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

Reza Ghaeini, Xiaoli Z. Fern, Prasad Tadepalli
School of Electrical Engineering and Computer Science, Oregon State University
1148 Kelley Engineering Center, Corvallis, OR 97331-5501, USA
{ghaeinim, xfern, tadepall}@eecs.oregonstate.edu

Abstract

Deep learning models have achieved remarkable success in natural language inference (NLI) tasks. While these models are widely explored, they are hard to interpret and it is often unclear how and why they actually work. In this paper, we take a step toward explaining such deep learning based models through a case study on a popular neural model for NLI. In particular, we propose to interpret the intermediate layers of NLI models by visualizing the saliency of attention and LSTM gating signals. We present several examples for which our methods are able to reveal interesting insights and identify the critical information contributing to the model decisions.

1 Introduction

Deep learning has achieved tremendous success for many NLP tasks. However, unlike traditional methods that provide optimized weights for human understandable features, the behavior of deep learning models is much harder to interpret. Due to the high dimensionality of word embeddings, and the complex, typically recurrent architectures used for textual data, it is often unclear how and why a deep learning model reaches its decisions.

There are a few attempts toward explaining/interpreting deep learning-based models, mostly by visualizing the representation of words and/or hidden states, and their importances (via saliency or erasure) on shallow tasks like sentiment analysis and POS tagging (Bahdanau et al., 2014; Denil et al., 2014; Li et al., 2016; Arras et al., 2017; Li et al., 2017; Rei and Søgaard, 2018). In contrast, we focus on interpreting the gating and attention signals of the intermediate layers of deep models in the challenging task of Natural Language Inference. A key concept in explaining deep models is saliency, which determines what is critical for the final decision of a deep model. So far, saliency has only been used to illustrate the impact of word embeddings. In this paper, we extend this concept to the intermediate layer of deep models to examine the saliency of attention as well as the LSTM gating signals to understand the behavior of these components and their impact on the final decision.

We make two main contributions. First, we introduce new strategies for interpreting the behavior of deep models in their intermediate layers, specifically, by examining the saliency of the attention and the gating signals. Second, we provide an extensive analysis of the state-of-the-art model for the NLI task and show that our methods reveal interesting insights not available from traditional methods of inspecting attention and word saliency.

In this paper, our focus was on NLI, which is a fundamental NLP task that requires both understanding and reasoning. Furthermore, the state-of-the-art NLI models employ complex neural architectures involving key mechanisms, such as attention and repeated reading, widely seen in successful models for other NLP tasks. As such, we expect our methods to be potentially useful for other natural understanding tasks as well.

2 Task and Model

In NLI (Bowman et al., 2015), we are given two sentences, a premise and a hypothesis, the goal is to decide the logical relationship (*Entailment*, *Neutral*, or *Contradiction*) between them.

Many of the top performing NLI models (Ghaeini et al., 2018b; Tay et al., 2018; Peters et al., 2018; McCann et al., 2017; Gong et al., 2017; Wang et al., 2017; Chen et al., 2017), are variants of the ESIM model (Chen et al., 2017), which we choose to analyze in this paper. ESIM reads the sentences independently using LSTM at first, and then applies attention to align/contrast

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4952–4957
Brussels, Belgium, October 31 - November 4, 2018.

