Mind your Partner: Cross-Attention between Premises and Hypotheses with BERT

Alan Dunetz, alan.dunetz@gmail.com

Abstract

We analyzed BERT's performance on the RTE task using the SNLI dataset. We compared the results to an LSTM model using GloVe vectors. In addition, we compared the performance of both models to hypothesis-only baselines and tested our fine-tuned BERT model on a "hard" dataset consisting of examples incorrectly classified by its hypothesis-only baseline. Finally, we analyzed the attention matrices generated by BERT to find evidence of cross-attention between premises and hypothesis.[1]

I. Introduction

Recognizing Textual Enhancement (RTE) is the task of determining whether for two sentences: 1) the first sentence (the premise) entails the second sentence; 2) the two sentences are unrelated; or 3) the two sentences are contradictory. The RTE task is intended to show whether a computer model understands the semantics/meaning of sentences. It is therefore a key challenge in the field of natural language inference (NLI).

Models that have been pertained on vast corpuses of data have performed well on a variety of Natural Language Procession (NLP)/Natural Language Understanding Tasks (NLU) with a minimal amount of fine-turning or additional training. We analyzed the performance of BERT on the RTE task and compared it to the performance of an LSTM model using GLoVe vectors.

Some researchers have found that crowd-sourced data sets may contain statistical anomalies that allow models to perform well without considering the premise and are therefore not really performing an NLI task. We created hypothesis-only baselines for both of our models in order to test how their performance improves when they have access to both the premise and the hypothesis. We further tested our BERT model on a 'hard' dataset consisting of instances which the hypothesis-only baseline failed to classify correctly.

Finally, we analyzed the attention weight matrices generated by BERT to examine the degree to which the model applies attention to the hypothesis when evaluating the premise and vice versa.

II. Related Literature

Dagan, et.al., introduced an early dataset to test NLI models. Bowman, et. al., introduced the SNLI corpus. which includes 570K labeled sentence pairs, over 400 times the size of the Dagan dataset. Their larger dataset provided sufficient data to train more complex models, including neural networks on the RTE task. Bowman evaluated a variety of NLI models on the corpus including rule-based systems, linguistic classifiers and neural networks and found the best results with two models. The first was feature-rich linguistic classifier and the second was an LSTM model.

Rocktanshel, et. al., extended this analysis on the SNLI dataset in a number of ways. Their LSTM model trained the premise and hypothesis together to determine entailment and extended the model to use an attention mechanism.

A number of authors explored statistical anomalies in NLI datasets. For example, Poliak et. al., argued that any interesting NLI model must depend on both the premise and the hypothesis. If a model which only has access to the hypotheses can perform

[1] Source code can be found at GitHub.com/dunetz/cs224u

well then it is only performing NLI to the extent that inference is based on prior background knowledge. Alternatively, such a model may be exploiting quirks or statistical irregularities in the dataset. The authors explored a number of data sets and found that for many of them a hypothesis-only model significantly outperformed the majority baseline. The authors identified a number of potential statistical irregularities in these datasets.

Similar to Poliak, Gururangen, et. al., noted that hypotheses generated from a crowdsourcing process may contain artifacts that enable a model to classify a hypothesis without looking at the premise. Rather than performing inference in such cases, they argue that the model is simply exploiting statistical irregularities. The authors attempted to identify some of these artifacts based on category label. In the entailment category, for example, the authors found more generic words such as 'animal' or 'instrument' that were probably selected to generalize over more specific premise words such as 'cat' or 'piano'.

Devlin, et.al., introduced the BERT model, a large transformer network (Vaswani, et. al.) designed to retrain deep bidirectional representations from unlabeled text. Transformers consist of multiple attention layers where each layer contains multiple attention heads.

Clark, et. al., proposed methods for analyzing the attention mechanisms of BERT and identified a number of patterns.

