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

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

1 Model

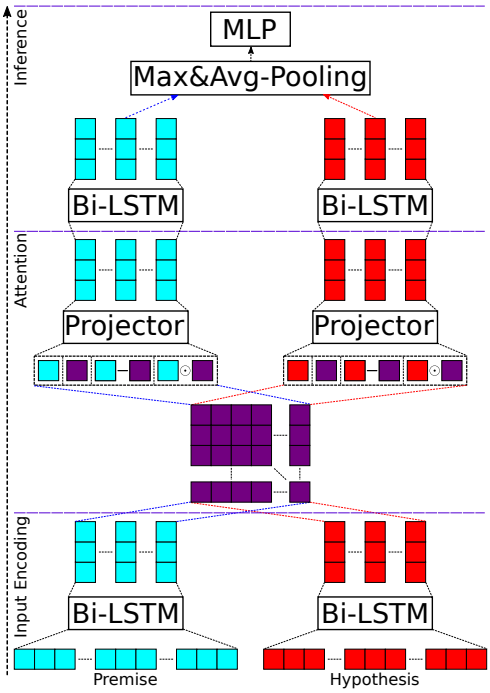

Figure 1: A high-level view of ESIM model.

In this section we describe the ESIM model. We divide ESIM to three main parts: 1) input encoding, 2) attention, and 3) inference. Figure 1 demonstrates a high-level view of the ESIM framework.

Let $u = [u_1, \cdots, u_n]$ and $v = [v_1, \cdots, v_m]$ be the given premise with length n and hypothesis with length m respectively, where $u_i, v_j \in \mathbb{R}^r$ 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.

1.1 Input Encoding

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

$$\hat{u} = \textit{BiLSTM}(u) \tag{1}$$

$$\hat{v} = \textit{BiLSTM}(v) \tag{2}$$

where $\hat{u} \in \mathbb{R}^{n \times 2d}$ and $\hat{v} \in \mathbb{R}^{m \times 2d}$ are the reading sequences of u and v respectively.

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.

$$e_{ij} = \hat{u}_i \hat{v}_j^T, \quad i \in [1, n], j \in [1, m] \tag{3}$$

where $\hat{u}_i$ and $\hat{v}_j$ 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 e_{ij}. Equations 4 and 5 provide formal and specific details of this procedure.

$$\tilde{u}_i = \sum_{j=1}^{m} \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{ik})} \hat{v}_j, \quad i \in [1, n] \tag{4}$$

$$\tilde{v}_j = \sum_{i=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{kj})} \hat{u}_i, \quad j \in [1, m] \tag{5}$$

where $\tilde{u}_i$ represents the extracted relevant information of $\hat{v}$ by attending to $\hat{u}_i$ while $\tilde{v}_j$ represents

ID	Premise	Hypothesis	Gold	Prediction	Category
1	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
2	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
3	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
4	A man just ordered a book from amazon.	A man ordered a book yesterday.	Neutral	Neutral	Chronology
5	A man ordered a book from amazon 30 hours ago.	A man ordered a book yesterday.	Entailment	Entailment	Chronology

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

the extracted relevant information of $\hat{u}$ by attending to $\hat{v}_j$. 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.

$$\begin{aligned} a_i &= [\hat{u}_i, \tilde{u}_i, \hat{u}_i - \tilde{u}_i, \hat{u}_i \odot \tilde{u}_i] \\ p_i &= \textit{ReLU}(W_p a_i + b_p) \end{aligned} \tag{6}$$

$$\begin{aligned} b_j &= [\hat{v}_j, \tilde{v}_j, \hat{v}_j - \tilde{v}_j, \hat{v}_j \odot \tilde{v}_j] \\ q_j &= \textit{ReLU}(W_p b_j + b_p) \end{aligned} \tag{7}$$

Here $\odot$ stands for element-wise product while $W_p \in \mathbb{R}^{8d \times d}$ and $b_p \in \mathbb{R}^d$ are the trainable weights and biases of the projector layer respectively. p and q indicate the output of attention devision for premise and hypothesis respectively.

1.3 Inference

During this phase, it uses another BiLSTM to aggregate the two sequences of computed matching vectors, p and q from the attention stage (Equations 8 and 9).

$$\hat{p} = \textit{BiLSTM}(p) \tag{8}$$

$$\hat{q} = \textit{BiLSTM}(q) \tag{9}$$

where $\hat{p} \in \mathbb{R}^{n \times 2d}$ and $\hat{q} \in \mathbb{R}^{m \times 2d}$ are the reading sequences of p and q respectively. Finally the concatenation max and average pooling of $\hat{p}$ and $\hat{q}$ 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.

