The Effect of Attention on NLI

Ankur Dhoot
March 2021

Abstract

Attention is a widely use mechanism in NLP that allows neural models to focus on particular areas of text. In the case of NLI, attention has been used to learn soft alignments between the hypothesis and premise. However, in the case of neutral relation, such alignments may not exist since the premise and hypothesis might be unrelated. Thus, this paper hypothesizes that attention models perform better on entailment / contradictory examples than neutral examples. The experiments run provide some evidence that this may be the case.

1 Introduction

Natural language understanding (NLU) focuses on the ability of machines to understand text. Natural language inference (NLI) is a subfield of NLU that focuses on whether a machine can reasonably infer a natural language hypothesis h from a given premise p (MacCartney and Manning, 2009).

For example,

p : A soccer game with multiple males playing.

$\Rightarrow$

h : Some men are playing a sport.

Entailment suggests that the hypothesis is true given the premise. Contradiction means the hypothesis is false given the premise. Neutral indicates that the hypothesis might be true given the premise.

Bowman et. al (2015) show that neural models can be as competitive on SNLI as the traditional lexicalized feature based classifiers. Rocktaschel et. al (2015) show that attention based neural models provide even better performace on SNLI. Chen et. al (2017) provide the current SOTA attention based model, called ESIM. This paper uses the Bowman et. al (2015) model as a baseline to compare with the attention based ESIM model.

As described in Bahdanau et. al (2015), attention based models help with the problem of encoding an entire sentence representation in a fixed length vector. When applied to neural machine translation, attention allows the decoding to focus on a particular part of the source sentence. When applied to NLI, attention allows the hypothesis to focus on a particular part of the premise (and vice-versa). The idea is that by allowing hypothesis words to focus in on words in the premise, the attention mechanism can figure out which words in the premise are important to each word in the hypothesis. As such, one might wonder how effective attention is at predicting neutral relations, since there might not be relevant words to attend to between the hypothesis and premise.

This original hypothesis that inspired this work is : *Attention helps neural models predict entailment / contradiction classes better than the neutral class*.

This hypothesis is worth investigating because it provides insights into when attention is effective and in what tasks attention is insufficient. In this paper, experiments will show that this hypothesis *could* be correct, although the answer cannot be definitely determined without further experimentation.

2 Related Work

Attention was first applied to NLI in Rocktaschel et al. (2015). This model drew inspiration from the neural machine translation models of Bahdanau et. al (2015) in which attention is applied when decoding. In the NLI case, decoding refers to processing the hypothesis sentence.

Further work was done by Parikh et al. (2016) to make attention models highly parallelizable. These new models were shown to achieve SOTA at the time without having the serial limitation of traditional RNNs.

Chen et al. (2017) proposed a new attention

model that used attention to form a representation of the premise and hypothesis. The novelty of their model was in applying attention both from premise to hypothesis and hypothesis to premise and forming input representations for a second layer of RNNs using attention.

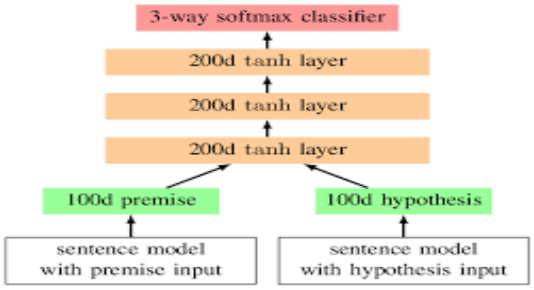
3 Data

To test the hypothesis, this paper uses the SNLI dataset (Bownman et. al, 2015). SNLI was the first large scale dataset released for NLI. The dataset consists of 570k premise-hypothesis pairs. Each pair contains one of three relations - entailment, contradiction or neutral. All premises are chosen from image captions on the Flickr30k corpus. Hypotheses come from Amazon Mechanical Turk workers being asked to provide alternative captions for each of the three relations. Thus, all relations are balanced in the dataset.

