Identify Duplicated Question Pairs for Quora

Yi Ning Liang, Zheya Feng

Code Our Way Out Of It

1 Goal

Classify each question pair as 1(duplicated) or 0(not duplicated)

2 Dataset

323164 question pairs for training, 81126 for testing

3 Text Preprocessing

Lemmatize words	e.g. ate ›› eat(v), men ›› man(n)
Replace special characters	e.g. 000,000 ›› m, % ›› percent
Replace abbreviation	e.g. won't ›› will not, what's ›› what is

4 Feature Engineering

30 Features	Common words/token feature， Similarity， Tfidf word share feature, Fuzzy feature， Longest substring/sequence characters/words ratio, Length of rare words, numbers intersection and union...

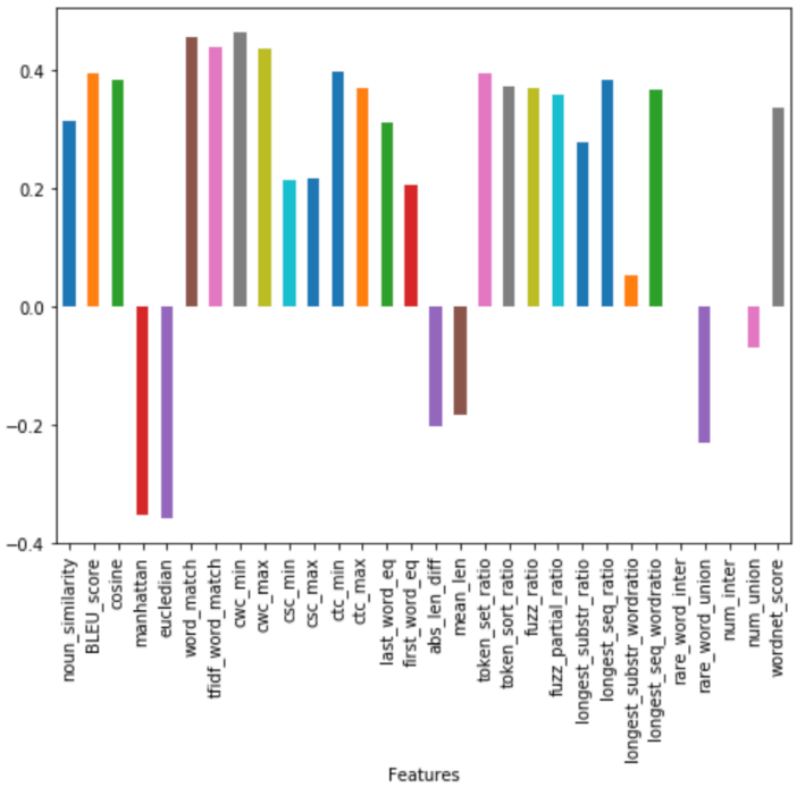


Fig.1 Feature-isDuplicated correlation

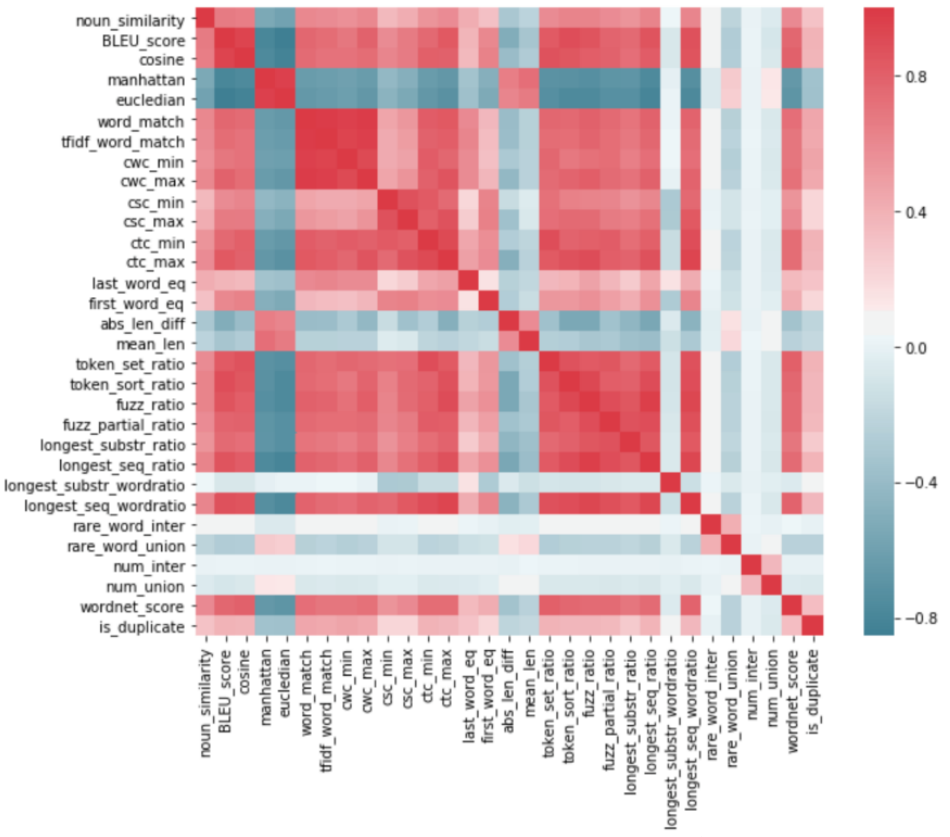


Fig.2 Feature-Feature correlation

Identify Duplicated Question Pairs for Quora

Yi Ning Liang, Zheya Feng

5 Model Comparison

We tried different algorithms on the 30 features, LSTM gained higher accuracy in our first attempt, so we only continued designing the LTSM structure and didn't further investigate the hyperparameters in other approach.

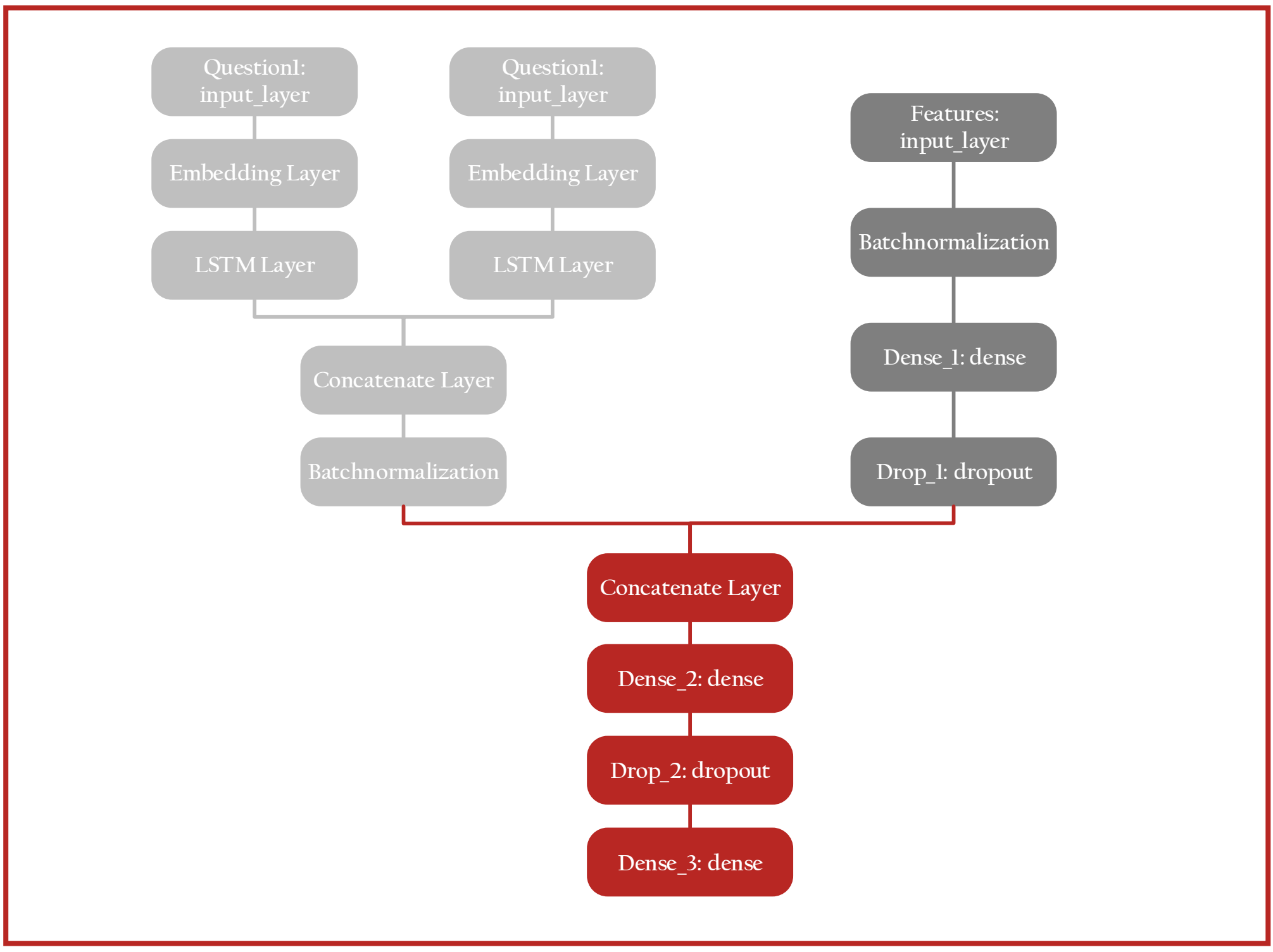
LSTM
Accuracy:0.85
(Final submission)