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.

3 LSTM Gating Signal

Finally, Figure 8 depicts the backward LSTM gating signals study.

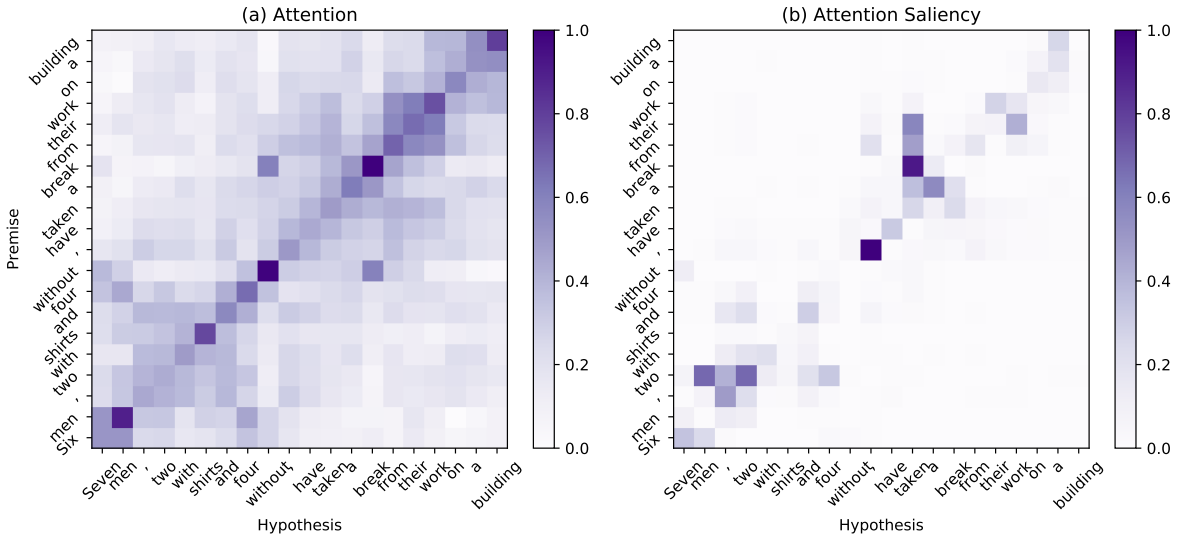

Figure 2: Normalized attention (a) and saliency attention (b) visualizations of Example 1. The gold relationship for this example is Contradiction. ESIM-50 also predicts Contradiction for this example.

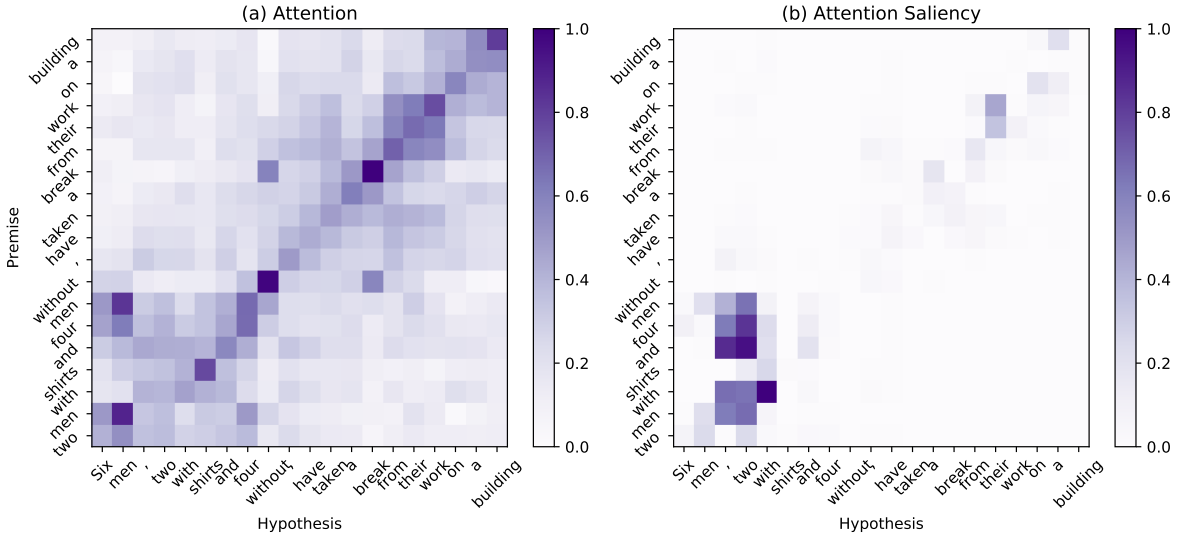

Figure 3: Normalized attention (a) and saliency attention (b) visualizations of Example 2. The gold relationship for this example is Entailment. ESIM-50 also predicts Entailment for this example.

