DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference: Supplementary Materials[*]

Reza Ghaeini[1 †], **Sadid A. Hasan**[2], **Vivek Datla**[2], **Joey Liu**[2], **Kathy Lee**[2], **Ashequl Qadir**[2], **Yuan Ling**[2], **Aaditya Prakash**[2], **Xiaoli Z. Fern**[1] **and Oladimeji Farri**[2]
[1]Oregon State University, Corvallis, OR, USA
[2]Artificial Intelligence Laboratory, Philips Research North America, Cambridge, MA, USA
{ghaeinim,xfern}@eecs.oregonstate.edu
{sadid.hasan,vivek.datla,joey.liu,kathy.lee_1,ashequl.qadir}@philips.com
{yuan.ling,aaditya.prakash,dimeji.farri}@philips.com

1 Ensemble Strategy Study

We use the following configurations in our ensemble model study:

- DR-BiLSTM (with different initialization seeds): here, we consider 6 DR-BiLSTMs with different initialization seeds.

- *tanh*-Projection: same configuration as DR-BiLSTM, but we use *tanh* instead of *ReLU* as the activation function in Equations 6 and 7 in the main paper:

$$p_i = tanh(W_p a_i + b_p) \tag{1}$$

$$q_j = tanh(W_p b_j + b_p) \tag{2}$$

- DR-BiLSTM (with 1 round of dependent reading): same configuration as DR-BiLSTM, but we do not use dependent reading during the inference process. In other words, we use $\tilde{p} = \bar{p}$ and $\tilde{q} = \bar{q}$ instead of Equations 10 and 11 in the main paper respectively.

- DR-BiLSTM (with 3 rounds of dependent reading): same configuration as the above, but we use 3 rounds of dependent reading. Formally, we replace Equations 1 and 2 in the main paper with the following equations respectively:

$$\begin{aligned} -, s_v &= \mathit{BiLSTM}(v, 0) \\ -, s_{vu} &= \mathit{BiLSTM}(u, s_v) \\ -, s_{vuv} &= \mathit{BiLSTM}(v, s_{vu}) \\ \hat{u}, - &= \mathit{BiLSTM}(u, s_{vuv}) \end{aligned} \tag{3}$$

$$\begin{aligned} -, s_u &= \mathit{BiLSTM}(u, 0) \\ -, s_{uv} &= \mathit{BiLSTM}(v, s_u) \\ -, s_{uvu} &= \mathit{BiLSTM}(u, s_{uv}) \\ \hat{v}, - &= \mathit{BiLSTM}(v, s_{uvu}) \end{aligned} \tag{4}$$

Our final ensemble model, DR-BiLSTM (Ensemble) is the combination of the following 6 models: tanh-Projection, DR-BiLSTM (with 1 round of dependent reading), DR-BiLSTM (with 3 rounds of dependent reading), and 3 DR-BiLSTMs with different initialization seeds.

We also experiment with majority voting and averaging the probability distribution strategies for ensemble models using the same set of models as our weighted averaging ensemble method (as described above). Figure 1 shows the behavior of the majority voting strategy with different number of models. Interestingly, the best development accuracy is also observed using 6 individual models including tanh-Projection, DR-BiLSTM (with 1 round of dependent reading), DR-BiLSTM (with 3 rounds of dependent reading), and 3 DR-BiLSTMs with varying initialization seeds that are different from our DR-BiLSTM (Ensemble) model.

We should note that our weighted averaging ensemble strategy performs better than the majority voting method in both development set and test set of SNLI, which indicates the effectiveness of our approach. Furthermore, our method could show more consistent behavior for training and test sets when we increased the number of models (see Figure 2 in Section 4.3 of the main paper). According to our observations, averaging the probability distributions fails to improve the development set accuracy using two and three models, so we did not study it further.

[*] ArXiv version of this work can be found here (arxiv.org/pdf/1802.05577.pdf).

[†] This work was conducted as part of an internship program at Philips Research.