II. Data

The premises for the SNLI data set are captions from the FLickr30k corpus, a dataset of approximately 160k crowd-sourced captions corresponding to about 30k photos. For each caption they were shown, workers were instructed to create a hypothesis for each of the three labels: entailment, neutral and contradiction. Consequently, the dataset is balanced The model uses the "base" sized BERT model which uses 768d embedding vectors.

between the labels. An additional round of validation was performed on 10% of the data which showed a high overall level of agreement with the original labels (91%).

The SNLI dataset is broken into training, dev and test subsets. The training set consists of approximately 550k premise/hypothesis pairs, and the dev and test sets each contain approximately 9,800 pairs. The average length of premises is approximately 66 words and the average length of hypothesis is approximately 38 words (i.e, the hypotheses are 57% as long as the premises).

III. Models

For the LSTM model, we used the LSTM chained model in the course repository . This model chained the premise and hypothesis as a pair. The model consists of a single LSTM layer followed by a linear layer. Our version of the model used 300d GloVe vectors, had a hidden layer of size 300 and was bidirectional. We performed a maximum of 100 epochs of training, used an Adam optimizer and a learning rate of 0.01. Our batch size was 512.

Given time constraints, we did not attempt to fine tune this model. For example we could have added additional feed-forward layers, explored drop-out and other regularization methods.

We also created a hypothesis-only version of the model which only provided the hypothesis for each pair as input to the model.

For the BERT model, we reviewed two applications based on the HuggingFace Transformer Library: 1) McCormack et. al, which applies BERT to sentence classification; and 2) Gao, which applies BERT to the NLI task. We used Gao's implementation as the basis of our own model.

The model has 12 attention layers, each with 12 multi-head attention transformers. A single

linear layer followed by a soft-max is aded for the classification problem. We performed one one epoch of fine tuning.

As with the LSTM, we also created a hypothesis only version of our BERT model.

All models were run on Google Colab Pro, using a single GPU, in most cases, Tesla P100-PCIE-16GB.

IV. Results and Discussion

We trained/fine-tuned four models in total on the full SNLI dataset:

1. LSTM model on premise/hypothesis pairs (LSTM model)
2. LSTM model using only the hypotheses (LSTM-HO model)
3. BERT model using premise/hypothesis pairs (BERT Model)
4. BERT model using only hypotheses (BERT-HO model)

We tested all models on test dataset. In addition, we tested our BERT model on a separate "hard" test dataset which consisted of all examples in the test dataset which the BERT-HO model failed to classify correctly. We report macro-F1 results in Table 1.[2]

The LSTM model only performed 11% percent better (73% vs 66%) than its hypotheses-only baseline, which suggests that the model may be on statistical anomalies in hypotheses.

The BERT model significantly outperformed the LSTM model (90% vs 73%). Also, the BERT model outperformed its hypothesis-only baseline by a larger percentage, 30% (90% vs 69%).

The BERT model also performed well on the "hard" dataset (77%) where the hypothesis-only model failed, suggesting that the BERT model is evaluating both premises and hypotheses.

	Test	"Hard" Test
LSTM Model	73%	
LSTM-HO Model	66%	
BERT Model	90%	77%
BERT-HO Model	69%	

Table 1: Macro-F1 results on test data.

V. Attention Head Visualization

We examined the attention matrix weights of our fine-tuned Bert Model to see the degree which BERT attends to the premise when evaluating tokens in the hypothesis. We performed this analysis at three levels of granularity.

First, we looked at attention-weights on a token by token basis in individual attention matrices. For each token in an instance, we examined attention weights across the premise-hypothesis pair. In Figure 1, we show two attention matrices for a single pair: 1) the first multi-head attention matrix in the first layer (figure 1a); and 2) the first multi-head attention at the 12th layer (Figure 1b). In the first matrix, we can see that the bulk of each token's attention is on other tokens in its sentence. But we can also see, particularly with the tokens in the hypothesis, that some attention is also paid to the partner sentence. In the second matrix, in the 12th layer, we see that separator characters and punctuation marks (i.e., the period at the end of each sentence) attract the highest attention. But, we can also see, in this example, some continued degree of cross attention between the pairs.

