Fake News Challenge Team 2

Group: Sahil Chopra, Saachi Jain, John Sholar
TA: Danqi Chen

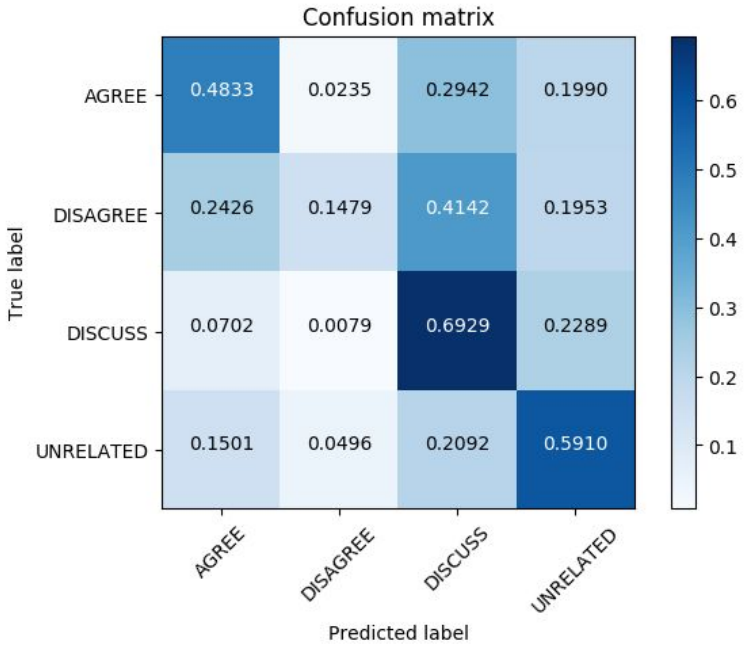
For our project, we are competing in the Fake News Challenge, which focuses on stance detection. We use the same random split utilized by the FNC-1 Baseline, which was released on March 1st. Their scoring metric assigns a weight .25 to correctly determining whether an article is related or unrelated from the headline and .75 to correctly determining whether an article agrees, disagrees, or discusses the headline. For comparison, the official baseline, a linear model, scores 79.53%. For our initial baselines, we implemented three simple models.

Baselines:

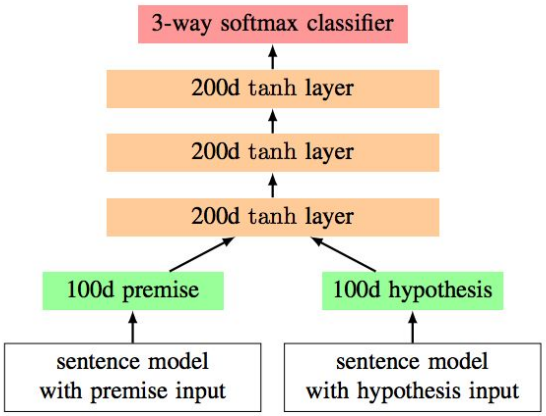
First, we implemented a linear classifier similar to the one described in the original SNLI Corpus paper (Bowman et. al. 2015). Our linear model uses the following features:

- Modified BLEU Scores
    - Since the article and the headline are of vastly different lengths, we apply a sliding window of (window_size = headline_length * 2) and then compute the BLEU Score between these windows and the headline, using equal weighting on unigrams, bigrams, and trigrams. We then take the max of these BLEU Scores for an article-headline pair and use this is a feature.
- Unigram Indicator Variables
    - We use a series of unigram-indicator-variable-features to represent the unigrams that appear in a given headline.
- Cross Unigram Indicator Variables
    - We use series of cross-unigram-indicator-variable features to represent cross unigrams, which are defined as unigram pairs from the headline and article that share the same part of speech - as defined in Bowman et. al. 2015.
- Cross Bigram Indicator Variables
    - We use a series of cross-bigram-indicator-variable features to represent cross bigrams, which are defined as bigram pairs from the headline and article where the second word share the same part of speech - as defined in Bowman et. al. 2015.
- Jaccard Distances
    - We use the jaccard distance of the sets of words used in the headline and article.