Original Sentence	Corrected Sentence
Froends ride in an open top vehicle together.	***Friends*** ride in an open top vehicle together.
A middle ***easten*** store.	A middle ***eastern*** store.
A woman is looking at a ***phtographer***	A woman is looking at a ***photographer***
The mother and daughter are ***fighitn***.	The mother and daughter are ***fighting***.
Two ***kiled*** men hold bagpipes	Two ***killed*** men hold bagpipes
A woman escapes a from a hostile ***enviroment***	A woman escapes a from a hostile ***environment***
Two ***daschunds*** play with a red ball	Two ***dachshunds*** play with a red ball
A black dog is running through a ***marsh-like*** area.	A black dog is running through a ***marsh like*** area.
a singer wearing a ***jacker*** performs on stage	a singer wearing a ***jacket*** performs on stage
There is a ***sculture***	There is a ***sculpture***
Taking a ***neverending*** break	Taking a ***never ending*** break
The woman has sounds ***emanting*** from her mouth.	The woman has sounds ***emanating*** from her mouth.
the lady is ***shpping***	the lady is ***shopping***
A Bugatti and a ***Lambourgini*** compete in a road race.	A Bugatti and a ***Lamborghini*** compete in a road race.

Table 1: Examples of original sentences that contain erroneous words (misspelled) in the test set of SNLI along with their corrected counterparts. Erroneous words are shown in ***bold and italic***.

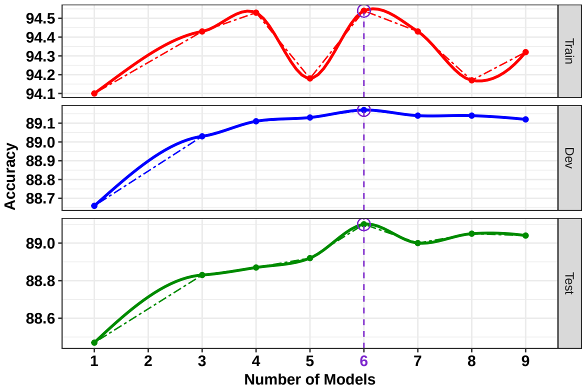


Figure 1: Performance of n ensemble models using majority voting on natural language inference reported for training set (red, top), development set (blue, middle), and test set (green, bottom) of SNLI. The best performance on development set is used as the criteria to determine the final ensemble. The best performance on development set is observed using 6 models.

2 Preprocessing Study

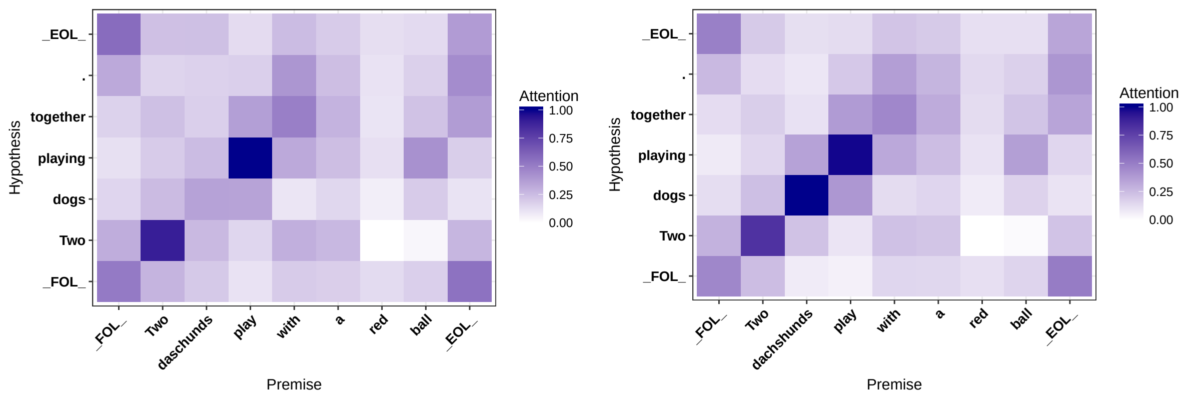
Table 1 shows some erroneous sentences from the SNLI test set along with their corrected equivalents (after preprocessing). Furthermore, we show the energy function (Equation 3 in the main paper) visualizations of 6 examples from the aforementioned data samples in Figures 2, 3, 4, 5, 6, and 7. Each figure presents the visualization of an original erroneous sample along its corrected version. These figures clearly illustrate that fixing the erroneous words leads to producing correct attentions over the sentences. This can be observed by comparing the attention for the erroneous words and corrected words, e.g. "daschunds" and "dachshunds" in the premise of Figures 2 and 3. Note that we add two dummy notations (i.e. _FOL_, and _EOL_) to all sentences which indicate their beginning and end.