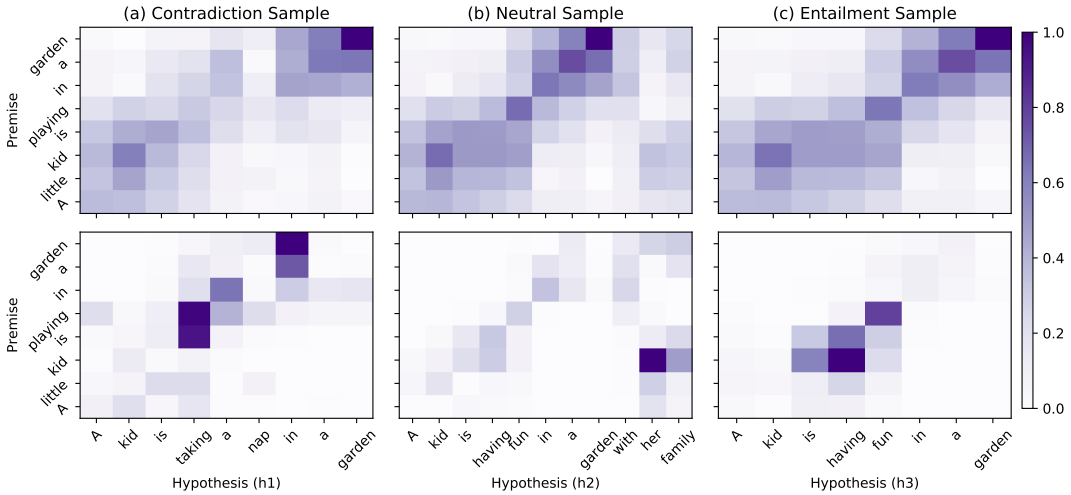

Figure 1: Normalized attention and attention saliency visualization. Each column shows visualization of one sample. Top plots depict attention visualization and bottom ones represent attention saliency visualization. Predicted (the same as Gold) label of each sample is shown on top of each column.

the sentences. Another round of LSTM reading then produces the final representations, which are compared to make the prediction. Detailed description of ESIM can be found in the Supplementary Materials.

Using the SNLI (Bowman et al., 2015) data, we train two variants of ESIM, with dimensionality 50 and 300 respectively, referred to as ESIM-50 and ESIM-300 in the remainder of the paper.

3 Visualization of Attention and Gating

In this work, we are primarily interested in the internal workings of the NLI model. In particular, we focus on the attention and the gating signals of LSTM readers, and how they contribute to the decisions of the model.

3.1 Attention

Attention has been widely used in many NLP tasks (Ghaeini et al., 2018a; Dhingra et al., 2017; Bahdanau et al., 2014) and is probably one of the most critical parts that affects the inference decisions. Several pieces of prior work in NLI have attempted to visualize the attention layer to provide some understanding of their models (Ghaeini et al., 2018b; Parikh et al., 2016). Such visualizations generate a heatmap representing the similarity between the hidden states of the premise and the hypothesis (Eq. 3 of the Supplementary Materials). Unfortunately the similarities are often the same regardless of the decision.

Let us consider the following example, where the same premise *"A kid is playing in the garden"*, is paired with three different hypotheses:

h1: *A kid is taking a nap in the garden*

h2: *A kid is having fun in the garden with her family*

h3: *A kid is having fun in the garden*

Note that the ground truth relationships are Contradiction, Neutral, and Entailment, respectively.

The first row of Fig. 1 shows the visualization of normalized attention for the three cases produced by ESIM-50, which makes correct predictions for all of them. As we can see from the figure, the three attention maps are fairly similar despite the completely different decisions. The key issue is that the attention visualization only allows us to see how the model aligns the premise with the hypothesis, but does not show *how such alignment impacts the decision.* This prompts us to consider the saliency of attention.

3.1.1 Attention Saliency

The concept of saliency was first introduced in vision for visualizing the spatial support on an image for a particular object class (Simonyan et al., 2013). In NLP, saliency has been used to study the importance of words toward a final decision (Li et al., 2016) .

We propose to examine the saliency of attention. Specifically, given a premise-hypothesis pair and the model's decision y, we consider the similarity between a pair of premise and hypothesis hidden states e_{ij} as a variable. The score of the decision $S(y)$ is thus a function of e_{ij} for all i and j. The saliency of e_{ij} is then defined to be $|\frac{\partial S(y)}{\partial e_{ij}}|$.

