CS 577 - Homework 2

Shahab Rahimirad

Apr 10th 2024

Note: 3 days of late submission used

1 Written Answers

1.1 General

1.1.1 Question 1

The main advantage of LSTM over RNN is its ability to handle long-term dependencies. Vanilla RNNs have to deal with vanishing and exploding gradient in long sequences. LSTM uses gates to control the gradient flow.

1.1.2 Question 2

While standard Word2Vec embeddings trained on a corpus can capture some information about word senses, they are generally not sufficient on their own to reliably identify different word senses for a given word.

The Word2vec embedding will create only one vector for each word regardless of the different meanings. So it will not distinguish between senses because it combines the different senses of a polysemous word. To distinguish between senses of a word, we should use sense-differentiated embeddings that learn a separate vector for each sense of a word.

1.1.3 Question 3

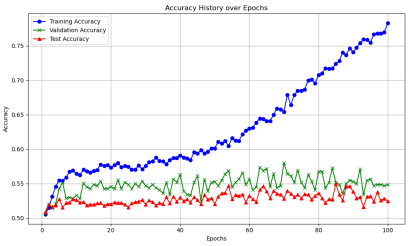
In textual entailment, the goal is to determine if a "hypothesis" text logically follows from a "premise" text. We can create a hypothesis for each sense of the target word and then assess it by checking whether each hypothesis is entailed by the original sentence.

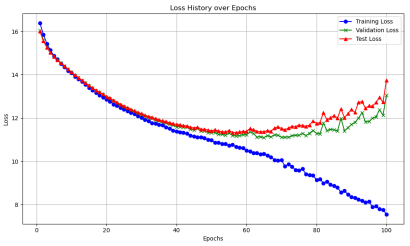
For example with the premise "I deposit money in *bank*." to get the sense of the word *bank* we create these hypothesis: H1: "*bank* is a financial institute" and H2: "*bank* is the land alongside a river" and compare the entailment of these two hypothesis with the premise sentence. Whichever has a higher confidence score, that refers to the correct sense of the word in that context.

1.2 Experiment

1.2.1 Question 1

The best results were with bidirectional LSTM and word embeddings. An embedding layer was used for each context sentence and the target word. Then each one of them were passed through a 2 layer LSTM with dropout. The resulting hidden layers were concatenated and go through a fully connected linear layer alongside the PoS tag of the target word. The final output is calculated through a sigmoid function. We can tell by the loss and accuracy plot that no overfitting is happening.





1.2.2 Question 2

All models performed similarly with around 50 to 54 percent accuracy which matches the baseline mentioned in the WiC paper. The best performance by a small margin belonged to bidirectional LSTM. Explicit ordering in the RNN improved the performance only by a little amount. This might be either a coincident, or might be because explicit ordering captures the semantic meaning of the context better in the phrases that exist in this dataset.

1.2.3 Question 3

Bidirectionality increases the models performance on trianing set significantly, but it causes the model to overfit with the same parameters. It increases the accuracy, but not significantly. This might be because the sense of a word in a context is not affected by the direction. In other words knowing if a word in a context happens before or after the target word does not change the interpretation about the meaning of that word.

1.2.4 Question 4

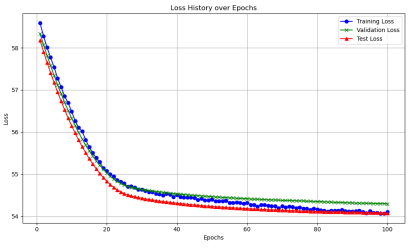
Using embedding such as GloVe improved the performance of the model. This is because we basically are dealing with the meaning of the words that are in the context. By using these embeddings, we incorporate the semantic similarity of sentences and words into our model. We can say that when the context of two target words are more semantically similar, it is more likely that they have the same meaning. These pretrained embeddings have these semantic similarities within them.

