Quora Question Pairs

The State of That Poster

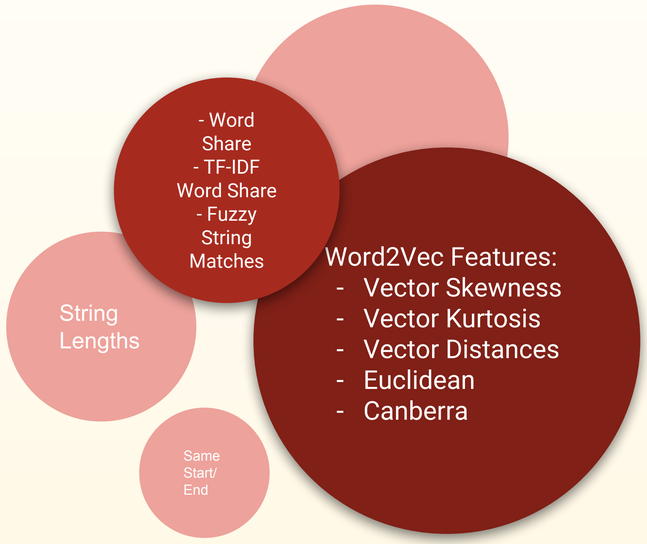
Dimitris Michailidis, Ben Wilmers

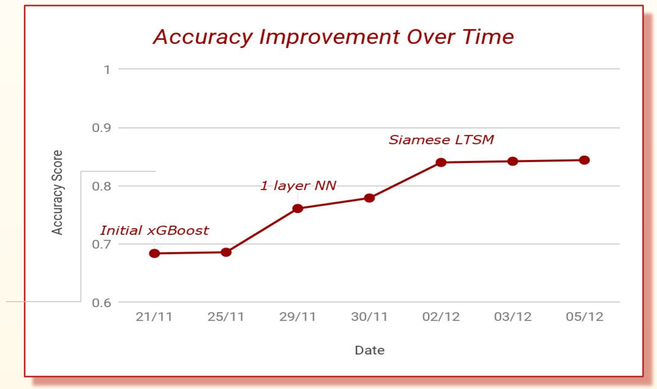
Problem Statement

Our goal was to build a model which could identify whether 2 questions on Quora were duplicates or not.

We oriented our approach around using Deep Learning methods.

Our final model is a combination of work done by [1], [2], [3].

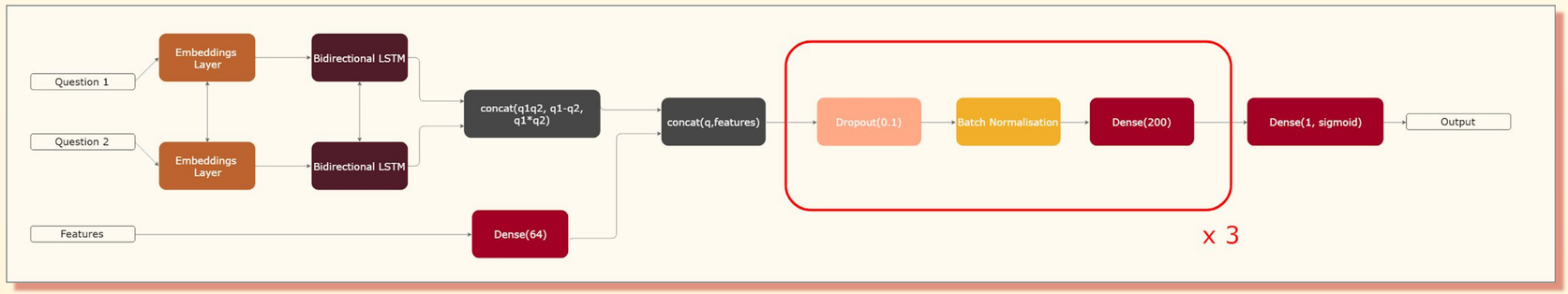




Summary

- GloVe Embeddings [100, **300**]
- Dropout rates [**0.1**, 0.2, 0.5]
- Sequence Length [20, **25**]
- LSTM Units [**256**]
- Features Dense [**64**, 100, 200]
- Batch Size [128, **256**, 1080, 2048]
- Fully Connected Layers [200]

Next: Try stacking different outputs together into a new model.



Text Cleaning

Clean Text

Can anyone increase their height at the age of 21 or 21?

Remove special character, hyperlinks, correct spelling

Remove Stopwords

Can anyone increase height age 21 21?

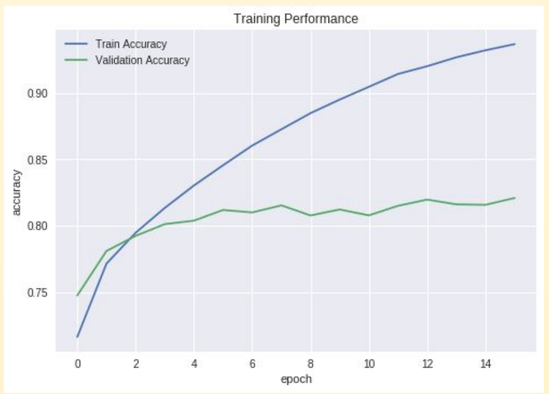
Get rid of common English words with no semantic significance

Stem Words

Can anyon increas height ag 21 21?

Reduce words to their base, root form

Final Model performance



Models we tried

84.4%	Bidirectional LSTM Features Dense Bowman Concat 2 Dense & BN	• Validation Loss: 0.444 • Validation Accuracy: 0.946
83.2%	Bidirectional LSTM Bowman Concat 2 Dense & BN	• Validation Loss: 0.394 • Validation Accuracy: 0.841
83.4%	Siamese LSTM Bowman Concatenate 2 Dense	• Validation Loss: 0.390 • Validation Accuracy: 0.839
83.1%	Siamese LSTM Concat 2 Dense	• Validation Loss: 0.383 • Validation Accuracy: 0.834
82.6%	Siamese LSTM Concat Dense	• Validation Loss: 0.394 • Validation Accuracy:0.827
77.9%	Simple Embeddings Flatten Layer	• Validation Loss: 0.510 • Validation Accuracy: 0.767

[1] Bowman, Samuel R., et al. "A fast unified model for parsing and sentence understanding." (2016)
[2] Bowman, Samuel R., et al. "A large annotated corpus for learning natural language inference." (2015)
[3] Addair, Travis. "Duplicate Question Pair Detection with Deep Learning."