CS 2470 Final Project

Semantic Textual Similarity using Dense CNN: Automated Response to User Questions

Ashish Rawat, Gurnaaz Kaur

Fall 2018

1 Introduction

Our project is inspired by a sub-problem that we were trying to solve as part of a product that helps international students in the process of applying to universities in the US. The process of university applications and migrating to a new country has a lot of unknowns and therefore, students will have a lot of questions in the mind regarding a number of processes. To help them with this, our product uses a chat bot window where users can ask us a question that will be answered by expert knowledge that has been accumulated over many years. The first problem in this approach is that there can be a large number of questions, but most of these questions are similar or duplicates of one another. Therefore, there is a small and unique set of questions in the domain of university applications. Any question that the students is a subset of these unique base questions. So, given a question, we hope to find a semantically similar question in our unique question set and provide automated answers to these questions without any human intervention.

2 Related Work

Our problem statement is very similar to the problem faced by Quora where they try to eliminate questions which are duplicates of one another. So we found the following paper by Yichen Gong titled "Natural Language Inference over Interaction Space", that achieves state of the art accuracy in the task of semantic textual similarity. The paper describes an end to end deep learning network that takes in two questions, where one is the premise and the other is the hypothesis. The output of the network is one of the labels between contradiction, entailment and neutral. Essentially, what the network does is that it accepts the premise and the hypothesis sentences and

tells us whether the hypothesis is an entailment of the premise or is a contradiction of the premise or is neither an entailment or a contradiction of the premise. The model described in the paper uses dense feature vectors created by concatenating word embedding, character embedding, parts of speech (POS) embedding for both questions. These are encoded and transformed into a three dimensional matrix that represents rich features of semantic similarity between the two questions. The model then uses dense convolution layers to extract these features and make a prediction.

3 Dataset

Quora Question Pairs: One of the datasets used by the paper was Quora question pairs which has over 400000 pairs of questions which have some semantically similar and dissimilar question pairs. The paper achieved state of the art performance on this dataset and was able to predict the correct label with an accuracy of 88%.

Question 1	Question 1	Label
How can I be a good geologist?	What can I do to become a good geologist?	1
What is the best way to make money online?	What is the best way to ask for money online?	0
How can I read and find my YouTube comments?	How can I see all my YouTube comments?	1
What's the one thing you would like to do better?	What's the one thing you do despite knowing better?	0

Figure 1: Quora Question Pairs.

Our Dataset- Student Questions: Over time, we scraped data from various admission forums and admission groups and that data was just a collection of questions that had been asked by students regarding the application process. However, all we had was a long list of questions that the students had asked, so manually labelling the questions would have required a lot of effort. So, we created 5000 labelled question pairs manually by grouping questions into different categories. The negative data points were question pairs from two different categories.

To make development and training simpler and feasible, we used only a subset of the Quora dataset supplemented by our own question pairs amounting to 100,000 question pairs.

Question 1	Question 2	Label
If a person doesn't have a paper published, would that make his or her chances of an admit very low?	Is it an unsaid rule to have a research paper to get into a good university?	1
Which is better for MS in CS? NCSU or Colorado Boulder?	Is NCSU a good university to study Computer Science?	0
Can I apply for US visa without paying the SEVIS fee?	Is a SEVIS account required before applying for US student visa?	1

Figure 2: Student Question Pairs.

4 Model

4.1 Network Architecture

The paper describes a network that creates a feature rich encoding of the two questions. The two encoded matrices are then combined together by means of a dot product into a three dimensional matrix. This 3D matrix contains rich contextual information of the semantic similarity of the two questions. The model then uses 3D convolution by means of a DenseNet architecture to find meaningful features from the interaction matrix. In other words, the 3D interaction matrix is down sampled using 3D convolution resulting in a 1D array of features. These features are then connected to a fully connected output layer which has three neurons allowing classification into three classes namely Contradiction, Entailment and Neutral.

In our implementation, we fused Contradiction and Neutral into one class called Non-Entailment, modifying the above model architecture into a binary classification model.

- **Embedding Layer**: Produces an embedding vector per word that consists of a pre-trained word embedding, character-level representation of the word and word level features. (Essentially, created representation matrices for sentences). The network uses a dense feature representation for each sentence which is a 2D matrix which is a concatenation of word embedding, character embedding, POS tagging and exact match vector. This rich feature representation is instrumental in achieving state of the art accuracy.
- **Encoding Layer**: The encoding layer applies self attention on the words in the sentence. The original paper applies self-attention on the word in the sentences and uses bidirectional LSTM in order to encode the embedding layer.