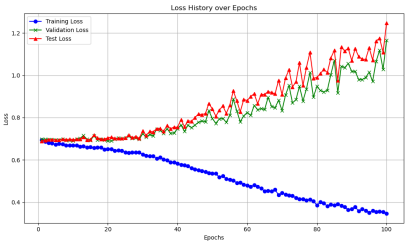
1.2.5 Question 5

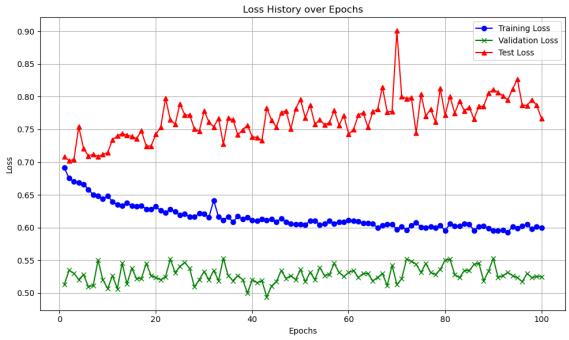
The extra feature that we added to the model was the wordnet lemma of the words in the contexts, hoping that words with the same lemma would end up with the same embedding and thus making the classification job easier. This did not improve the performance by any significant margin.

1.2.6 Question 6

A major hyperparameter that was tuned was the regualrization term either through `lambda_l1` for l1 regularization or `weight_decay` for l2. Before tuning this hyper parameter, The learning curves very obviously showed that the model was overfitting on the training data and the validation loss was starting to increase. But after finetuning this parameter and increasing it we saw that the model was no longer overfitting even though the accuracy did not change that much. However, it should be noted that if this value gets too high the model does not learn the training set well and the loss does not decrease even for the training data. As it can be seen in the plots below showing a good fitting, over fitting, or not learning due to high regularizaiton.







2 Paper Analysis

The paper selcted for this section is the WiC paper.

2.1 Question 1

The main contribution of this paper is the introduction of a benchmark dataset called WiC for evaluating context-sensitive word embeddings. This dataset aims to assess the ability of models to capture fine-grained semantics of words in different contexts, highlighting the need for further research in context-sensitive word embeddings. Various state-of-the-art models were tested on this dataset, indicating room for improvement in discerning different meanings of words in context.

The author mentions that another dataset (SCWS) exists for a similar task, but claims that the quality of that dataset is subpar. The author claims that because most of the context pairs have different target words, the performance of models on that dataset is independant from the context.

2.2 Question 2

The paper outlines the methodology for constructing the WiC dataset, which involves extracting contextual sentences from three major lexical resources: WordNet, VerbNet, and Wiktionary. By leveraging BabelNet's mappings as a bridge between these resources, the authors were able to create a dataset that is not only large in scale but also rich in semantic diversity. This construction methodology contributes to the novelty and utility of the WiC dataset. mappings as a bridge.

The binary classification framework of the dataset is a departure from traditional word similarity benchmarks and introduces a novel approach to evaluating word embeddings and sense representations' ability to capture context-sensitive semantics.

The authors experimented with a variety of Contextualized word-based models and Multi-prototype models to demonstrate these models performance on the introduced dataset.

2.3 Question 3

The paper evaluates a range of models including contextualized word embeddings (Context2Vec, ELMo, BERT), multi-prototype embeddings (DeConf, SW2V, JBT), and baseline models (Bag of Words, Sentence LSTM). Performance is measured using accuracy, comparing model predictions against ground truth labels in the WiC dataset.

The goal of the paper was to show that current models have surpassed the performance ceilings possible with the current evaluation datasets (Stanford SCWS) and introduce a new dataset. The evaluation seeks to validate the dataset's complexity and its effectiveness in pushing the boundaries of research in context-sensitive meaning representation. However, they did not calculate the Inter-rater agreement (IRA) of the WiC dataset independently eventhough IRA was used to evidence towards weakness of other datasets.

2.4 Question 4

The weakness and strength of the paper will be analyzed here.

The main strength of the paper is in its clear definition of the problems with previous benchmarks and introducing a new one to replace them. Obviously in such a situation the quality of the new dataset is important. The sources used for the dataset, the pruning step, and the difference between current models performance and human performance can point to the high quality of the dataset. However, there can be some issues with the way the dataset is formatted. For example there has been a separation only between nouns and verbs and other parts of speech have not been considered. Another weakness that can be associated with the dataset is that it uses examples lexical resources instead of the usual, evey-day use of words. The examples used in the lexical resources might deliberately selected to be as semantically obvious and easy as possible. But in real life situations with longer contexts it might be even more difficult for humans to detect the words meaning.

The dataset and its framework can be used in some other situations/problems with some change. The dataset and the benchamrk itself can be seen as a repackaging of the word sense disambiguation problem and the (target-word, context-1, context-2) binary classification format can be used in other problems in which disambiguation a word's meaning can be important.