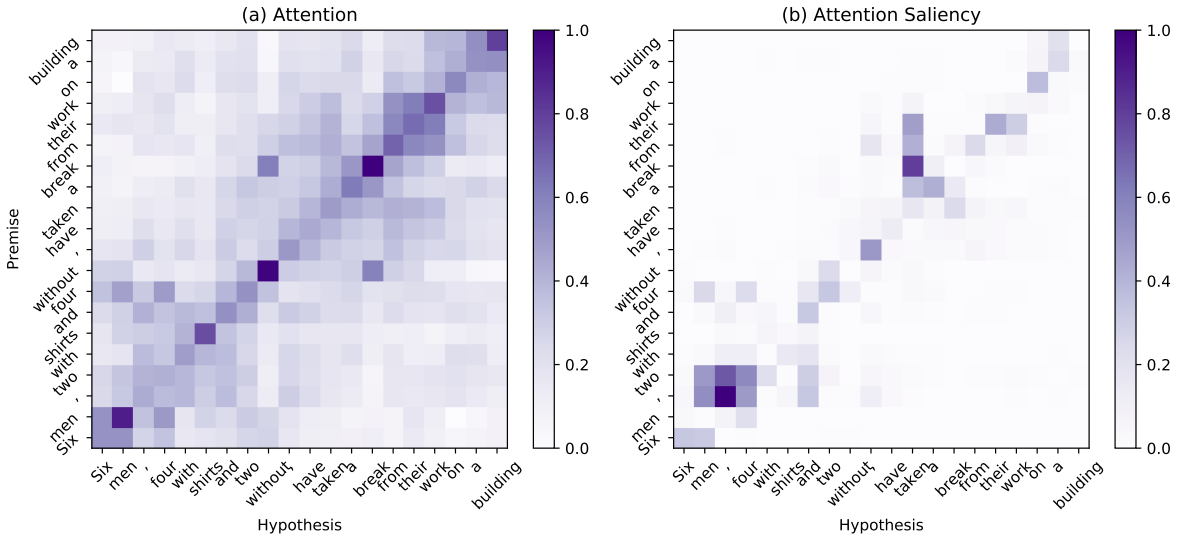

Figure 4: Normalized attention (a) and saliency attention (b) visualizations of Example 3. The gold relationship for this example is Contradiction. ESIM-50 also predicts Contradiction for this example.

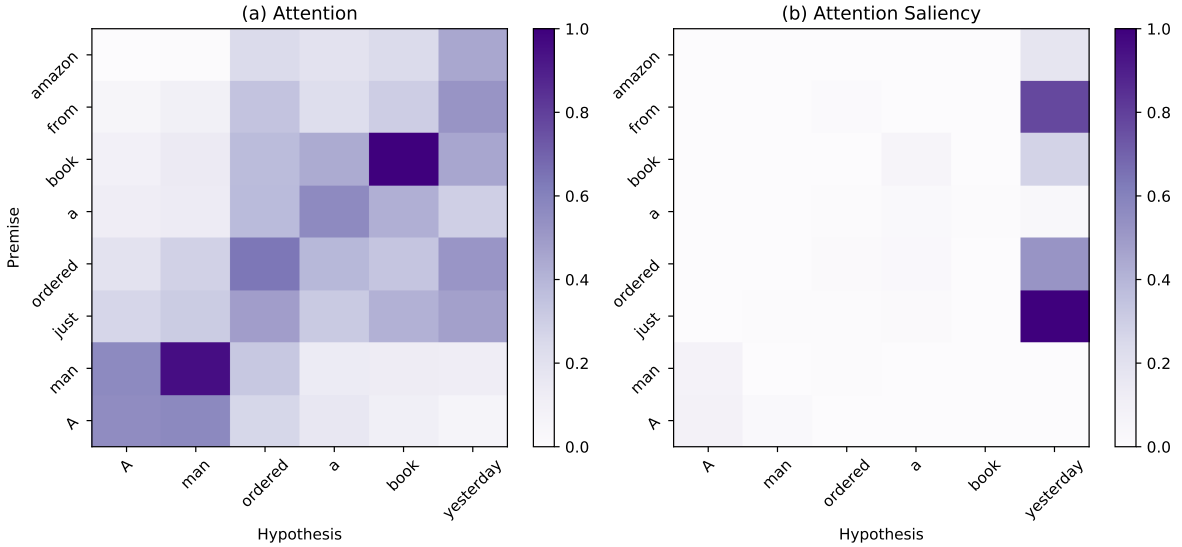

Figure 5: Normalized attention (a) and saliency attention (b) visualizations of Example 4. The gold relationship for this example is Neutral. ESIM-50 also predicts Neutral for this example.