Second, for an individual example, we averaged across all multi-head attention matrices at each level for all of the tokens in the premise; and separately for all of the tokens in the hypothesis. We found a similar pattern a in the more granular level: at

[2] Our results on the test data set were close to the results on the dev set (not reported). We believe this occurred because only minimal fine-tuning was performed on model parameters during training.

lower level he bulk of each token's attention is on other tokens in the same sentence. However, we also see a degree of cross-attention. At higher levels, separators and sentence periods attract the higher attention. But there is still observable attention across sentences.

Finally, we aggregated across all examples in the test set. We divided tokens into three categories: 1) premise tokens (sentence 1); 2) hypothesis tokens (sentence 2); and 3) separator and punctation tokens.[3] For each category we averaged all tokens across all multi-head attention matrices at each level. We show the results for the first two categories in Figure 3. Results are consistent with those described for the two more granular analyses.

VI. Conclusion

We applied BERT to the NLI task using the SNLI dataset and compared it to LSTM RNN model using GLoVE vectors. With one epoch of training, the BERT model significantly out performed the LSTM model.

We also trained/fine-tuned both models on hypothesis-only baselines and found that these hypothesis-only models performed well, suggesting the presence of statistical anomalies in the crowd-sourced dataset. However, when we compared the models trained on the full dataset versus the hypothesis-only models, we found a much larger improvement in the BERT model than the LSTM model, which suggests that BERT is less reliant on statistical anomalies in the hypothesis.

In addition, we created a "hard" dataset which consisted of examples which the hypothesis-only fine-tuned BERT model failed to label correctly. The premise-hypothesis fine-tuned BERT model performed well on these instances, suggesting that the model is drawing on both the premise and the hypothesis when performing the NLI task.

Finally, we examined the weights of attention matrices of the fine-tuned BERT model and found evidence that BERT applies attention across sentences in premise/hypothesis pairs.

[3] We classified all periods as punctuation and did not distinguish between periods at the end of a sentence and periods that occurred after abbreviations (e.g., 'Mr.'). We only classified periods as punctation (e.g., not exclamation points or question marks), but we believe that non-period punctuation was relatively rare in the dataset.

['A person wearing a straw hat, standing outside working a steel apparatus with a pile of coconuts on the ground.', 'A person is selling coconuts.']