4 Models

4.1 Baseline

The baseline model will be the original neural model from Bowman et. al (2015). This model runs the premise and hypothesis through an RNN and uses the RNN output vector as the representation for the premise and hypothesis. The vectors are then concatenated and fed into a deep network with a softmax classifier at the top.



4.2 ESIM

The attention model used to test the hypothesis is ESIM (Chen et al., 2017).

This model first runs the premise and hypothesis through separate BiLSTMS. The hidden representations produced by the BiLSTMS are used to compute attention weights between the premise and hypothesis using simple dot-product attention. The idea is that the hidden representations are in the same "attention space". The attention vectors are then computed using softmax on the attention weights.

The resulting attention vectors are concatenated with the original hidden representations. These vectors form the input to a second set of BiLSTMS. The resulting premise and hypothesis vectors are fed through a pair of BiLSTMs and the resulting output vectors are concatenated and fed through a projection layer with a softmax on top to form the final predictions.

The intuition behind the second set of LSTMs is to perform *inference composition*. As the second set of LSTMs sequentially process the attention vectors from the first set of LSTMs, they might be looking for contradiction or entailment hints and updating their hidden states as such.

5 Experiments

All experiments used 100-D Glove embeddings for input token representation. Unknown words were mapped to an UNK token which had a randomly initialized vector. Embeddings were fine tuned during training. Optimization was done with Adam optimizer using a learning rate of .004. Batch size was 32. Training was done until dev score accuracy no longer increased. Sentences were tokenized on whitespace.

Baseline	Precision	Recall	F1	Support
neutral	.68	.66	.67	3235
contradiction	.69	.68	.69	3278
entailment	.69	.72	.70	3329
accuracy			.69	9842

ESIM	Precision	Recall	F1	Support
neutral	.68	.70	.69	3235
contradiction	.78	.67	.72	3278
entailment	.70	.78	.74	3329
accuracy			.72	9842

6 Analysis

The results presented might deviate from the higher numbers reported in the original papers for a num-

ber of reasons. However, the goal here is to interpret relative results between models, not absolute numbers. Thus, the deviation from the original papers should be of lesser concern.

- 100-D Glove embeddings were used due to model space constraints compared to 300-D in the papers.
- The hidden dimensions were 100 instead of 300 due to space constraints on the model.
- Minimal hyperparameter tuning was done.
- Tokenization was just done using whitespace. The papers used slightly more advanced tokenization schemes.

As can be seen from the experiments, the attention based models performed better overall. Interestingly, both the baseline and ESIM model seem to do better on contradiction and entailment than neutral which could maybe suggest that neutral is a harder task than the others or that the SNLI dataset is constructed in such a manner. ESIM does about .03 F1 better on contradiction than neutral while baseline does about .02 F1 better. ESIM does .05 F1 better on entailment than neutral while baseline does about .03 better. ESIM also does .02 F1 better than baseline on the neutral class indicating attention could be helping the neutral case as well. Overall, the results suggest that attention could help the models perform marginally better on entailment / contradiction than neutral, but the evidence isn't overwhelming.

Another interesting observation is the high precision produced by ESIM on the contradiction class and the high recall produced by ESIM on the entailment class. At this point, it's unclear to me the reason for this.

7 Conclusion

In this paper, it was shown that attention *may* be more beneficial to predicting entailment / contradiction relations than neutral relations. Further work could be done to investigate what the attention mechanism is learning, and if there are patterns to the neutral examples attention fails on.

8 Authorship

All work was performed by Ankur Dhoot.

9 Code

https://github.com/ankurdhoot/CS224U

10 References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 1657–1668.

Ankur P Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016.

Tim Rocktaschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. 2016. Reasoning about entailmentwith neural attention. In Proceedings of ICLR.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting Recurrent and Attention-Based Neural Models: a Case Study on Natural Language Inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 4952–4957