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

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. we take a step toward explaining such deep learning based models through a case study on a popular neural model for NLI. 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. 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. 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.

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, 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, are variants of the ESIM model, which we choose to analyze. ESIM reads the sentences independently using LSTM at first, and then applies attention to align/contrast the sentences. Another round of LSTM reading then produces the final representations, which are compared to make the prediction.

3 Visualization of Attention and Gating

we are primarily interested in the internal workings of the NLI model. 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 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. Such visualizations generate a heatmap representing the similarity between the hidden states of the premise and the hypothesis. 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 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. In NLP, saliency has been used to study the importance of words toward a final decision.

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 eij as a variable. The score of the decision S(y) is thus a function of eij for all i and j. The saliency of eij is then defined to be |S(y) / eij|.

, 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.

Although the two models make different predictions, their attention maps appear qualitatively similar.

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". 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. It is

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 due to their high dimensionality and complexity. 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.

, 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. 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, suggesting the forward and backward readings provide complementary understanding of the sentence.

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.

5 Appendix

5.1 Model

In this section we describe the ESIM model. We divide ESIM to three main parts: 1) input encoding, 2) attention, and 3) inference.

Let u = [u1, · · · , un] and v = [v1, · · · , vm] be the given premise with length n and hypothesis with length m respectively, where ui, vj Rr are word embeddings of r-dimensional vector. The goal is to predict a label y that indicates the logical relationship between premise u and hypothesis v. Below we briefly explain the aforementioned parts.

5.1.1 Input Encoding

It utilizes a bidirectional LSTM (BiLSTM) for encoding the given premise and hypothesis using Equations 1 and 2 respectively.

(1) u^ = BiLSTM(u)

(2) v^ = BiLSTM(v)

where u^ Rn×2d and v^ Rm×2d are the reading sequences of u and v respectively.

5.1.2 Attention

It employs a soft alignment method to associate the relevant sub-components between the given premise and hypothesis. Equation 3 (energy function) computes the unnormalized attention weights as the similarity of hidden states of the premise and hypothesis.

(3) eij = u^Ti v^j, i [1, n], j [1, m]

where u^i and v^j are the hidden representations of u and v respectively which are computed earlier in Equations 1 and 2. Next, for each word in either premise or hypothesis, the relevant semantics in the other sentence is extracted and composed according to eij. Equations 4 and 5 provide formal and specific details of this procedure.

(4) u~i = sum(exp(eij) / sum(exp(eik))) * uj, i [1, n]

(5) v~j = sum(exp(eij) / sum(exp(ekj))) * ui, j [1, m]

where u~i represents the extracted relevant information of v^ by attending to u^i while v~j represents the extracted relevant information of u^ by attending to v^j. Next, it passes the enriched information through a projector layer which produce the final output of attention stage. Equations 6 and 7 formally represent this process.

(6) ai = [ui, u~i, ui u~i, ui u~i] ; pi = ReLU(Wpai + bi)

(7) bj = [vj, v~j, vj v~j, vj v~j] ; qj = ReLU(Wqbj + byj)

Here stands for element-wise product while Wp, Wq R4d×d and bp, by Rd are the trainable weights and biases of the projector layer respectively. p and q indicate the output of attention de- vision for premise and hypothesis respectively.

5.1.3 Inference

During this phase, it uses another BiLSTM to aggregate the two sequences of computed matching

(8) p^ = BiLSTM(p)

(9) q^ = BiLSTM(q)

where p^ Rn×2d and q^ Rm×2d are the reading sequences of p and q respectively. Finally the concatenation max and average pooling of p^ and q^ are pass through a multilayer perceptron (MLP) classifier that includes a hidden layer with tanh activation and softmax output layer. The model is trained in an end-to-end manner.

5.2 Attention Study

Here we provide more examples on the NLI task which intend to examine specific behavior in this model. Such examples indicate interesting observation that we can analyze them in the future works. Table 1 shows the list of all example.

Table 1: Examples along their gold labels, ESIM-50 predictions and study categories.

Premise	Hypothesis	Gold	Prediction	Category
Six men, two with shirts and four without, have taken a break from their work on a building.	Seven men, two with shirts and four without, have taken a break from their work on a building.	Contradiction	Contradiction	Counting
two men with shirts and four men without, have taken a break from their work on a building.	Six men, two with shirts and four without, have taken a break from their work on a building.	Entailment	Entailment	Counting
Six men, two with shirts and four without, have taken a break from their work on a building.	Six men, four with shirts and two without, have taken a break from their work on a building.	Contradiction	Contradiction	Counting
A man just ordered a book from amazon.	A man ordered a book yesterday.	Neutral	Neutral	Chronology
A man ordered a book from amazon 30 hours ago.	A man ordered a book yesterday.	Entailment	Entailment	Chronology