3 Category Study

Here we investigate the normalized attention weights of DR-BiLSTM and ESIM for 3 samples that belong to Negation and/or Quantifier categories (Figures 8, 9, and 10). Each figure illustrates the normalized energy function of DR-BiLSTM (left diagram) and ESIM (right diagram) respectively. Provided figures indicate that ESIM assigns somewhat similar attention to most of the pairs while DR-BiLSTM focuses on specific parts of the given premise and hypothesis.

4 Attention Study

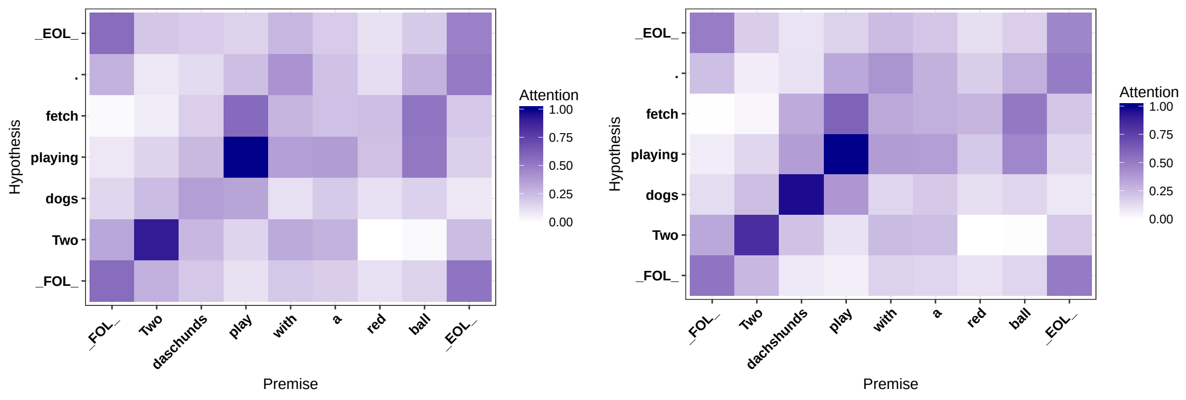
In this section, we show visualizations of 18 samples of normalized attention weights (energy function, see Equation 3 in the main paper). Each column in Figures 11, 12, and 13, represents three data samples that share the same premise but differ in hypothesis. Also, each row is allocated to a specific logical relationship (Top: Entailment, Middle: Neutral, and Bottom: Contradiction). DR-BiLSTM classifies all data samples reported in these figures correctly.



(a) Erroneous sample (daschunds in premise).

(b) Fixed sample (dachshunds in premise).

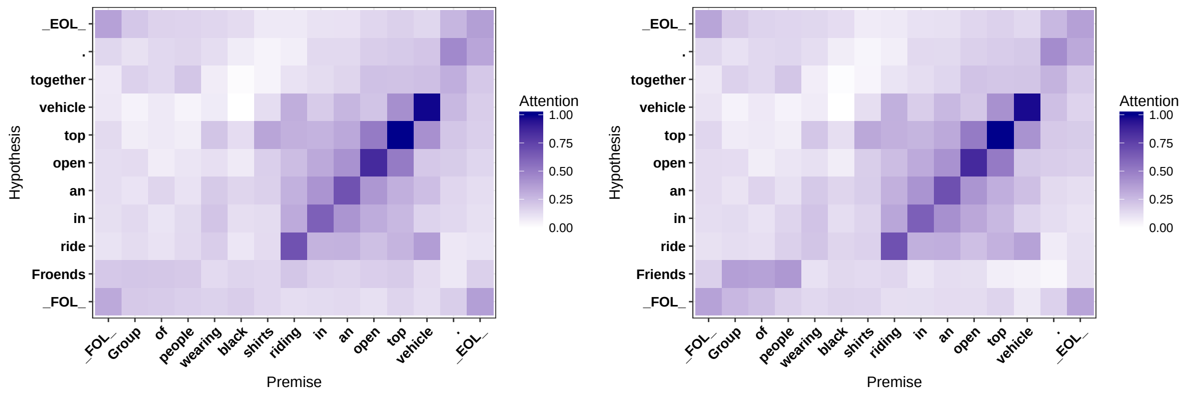
Figure 2: Visualization of the energy function for one erroneous sample (a) and the fixed sample (b). The gold label is *Entailment*. Our model returns *Contradiction* for the erroneous sample, but correctly classifies the fixed sample.