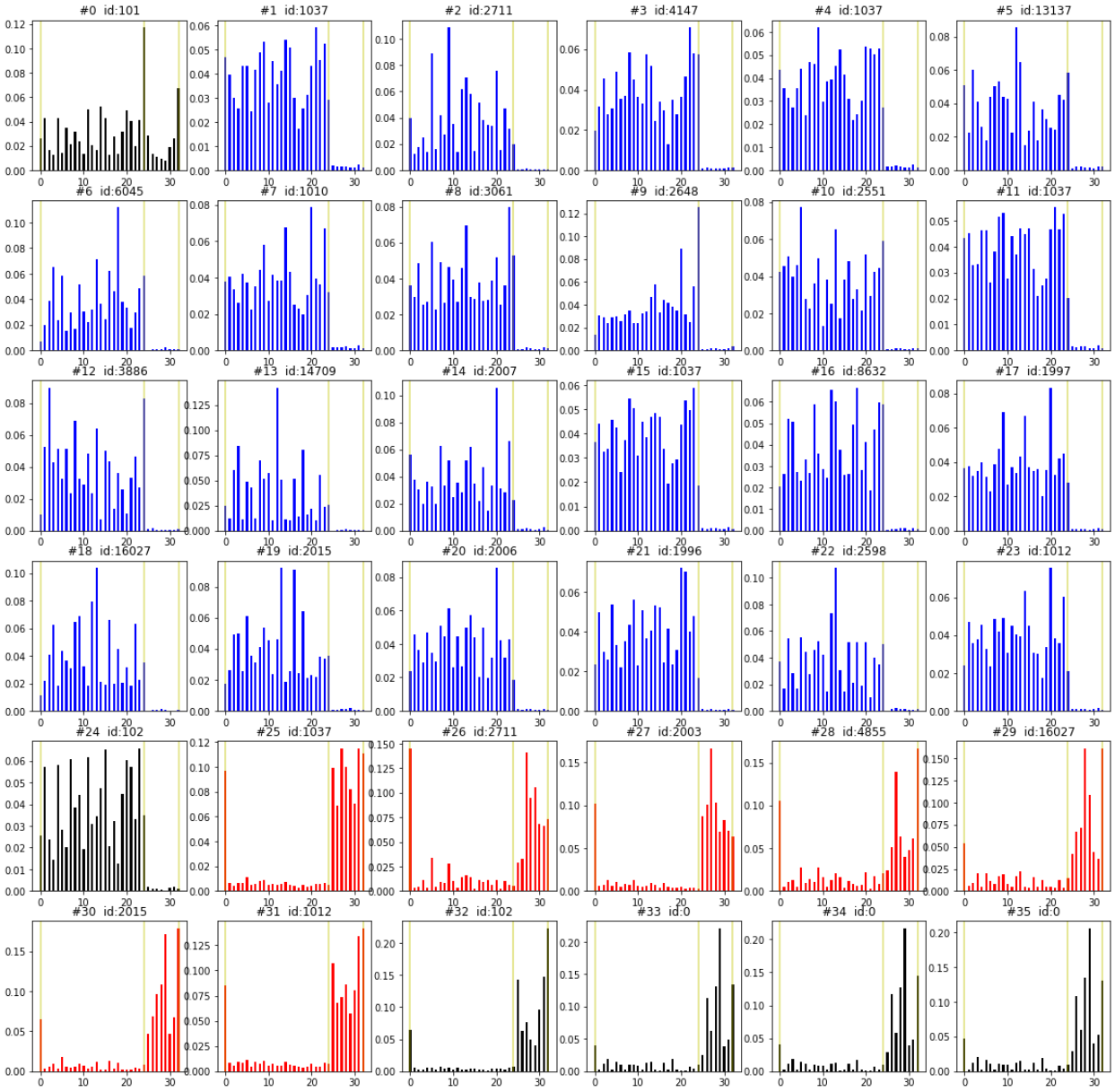

Figure 1a: For the premise/hypothesis pair above, for a single attention matrix, each figure shows the attention weights for one token. Separator and fill tokens are in black. Tokens in the premise are blue. Tokens in the hypothesis are red. The yellow line marks the separator between the premise and the hypothesis. Figure 1a shows the attention matrix for level 1, multi-head 1. Figure 1b shows the attention matrix for layer 12, multi-head 1 for the same pair.