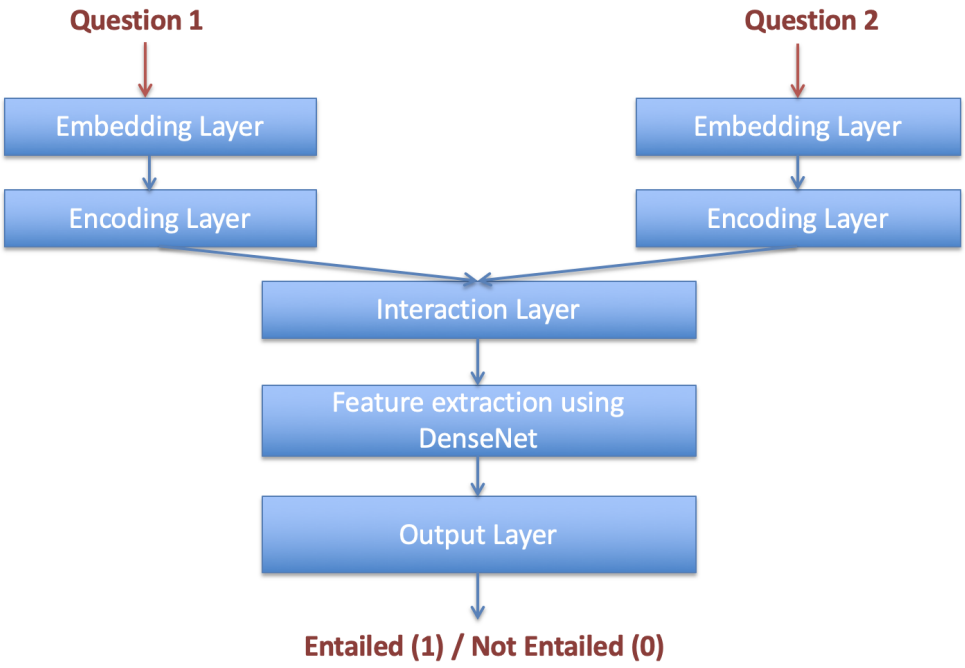

Figure 3: Dense Interactive Inference Network. Accepts two questions as input and outputs whether the questions are similar or not.

Since this part of the model was very tricky to replicate, we instead ended up using a pre-trained LSTM to encode the embedding layer and produce a 2D encoding matrix. This is partly the reason why our model performs weakly as compared to the original implementation.

- **Interaction Layer**: This layer creates an word-by-word interaction vector by both premise and hypothesis representation matrix. The interaction can be modeled in different ways. A common approach is to compute the cosine similarity or dot product between each pair of feature vector. On the other hand, a high-order interaction tensor can be constructed with the outer product between two matrix representations. Our implementation uses a dot product of the two encoding matrices which results in a 3D interaction matrix.

- **Feature Extraction**: From the interaction vector, we extract features using the process of convolution by using DenseNet feature extractor. The DenseNet is composed of 3D convolution layers where each layer is connected to layers that are forward in the network. Using 3D convolution and down sampling, we reduce the interaction vector into a 1D array of feature logits.

- **Output Layer**: The feature logits are connected to the output layer in a

fully connected fashion. As opposed to the original implementation which has three output neurons, our implementation has a single neuron since we have a binary classification task at hand.

4.2 Training Procedure

For the feasibility of developing and testing the model iteratively, we used only 100,000 rows of question pairs, 95% of which were Quora question pairs, and 5% were the question pairs that we manually labelled from our own dataset. For the purposes of this project, since the training time was still very large, we did not split the data into train-test set. The accuracy that we report is the cross-validation accuracy that is obtained while the model is being trained. After 3 epochs, the accuracy with which the model was predicting correct label was between 66%.

The hyper-parameters that we used were exactly the same as the ones used by the original paper. We did not change or fine tune any of the hyper-parameters.

The time taken to train the model with 100,000 rows of data was 6 hours per epoch when trained on a CPU, with 16 GB of memory, while on a GPU with 8 CPU'S with 64 GB memory, it took 2 hours per epoch.

5 Results

The model achieved an accuracy of 66% after the third epoch, when using 100,000 rows of question pairs. The sub-optimal accuracy is a result of the fact that only a subset of data was used while training, and since training took a lot of time, the model was run only for 3 epochs. Using the full dataset of 400,000 rows of question pairs, and training a model for a larger number of epochs would increase the accuracy at least by 6-7%. The reason why state of the art accuracy of 88% that the paper mentions could not be obtained was because we could not exactly replicate the encoding layer that the original paper was used, and instead used a pre-trained LSTM to encode sentence embedding. Also, since the DenseNet architecture is extremely complicated to replicate, we used the library code for DenseNet provided by PyTorch without customizing or fine-tuning it to work in accordance with the original paper.

6 Discussion and Future Work

There is a lot of room for improvement in the model that we have trained. However, to solve the sub problem of semantic textual similarity for the University Application product, we could use Transfer Learning and replace our DenseNet with a DenseNet

having pre-trained weights as provided by the original paper. This would improve the accuracy by a significant amount.

Further improvement can be achieved by collecting new question pairs pertaining to the domain of University applications for international students. Using this model with human supervision would help us collect more student question pairs and this new data would allow further improvement in accuracy of domain specific questions.

****