(a) Erroneous sample (daschunds in premise).

(b) Fixed sample (dachshunds in premise).

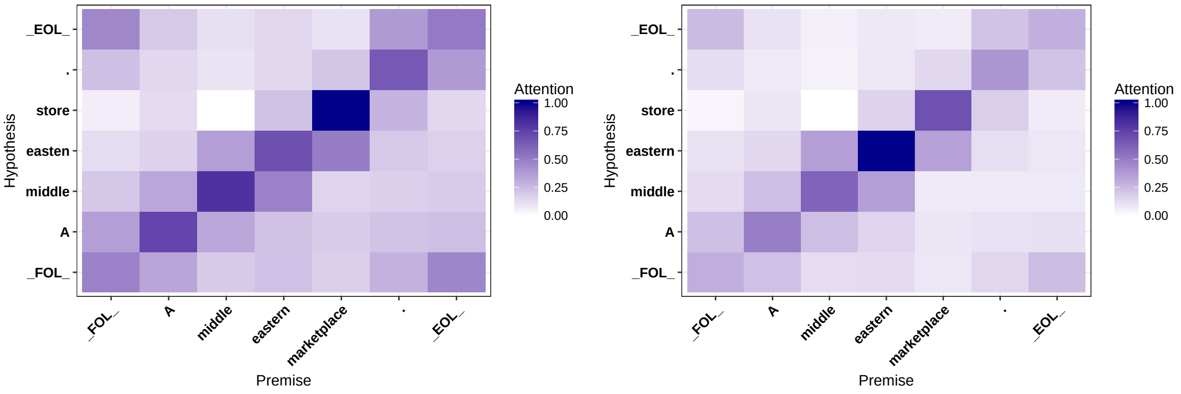
Figure 3: Visualization of the energy function for one erroneous sample (a) and the fixed sample (b). The gold label is *Neutral*. Our model returns *Contradiction* for the erroneous sample, but correctly classifies the fixed sample.



(a) Erroneous sample (Froends in hypothesis).

(b) Fixed sample (Friends in hypothesis).

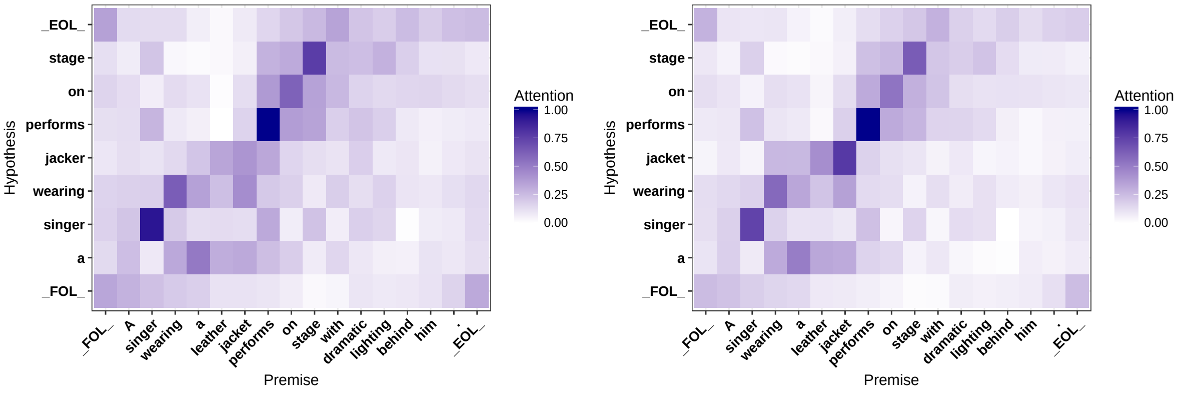
Figure 4: Visualization of the energy function for one erroneous sample (a) and the fixed sample (b). The gold label is *Neutral*. Our model returns *Entailment* for the erroneous sample, but correctly classifies the fixed sample.