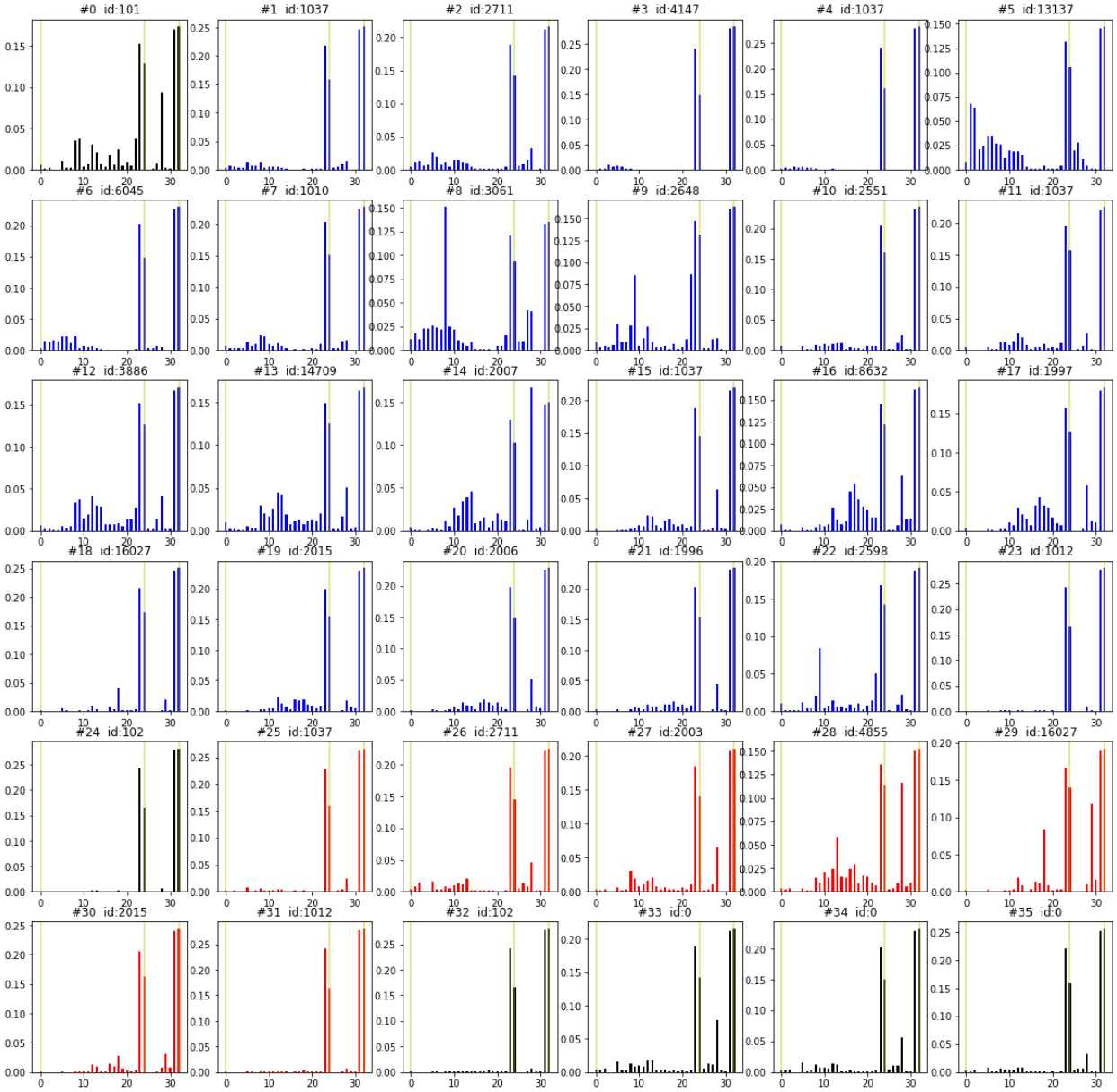

Figure 1b: Attention matrix for layer 12, multi-head 1. See caption for Figure 1a and the text for additional detail.

['A young family enjoys feeling ocean waves lap at their feet.', 'A young man and woman take their child to the beach for the first time.']

