XCS224U - Natural Language Understanding

Literature Review

Natural Language Inference

Ankur Dhoot

February 2021

Contents

1 General Problem And Task Definition

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.

This literature review covers neural models for NLI with a focus on those that use some form of attention.

2 Article Summaries

This section summarizes some papers where neural models are applied to NLI. We start with the original SNLI paper (Bowman et al., 2015) to establish some neural model baselines. This is followed by the paper that popularized attention as it applies to neural models (Bahdanau et al., 2015) to give an overview of what attention is and why it is useful. We then look at a few papers that combine neural models with attention and apply it to NLI. The first builds on the original SNLI paper using recurrent models with attention (Rocktaschel et al., 2015). The second uses attention in a manner that is not subject to the serial nature of recurrent networks (Parikh et al., 2016). The third introduces a new way of interpreting attention based neural models (Ghaeini et al., 2018).

2.1 NLI Background

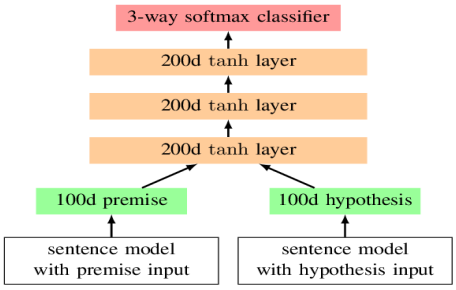
2.1.1 A Large Annotated Corpus For Learning Natural Language Inference

SNLI was the first large scale dataset released for NLI. The dataset consisted 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.

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 show that neural models can be as competitive on SNLI as the traditional lexicalized feature based classifiers. In particular, they build models that encode the premise sentence and separately encode the hypothesis sentence. These representations are concatenated and then fed into a deep neural network. The models differ in how these premise and hypothesis encodings are constructed.



The different encoding schemes tried are a sum of the embedding words, the final output embedding of a vanilla RNN, and the final output embedding of a LSTM RNN. These models provide important baselines for future papers to build on.

2.2 Attention Background

Looking at the baseline SNLI models from Bowman et. al, one might reasonably wonder about the difficulty of trying to encode an entire sentence into a fixed length vector. Additionally, one might notice that the premise and hypothesis are encoded *independently* before concatenation. Could an encoding that utilizes the premise and hypothesis together produce better representations? Here is where the idea of attention comes in.

2.2.1 Neural Machine Translation By Jointly Learning to Align And Translate

Attention was originally popularized in the context of encoder-decoder systems for neural machine translation (Bahdanau et al., 2015). Such systems are used to translate between two languages by passing the source language sentence through an RNN and then passing the resulting encoding through a decoder RNN in an effort to produce the target language translation.

Attention mitigates the bottleneck of representing the entire source sentence with a fixed-length vector by allowing the decoder RNN to look back at the hidden representations corresponding to each word in the source sentence. When the decoder is predicting the next word, it can zoom in on the representation of the word / words from the source sentence that should be translated next.

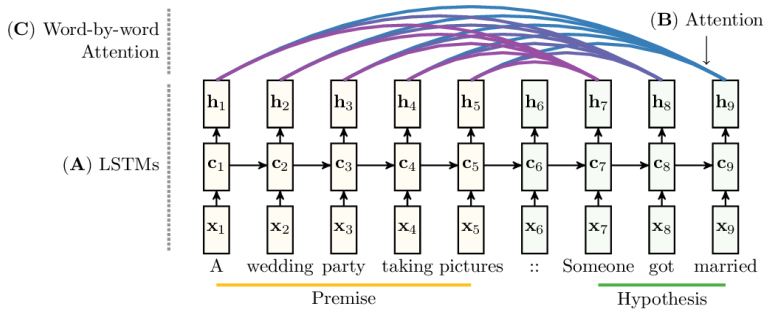
2.3 NLI And Attention

2.3.1 Reasoning about Entailment with Neural Attention

Attention was first applied to neural models for NLI by constructing an architecture similar to the encoder-decoder (Rocktaschel et al., 2015). In this case, the objective wasn't to predict the next translated word, but rather to determine the relation of a premise and hypothesis. Attention was applied in two similar ways - based on the last hidden state of the decoder and based on the hidden states in the decoder after *each* hypothesis word was processed. The attention vector was added to the last hidden state vector and then fed through a non-linear layer followed by a softmax layer to predict the relation class.

The authors report that both attention methods improve over a vanilla encoder-decoder with the word by word attention being better for examples with longer premises. This intuitively makes sense as attention is trying to relieve the bottleneck of representing the premise in a fixed length vector.

The model architecture is as follows.