Xgboost
Accuracy:0.79

Logistic regression
Accuracy: 0.77

❖ *LSTM: 3 Dense layers, 2 Drop out layers (20%)*

6 LSTM Network Structure



7 Model Performance

Model Evaluation	3-fold cross validation Epoch: 5 Train on 215,442 samples Validate on 107,722 samples Kaggle score : **0.85474**

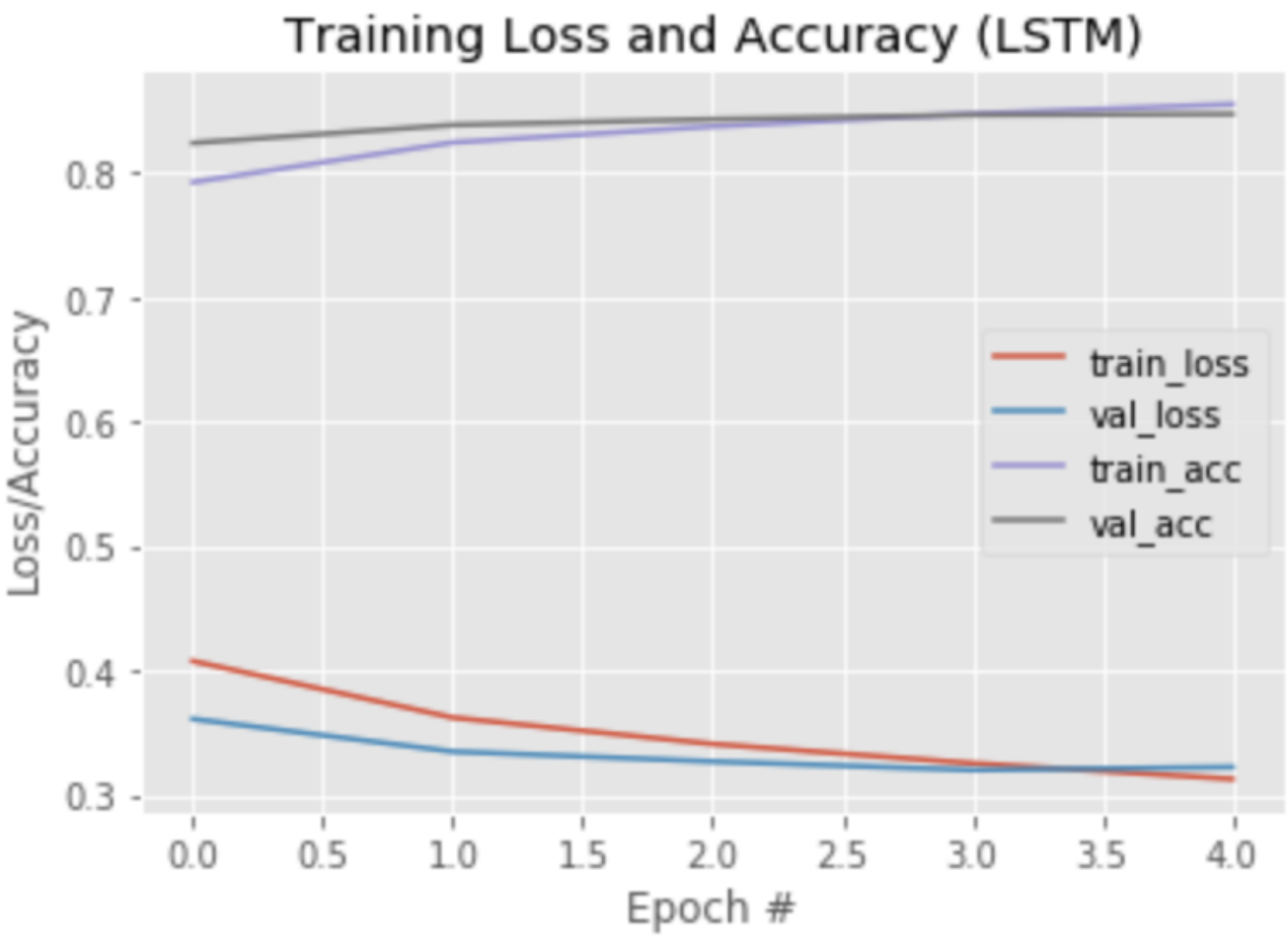


Fig.3. Epoch #-Loss/Accuracy

8 Reflection

- The 0.88 accuracy score we got on Kaggle is because we added 4 features that combine train data and test data. In the end we decided not to do that and only stick with 30 features.

- Due to the limitation of time and computational power, we only trained the model with 3-fold cross validation and with 5 epochs each.

- Imbalance data problem can be further investigated (Not_duplicate = Is_duplicate*1.7)