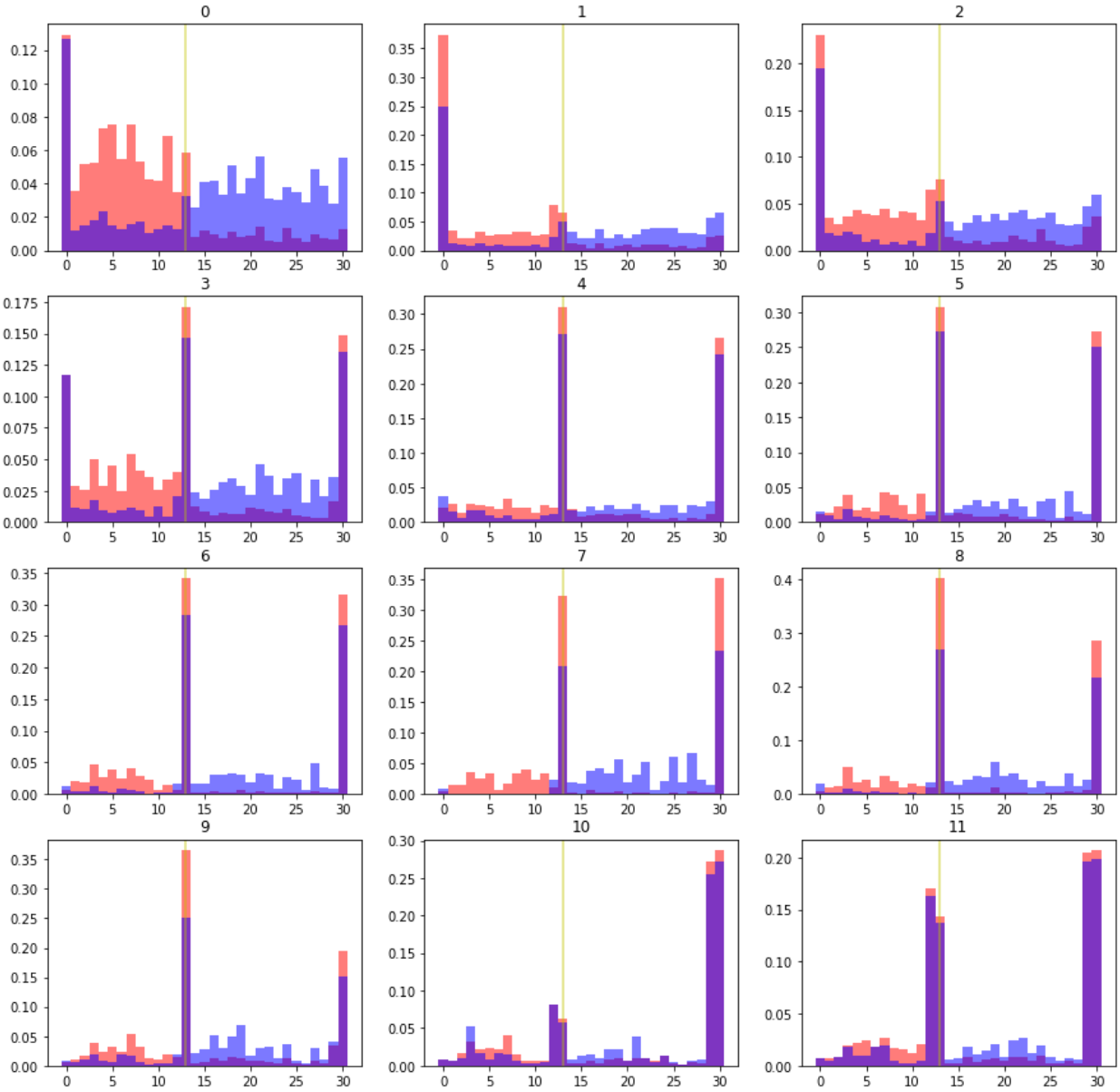

Figure 2: For the premise/hypothesis instance above, each figure shows, for one attention level: 1) the tokens to which the premise tokens, attend (red) on average,; and 2) the tokens to which the hypothesis tokens attend (blue) on average. In each figure, the 12 multi-head attention weight matrices for that level are averaged. The yellow line marks the separator between the two sentences.