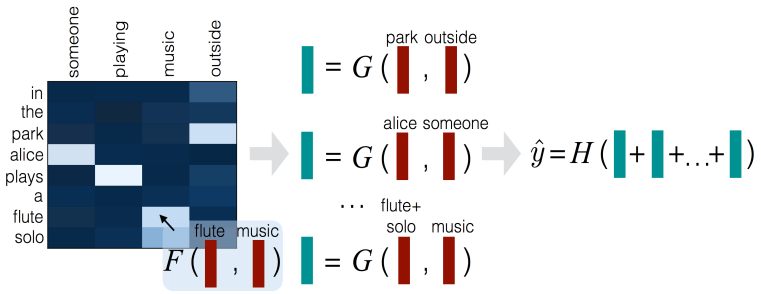


2.3.2 A Decomposable Attention Model For Natural Language Inference

A shortcoming of the encoder-decoder RNN architecture is that RNNs are inherently serial. The output of the next step depends on the previous one. Parikh et al. (2016) propose a parallelizable model based solely on attention and feed forward neural networks. Their proposed architecture consists of three steps - attend, compare, and aggregate.

In the attend step, attention weights e_{ij} are calculated between the ith premise word and jth hypothesis word by taking the dot product of $F(p_i)$ and $F(h_j)$ where F is a feed forward neural network. The corresponding attention vectors are then computed using a softmax on the attention weights (as usual). In the compare step, the attention vector corresponding to a particular word is then compared with the word itself by feeding both as inputs to a feed forward network. The motivation here is that the compare step can give some indication as to whether the attention alignment for a particular word indicated some form of entailment or contradiction. Finally, in the aggregate step, the comparison vectors for each word are summed and the resulting summation is fed through a classifier to produce the final relation prediction.

The model architecture looks as follows.



2.3.3 Interpreting Recurrent and Attention-Based Neural Models: A Case Study on Natural Language Inference

Attention produces great results, but as with many deep learning models, there is sometimes a problem of interpretability. In this paper, Ghaeini et al. (2018) introduce the concept of saliency to attention. Saliency is a measure of how much the attention alignment impacts the final classification decision. In other words, for a decision score S(y) and an attention weight e_{ij}, the saliency is $\frac{\partial S(y)}{\partial e_{ij}}$.

As opposed to the attention weights e_{ij} that tell us how much weight there is between two words, the saliency tells us which of the weights are the most important to the classification decision. This is useful in interpreting model results and informing future research.

3 Compare And Contrast

To begin comparing these papers, here is a table from Parikh et al. (2016) summarizing the performance of all of the models discussed along with a few others.

Method	Train Acc	Test Acc	#Parameters
Lexicalized Classifier (Bowman et al., 2015)	99.7	78.2	–
300D LSTM RNN encoders (Bowman et al., 2016)	83.9	80.6	3.0M
1024D pretrained GRU encoders (Vendrov et al., 2015)	98.8	81.4	15.0M
300D tree-based CNN encoders (Mou et al., 2015)	83.3	82.1	3.5M
300D SPINN-PI encoders (Bowman et al., 2016)	89.2	83.2	3.7M
100D LSTM with attention (Rocktäschel et al., 2016)	85.3	83.5	252K
300D mLSTM (Wang and Jiang, 2016)	92.0	86.1	1.9M
450D LSTMN with deep attention fusion (Cheng et al., 2016)	88.5	86.3	3.4M
Our approach (vanilla)	89.5	86.3	382K
Our approach with intra-sentence attention	90.5	**86.8**	582K

Table 1: Train/test accuracies on the SNLI dataset and number of parameters (excluding embeddings) for each approach

As can be seen, the models with attention seem to do better than the models without. All models are evaluated on the same SNLI dataset which makes the comparisons meaningful.

3.1 Runtime

The models of Bownman et al. (2015) and Rocktaschel et al. (2015) are serial RNNs, while the models of Parikh et al. (2016) are parallelizable. Because of this, the models of Parikh et al. (2016) have limited word order information. However, they still seem to do better. Perhaps this suggests something about the nature of the dataset?

3.2 Sentence Length

Bahdanau et al. (2015) and Rocktaschel et al. (2015) briefly comment that attention seems to especially benefit longer length sentences. Bowman et al. (2015) and Parikh et al. (2016) don't report whether performance differs based on sentence length.

3.3 Interpretability

Rocktaschel et al. (2015) and Bahdanau et al. (2015) both provide diagrams showing the attention weights e_{ij} between some selected premises and hypotheses. Ghaeini et al. (2018) take it a step further and introduce the notion of saliency to demonstrate which weights are the most important.

4 Future Work

4.1 Sentence Length

While a few papers briefly mention the effect of sentence length on attention, having a study dedicated to quantifying this could be beneficial. For example, is it only premise length that attention helps with, or does hypothesis length matter also?

4.2 Interpretability

Having interpretable models is important. Knowing when and how systems fail is the first step in building more resilient systems. Ghaeini et al. (2018) apply their saliency notion to the ESIM model (Chen et al., 2017). It could be illustrative to apply attention saliency to other models to compare whether certain attention weights are universally useful across models.

4.3 Relation Class

None of the papers compare how effective models are between the relation classes. For example, does attention help more with entailment / contradiction than neutral? Such a study could reveal shortcomings of attention if it fails to help with certain classes.

5 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.

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

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.

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