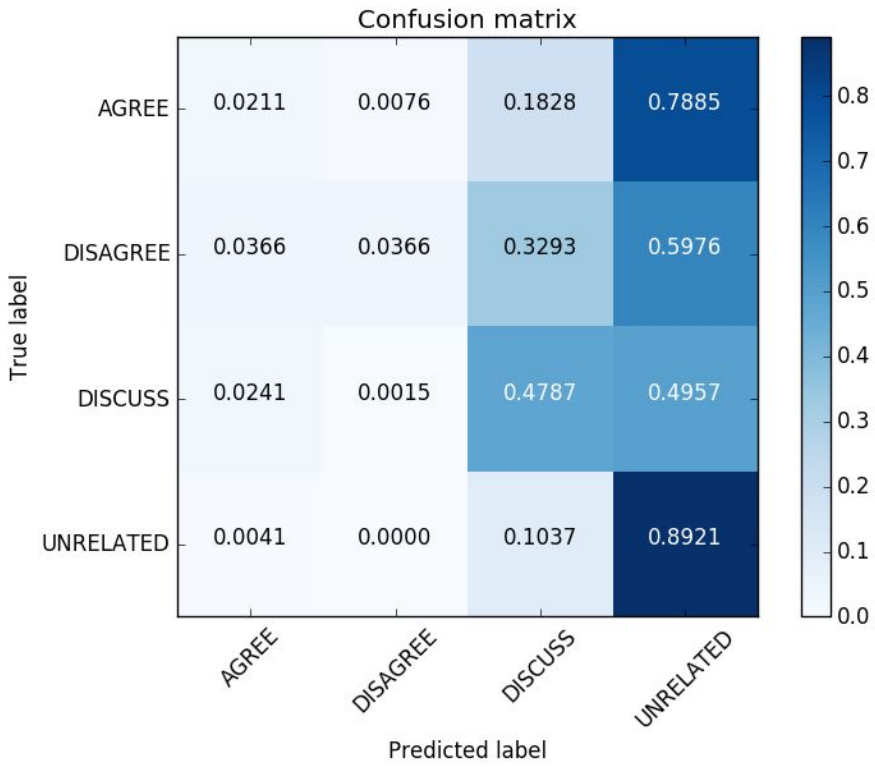
Our simplistic linear model scored ~60% using the official Fake News Challenge weighted score metric and the confusion matrix is provided below.



Second, we implemented the simple multilayer perceptron described in the original SNLI Corpus Paper (Bowman et. al. 2015). This neural network encodes the article and headline using a bag of words model, where each word is represented as a 300-dimensional GloVe vector. We then project these two embedding vectors (1 for headline and 1 for the article body) into 100-dimensional space, concatenate them together, and finally run this 200-dimensional vector through a 3-layer deep neural network that utilizes tanh activation functions. Finally, we apply softmax to generate predictions and cross entropy to determine loss (see image below from Bowman et. al. 2015).



Our simplistic bag-of-words neural net model scored 71.67% and the confusion matrix is provided below.

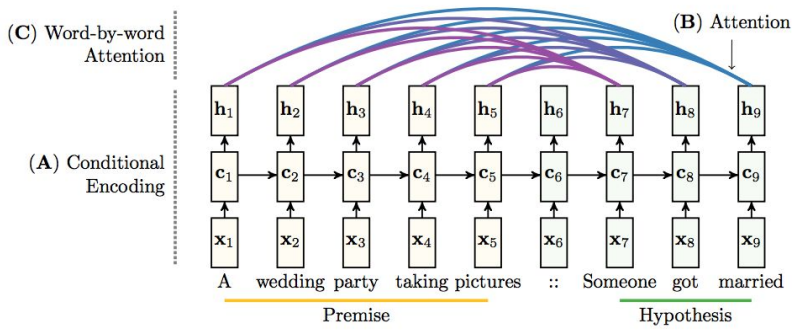


Finally, we are running an LSTM using 300-dimensional GloVe vectors as word-embeddings. As input we fed a concatenated headline-article pairs, simply ignoring words beyond a max length of 725. We're about to kick off this on Azure today.

Next Steps:

Moving forward, we hope to implement the following models:

- Conditional-Encoding LSTM model (Rocktaschel et. al 2016)
    - Run 2 LSTMs - 1 for the headline and 1 for the article body. We shall warm up the LSTM for the article body with the hidden state of the headline encoding.
    - Apply softmax on the final hidden state of 2nd LSTM (article body encoding) to determine predictions and apply cross entropy loss.
- Conditional-Encoding LSTM Model w/ Attention Network (Rocktaschel et. al 2016)
- Conditional-Encoding LSTM Model w/ Word-by-Word Attention Network (Rocktaschel et. al 2016)



Rocktaschel et. al 2016 conditional LSTM with Attention/WxW Attention

We hypothesize that the conditional LSTM model will work better than simply concatenating the article and headline in our baseline LSTM model because the conditional model separates processing of the headline and the article. We further hypothesize that both types of attention will increase the accuracy, but that the general Attention network will be more useful than Word by Word Attention Networks because we are looking more at holistic statements by the article and headline.

Rocktaschel et. al 2016 also tried a conditional LSTM model with 2 Word by Word attention networks. Essentially, rather than just creating attention vectors for the premise which are used in the hypothesis LSTM, they also created attention vectors for the hypothesis that were used in the premise. They found that the double attention vector system did not work well because entailment is asymmetric (premise implies hypothesis). However, stance detection is symmetric. Therefore, we will aim to finally implement a conditional LSTM that uses a 2 way general attention network on top.

Additionally, we are going to examine the potential use of convolutional neural networks to solve this problem. Specifically, we plan to look at the model in (Chen et. al 2016) that was proposed for solving reading comprehension and question-answering tasks, which are somewhat related to our stance detection task.

Github Code: https://github.com/schopra8/fnc-1_cs224n-