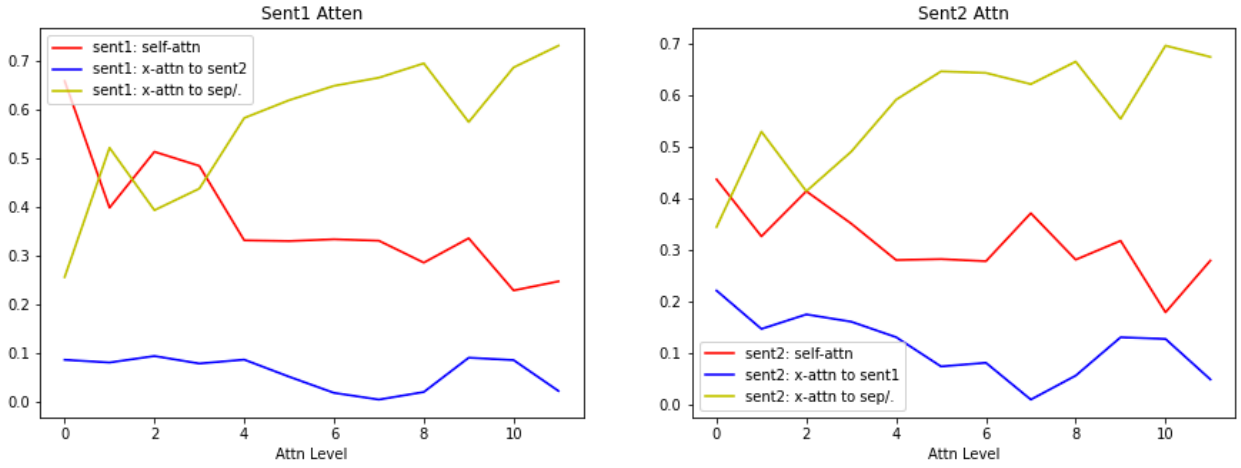

Figure 3: The figure on the left shows 1) *premise self attention* (red):i.e., how much do tokens in the premise, on average, pay attention to premise tokens, across each of the 12 attention levels; 2) *premise cross-attention (blue):* i.e., how much attention do tokens in the premise, on average, pay to tokens in the hypothesis (red); and 3) premise separator/punctation attention : i.e., how much attention do tokens in the premise, on average pay to sentence separators and periods (yellow). The figure on the right shows the same thing for the hypothesis. Data is averaged across all instances in the SNLI test data set.

References

Ido Dagan, Oren Glickman, Bernando Magnini, The PASCAL Recognising Textual Entailment Challenge. In Machine Learning Challenges pages 177-190. https://u.cs.biu.ac.il/~nlp/RTE1/Proceedings/dagan_et_al.pdf

Samuel R. Bowman, Christopher Potts, Gabor Angeli and Christopher Manning, A large annotated corpus for learning natural language inference, 2015. In Proc of EMNLP https://arxiv.org/abs/1508.05326

Tim Rocktaschel, Edward Grefenstette, Karl Moritz Herman, Tomas Kocisky and Phil Blumson, Reasoning about Entailment with Neural Attention, 2015. In ICLR 2016 https://arxiv.org/pdf/1509.06664.pdf

Adam Poliak, Jason Nardowsky, Aparajita Haldar, Rachel Rudinger and Benjamin Van Durme, Hypothesis Only Baselines in Natural Language Inference, 2018.In Proc of *SEM https://www.aclweb.org/anthology/S18-2023.pdf

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman and Noah A. Smith, Annotation Artifacts in Natural Language Inference Data, 2018 In Proc of NAACL. https://arxiv.org/abs/1803.02324

Jacob Devlin, MIng-Wi Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Submitted on October 11, 2018 (v1), last revised May 24, 2019 https://arxiv.org/abs/1810.04805

Chris McCormick, Nick Ryan, BERT Fine-Tuning Tutorial with PyTorch, July 22, 2019, Revised 3/20/20
https://mccormickml.com/2019/07/22/BERT-fine-tuning/

Yang Gao, A Natural Language Inference (NLI) model based on BERT
https://github.com/yg211/bert_nli

Kevin Clark, Urvashi Kandelwal, Omer Levy, Christopher Manning, What does BERT Look At? An Analysis of Bert's Attention, June 2019
https://arxiv.org/pdf/1906.04341.pdf

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukas Kaiser and Illia Polusukhin, Attention is al you need, June 2017
https://arxiv.org/pdf/1706.03762.pdf