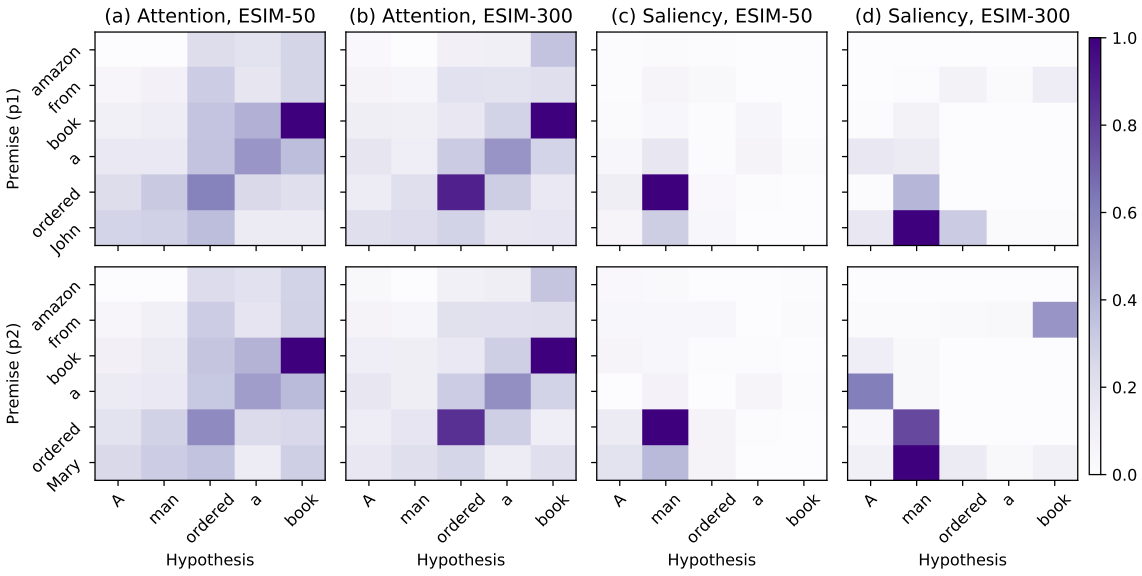

Figure 2: Normalized attention and attention saliency visualizations of two examples (p1 and p2) for ESIM-50 (a) and ESIM-300 (b) models. Each column indicates visualization of a model and each row represents visualization of one example.

The second row of Fig. 1 presents the attention saliency map for the three examples acquired by the same ESIM-50 model. Interestingly, the saliencies are clearly different across the examples, each highlighting different parts of the alignment. Specifically, for h1, we see the alignment between "is playing" and "taking a nap" and the alignment of "in a garden" to have the most prominent saliency toward the decision of Contradiction. For h2, the alignment of "kid" and "her family" seems to be the most salient for the decision of Neutral. Finally, for h3, the alignment between "is having fun" and "kid is playing" have the strongest impact toward the decision of Entailment.

From this example, we can see that by inspecting the attention saliency, we effectively pinpoint which part of the alignments contribute most critically to the final prediction whereas simply visualizing the attention itself reveals little information.

3.1.2 Comparing Models

In the previous examples, we study the behavior of the same model on different inputs. Now we use the attention saliency to compare the two different ESIM models: ESIM-50 and ESIM-300.

Consider two examples with a shared hypothesis of *"A man ordered a book"* and premise:

p1: *John ordered a book from amazon*

p2: *Mary ordered a book from amazon*

Here ESIM-50 fails to capture the gender connections of the two different names and predicts Neutral for both inputs, whereas ESIM-300 correctly predicts Entailment for the first case and Contradiction for the second.

In the first two columns of Fig. 2 (column a and b) we visualize the attention of the two examples for ESIM-50 (left) and ESIM-300 (right) respectively. Although the two models make different predictions, their attention maps appear qualitatively similar.

In contrast, columns 3-4 of Fig. 2 (column c and d) present the attention saliency for the two examples by ESIM-50 and ESIM-300 respectively. We see that for both examples, ESIM-50 primarily focused on the alignment of "ordered", whereas ESIM-300 focused more on the alignment of "John" and "Mary" with "man". It is interesting to note that ESIM-300 does not appear to learn significantly different similarity values compared to ESIM-50 for the two critical pairs of words ("John", "man") and ("Mary", "man") based on the attention map. The saliency map, however, reveals that the two models use these values quite differently, with only ESIM-300 correctly focusing on them.

3.2 LSTM Gating Signals

LSTM gating signals determine the flow of information. In other words, they indicate how LSTM reads the word sequences and how the information from different parts is captured and combined. LSTM gating signals are rarely analyzed, possibly