(a) Erroneous sample (easten in hypothesis).

(b) Fixed sample (eastern in hypothesis).

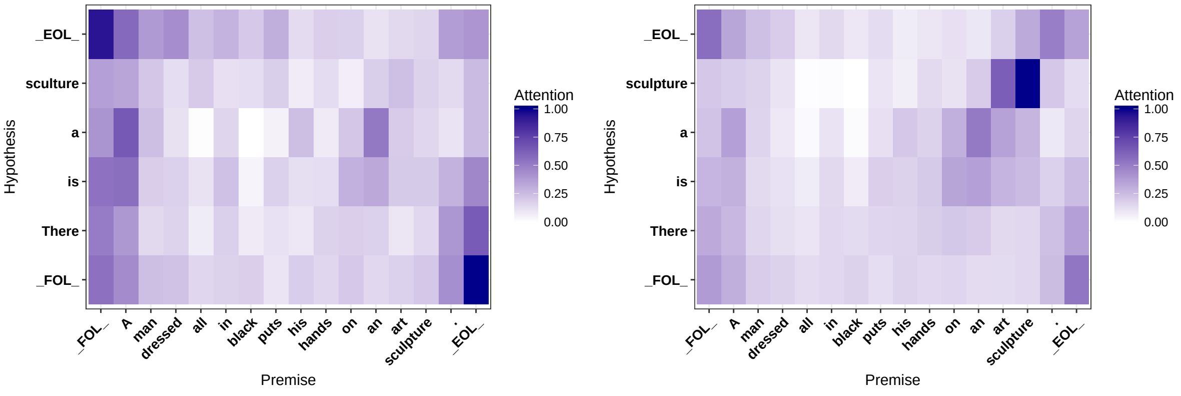
Figure 5: Visualization of the energy function for one erroneous sample (a) and the fixed sample (b). The gold label is *Entailment*. Our model returns *Contradiction* for the erroneous sample, but correctly classifies the fixed sample.



(a) Erroneous sample (jacker in hypothesis).

(b) Fixed sample (jacket in hypothesis).

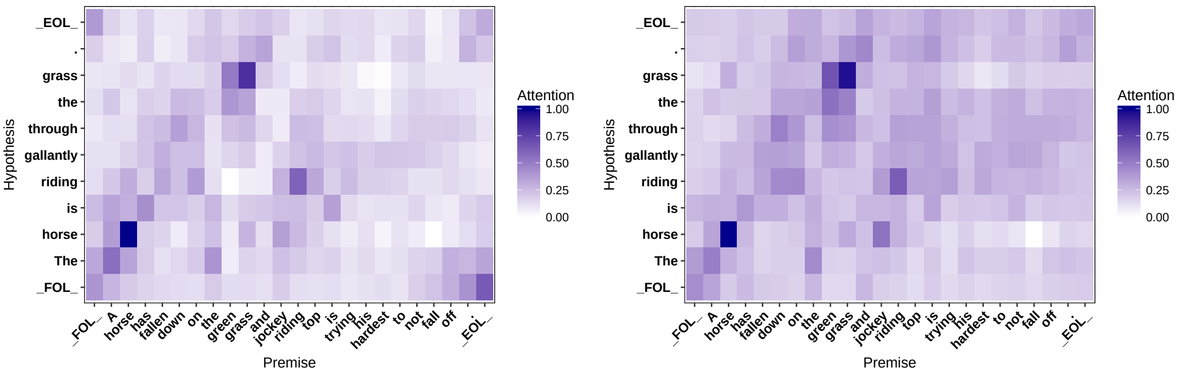
Figure 6: Visualization of the energy function for one erroneous sample (a) and the fixed sample (b). The gold label is *Entailment*. Our model returns *Neutral* for the erroneous sample, but correctly classifies the fixed sample.



(a) Erroneous sample (sculture in hypothesis).

(b) Fixed sample (sculpture in hypothesis).