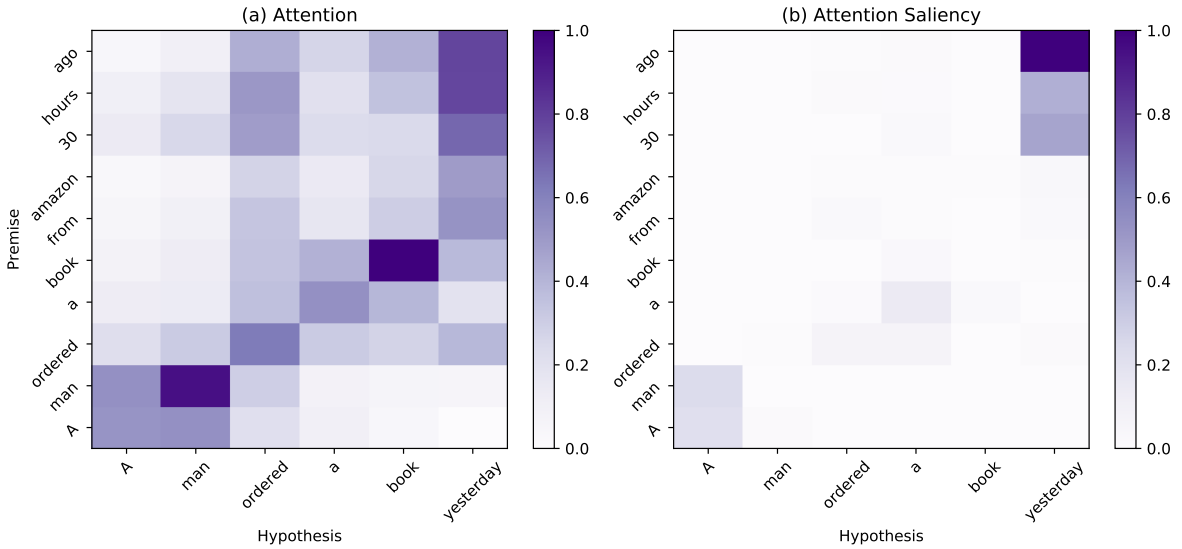

Figure 6: Normalized attention (a) and saliency attention (b) visualizations of Example 5. The gold relationship for this example is Entailment. ESIM-50 also predicts Entailment for this example.