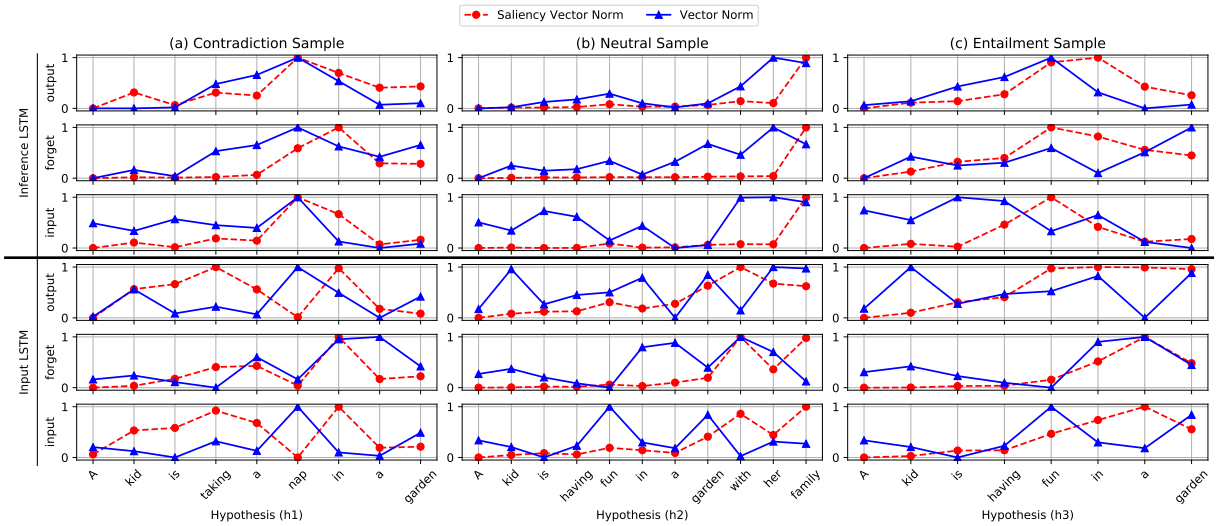

Figure 3: Normalized signal and saliency norms for the input and inference LSTMs (forward) of ESIM-50 for three examples. The bottom (top) three rows show the signals of the input (inference) LSTM. Each row shows one of the three gates (input, forget and output).

due to their high dimensionality and complexity. In this work, we consider both the gating signals and their saliency, which is computed as the partial derivative of the score of the final decision with respect to each gating signal.

Instead of considering individual dimensions of the gating signals, we aggregate them to consider their norm, both for the signal and for its saliency. Note that ESIM models have two LSTM layers, the first (input) LSTM performs the input encoding and the second (inference) LSTM generates the representation for inference.

In Fig. 3 we plot the normalized signal and saliency norms for different gates (input, forget, output)[1] of the Forward input (bottom three rows) and inference (top three rows) LSTMs. These results are produced by the ESIM-50 model for the three examples of Section 3.1, one for each column.

From the figure, we first note that the saliency tends to be somewhat consistent across different gates within the same LSTM, suggesting that we can interpret them jointly to identify parts of the sentence important for the model's prediction.

Comparing across examples, we see that the saliency curves show pronounced differences across the examples. For instance, the saliency pattern of the Neutral example is significantly different from the other two examples, and heavily concentrated toward the end of the sentence ("with her family"). Note that without this part of the sentence, the relationship would have been Entailment. The focus (evidenced by its strong saliency and strong gating signal) on this particular part, which presents information not available from the premise, explains the model's decision of Neutral.

Comparing the behavior of the input LSTM and the inference LSTM, we observe interesting shifts of focus. In particular, we see that the inference LSTM tends to see much more concentrated saliency over key parts of the sentence, whereas the input LSTM sees more spread of saliency. For example, for the Contradiction example, the input LSTM sees high saliency for both "taking" and "in", whereas the inference LSTM primarily focuses on "nap", which is the key word suggesting a Contradiction. Note that ESIM uses attention between the input and inference LSTM layers to align/contrast the sentences, hence it makes sense that the inference LSTM is more focused on the critical differences between the sentences. This is also observed for the Neutral example as well.

It is worth noting that, while revealing similar general trends, the backward LSTM can sometimes focus on different parts of the sentence (e.g., see Fig. 8 of the Supplementary Materials), suggesting the forward and backward readings provide complementary understanding of the sentence.

[1] We also examined the memory cell but it shows very similar behavior with the output gate and is hence omitted.

4 Conclusion

We propose new visualization and interpretation strategies for neural models to understand how and why they work. We demonstrate the effectiveness of the proposed strategies on a complex task (NLI). Our strategies are able to provide interesting insights not achievable by previous explanation techniques. Our future work will extend our study to consider other NLP tasks and models with the goal of producing useful insights for further improving these models.

References

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 159–168.

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642.

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.

Misha Denil, Alban Demiraj, Nal Kalchbrenner, Phil Blunsom, and Nando de Freitas. 2014. Modelling, visualising and summarising documents with a single convolutional neural network. *CoRR*, abs/1406.3830.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. *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 1832–1846.

Reza Ghaeini, Xiaoli Z. Fern, Hamed Shahbazi, and Prasad Tadepalli. 2018a. Dependent gated reading for cloze-style question answering. *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3330–3345.

Reza Ghaeini, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z. Fern, and Oladimeji Farri. 2018b. Dr-bilstm: Dependent reading bidirectional LSTM for natural language inference. *NAACL HLT 2018, The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yichen Gong, Heng Luo, and Jian Zhang. 2017. Natural language inference over interaction space. *CoRR*, abs/1709.04348.

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 681–691.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6297–6308.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2249–2255.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237.

Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 293–302.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. A compare-propagate architecture with alignment factorization for natural language inference. *CoRR*, abs/1801.00102.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4144–4150.