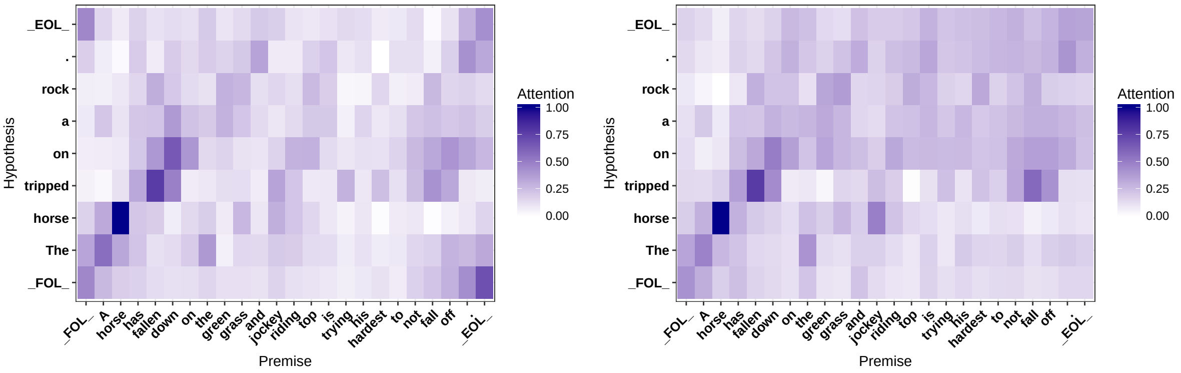
Figure 7: Visualization of the energy function for one erroneous sample (a) and the fixed sample (b). The gold label is *Entailment*. Our model returns *Neutral* for the erroneous sample, but correctly classifies the fixed sample.



(a) Normalized attention of DR-BiLSTM.

(b) Normalized attention of ESIM

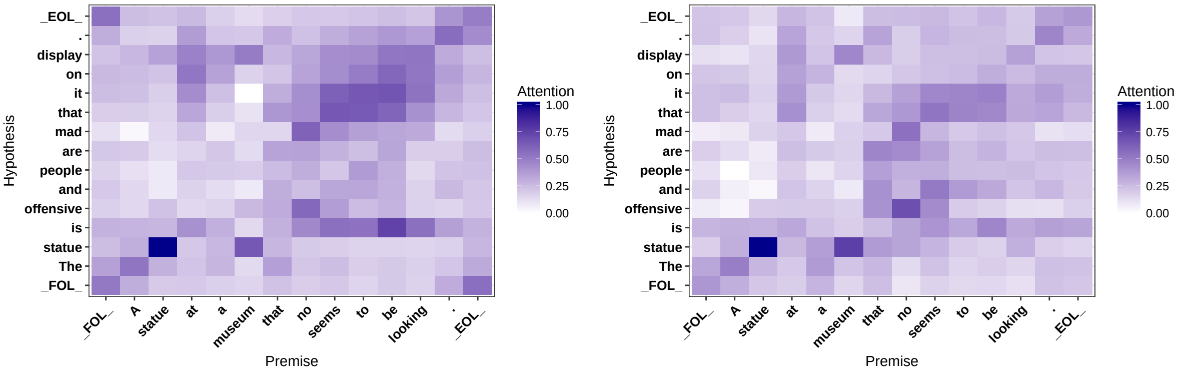
Figure 8: Visualization of the normalized attention weights of DR-BiLSTM (a) and ESIM (b) models for one sample from the SNLI test set. This sample belongs to the Negation category. The gold label is *Contradiction*. Our model returns *Contradiction* while ESIM returns *Entailment*.



(a) Normalized attention of DR-BiLSTM.