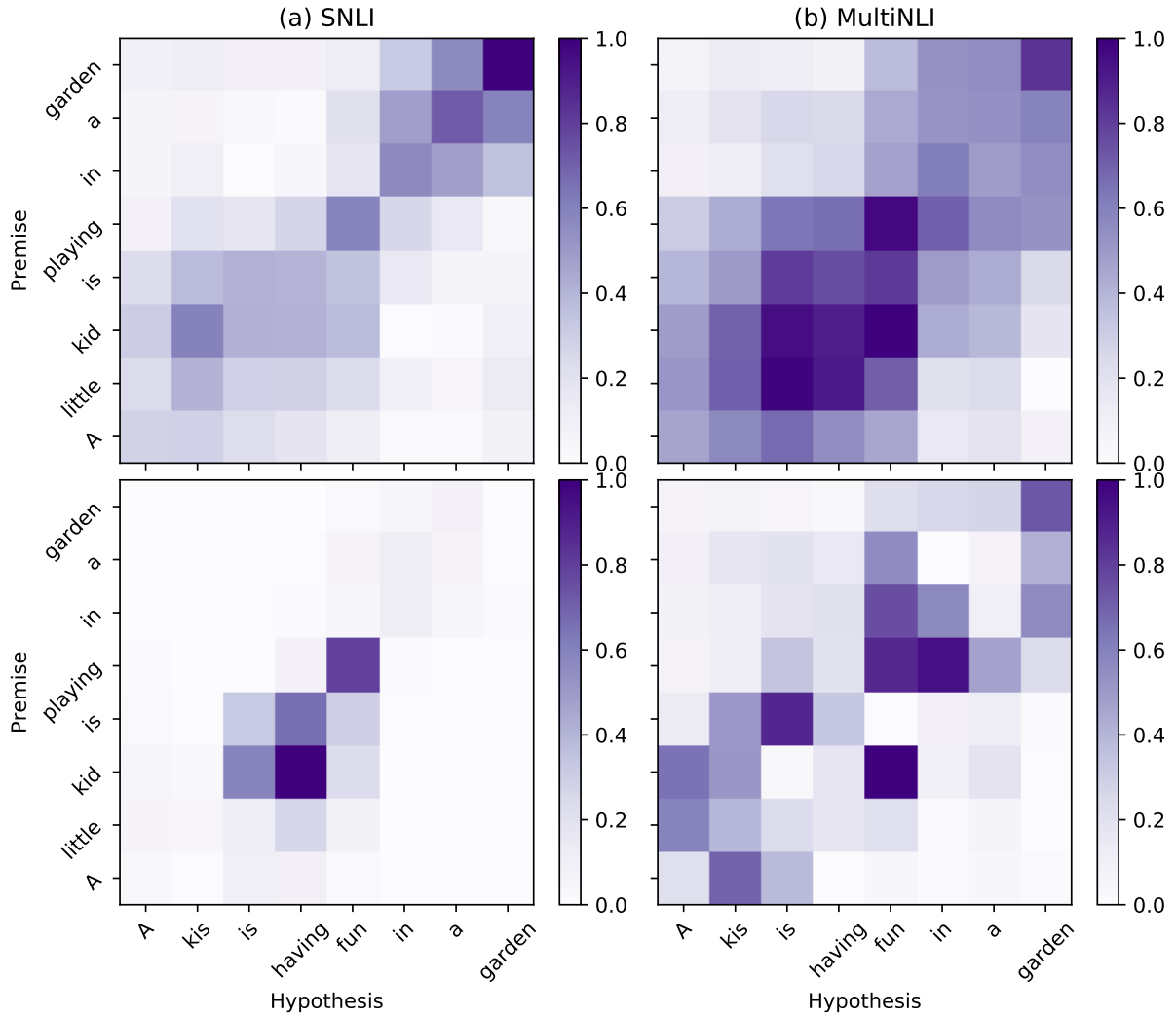

Figure 7: Normalized attention and saliency attention visualizations of an example (h3 in the main paper) for ESIM-50 learned on SNLI (a) and learned on MultiNLI (b). Top plots indicates the attention visualization and bottom ones shows the saliency attention visualization. Both systems correctly predict entailment for this example.

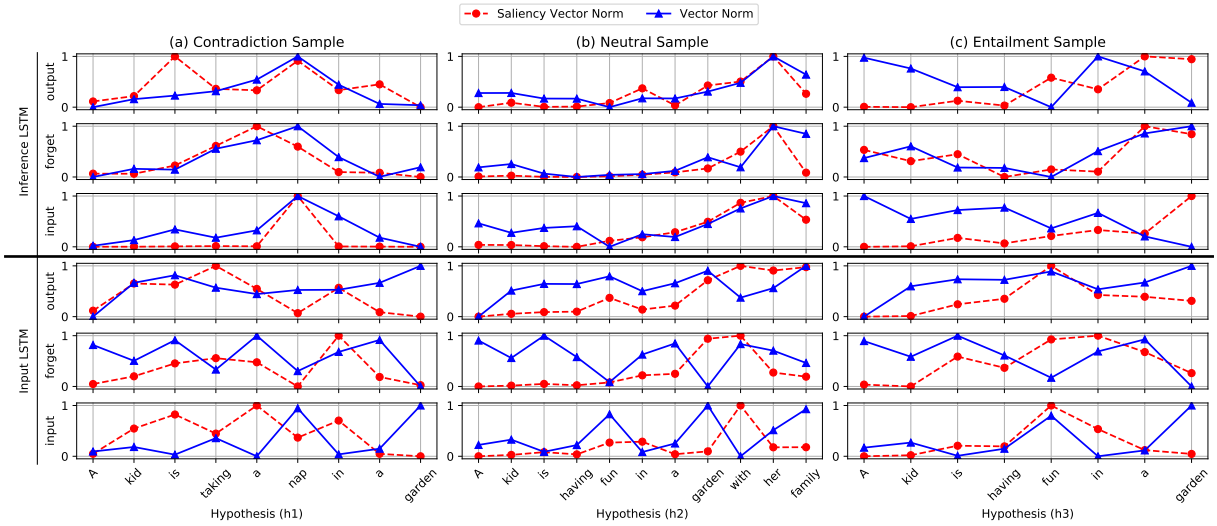

Figure 8: Normalized signal and saliency norms for the input and inference LSTMs (backward) for three examples, one for each column. The bottom (top) three rows show the signals of the input (inference) LSTM, where each row shows one of the three gates (input, forget and output).