicate questions with deep learning.

[3] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[4] Quora Question Pair Duplicate Feature Engineering By Abhishek Thakur https://www.linkedin.com/pulse/duplicate-quora-question-abhishek-thakur/