(b) Normalized attention of ESIM

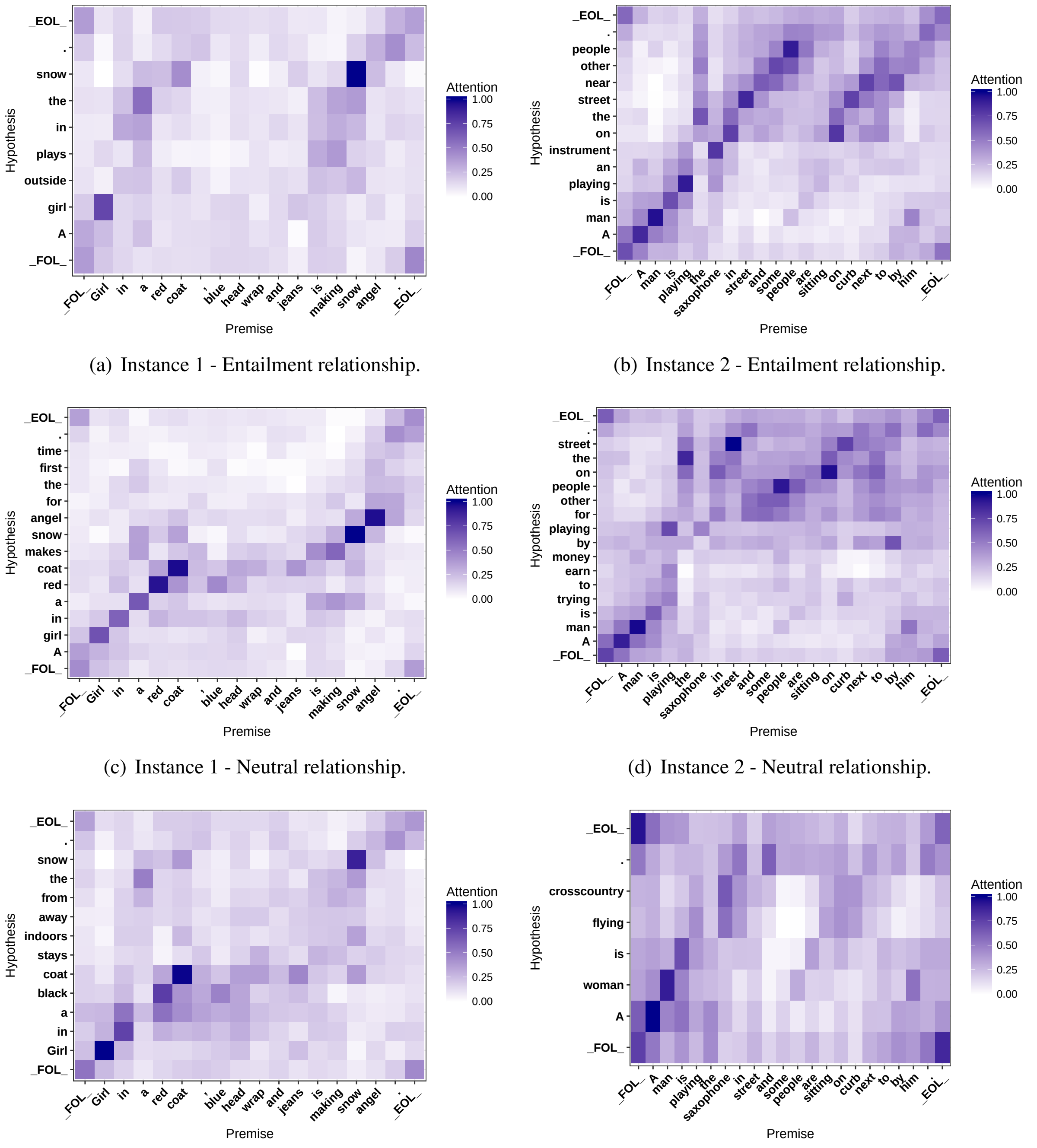
Figure 9: Visualization of the normalized attention weights of DR-BiLSTM (a) and ESIM (b) models for one sample from the SNLI test set. This sample belongs to the Negation category. The gold label is *Contradiction*. Our model returns *Contradiction* while ESIM returns *Entailment*.



(a) Normalized attention of DR-BiLSTM.

(b) Normalized attention of ESIM

Figure 10: Visualization of the normalized attention weights of DR-BiLSTM (a) and ESIM (b) models for one sample from the SNLI test set. This sample belongs to both Negation and Quantifier categories. The gold label is *Neutral*. Our model returns *Neutral* while ESIM returns *Contradiction*.



(a) Instance 1 - Entailment relationship.

(b) Instance 2 - Entailment relationship.

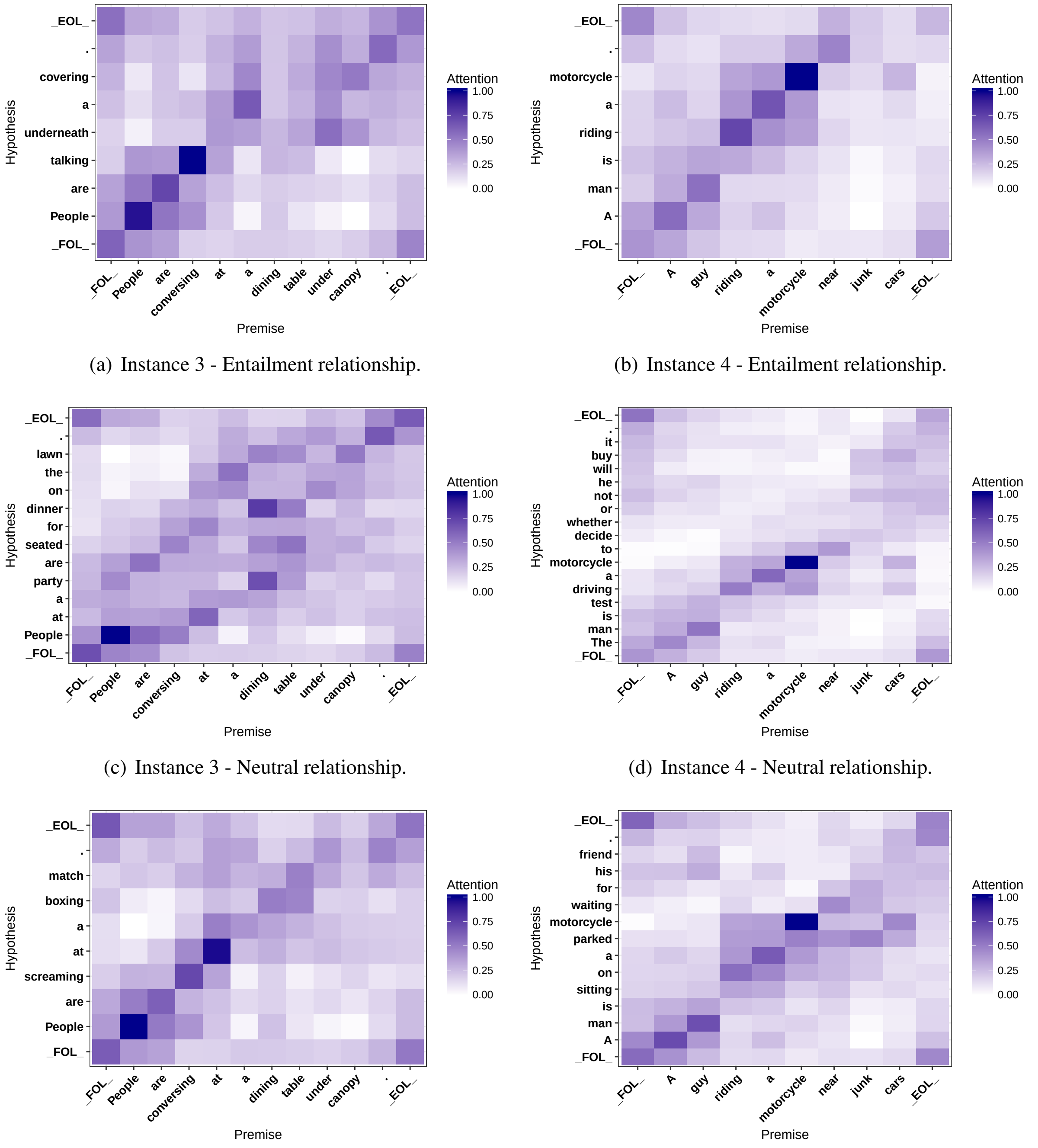
(c) Instance 1 - Neutral relationship.

(d) Instance 2 - Neutral relationship.

(e) Instance 1 - Contradiction relationship.

(f) Instance 2 - Contradiction relationship.

Figure 11: Normalized attention weights for 6 data samples from the test set of SNLI dataset. (a,c,e) and (b,d,f) represent the normalized attention weights for *Entailment*, *Neutral*, and *Contradiction* logical relationships of two premises (Instance 1 and 2) respectively. Darker color illustrates higher attention.



(a) Instance 3 - Entailment relationship.

(b) Instance 4 - Entailment relationship.

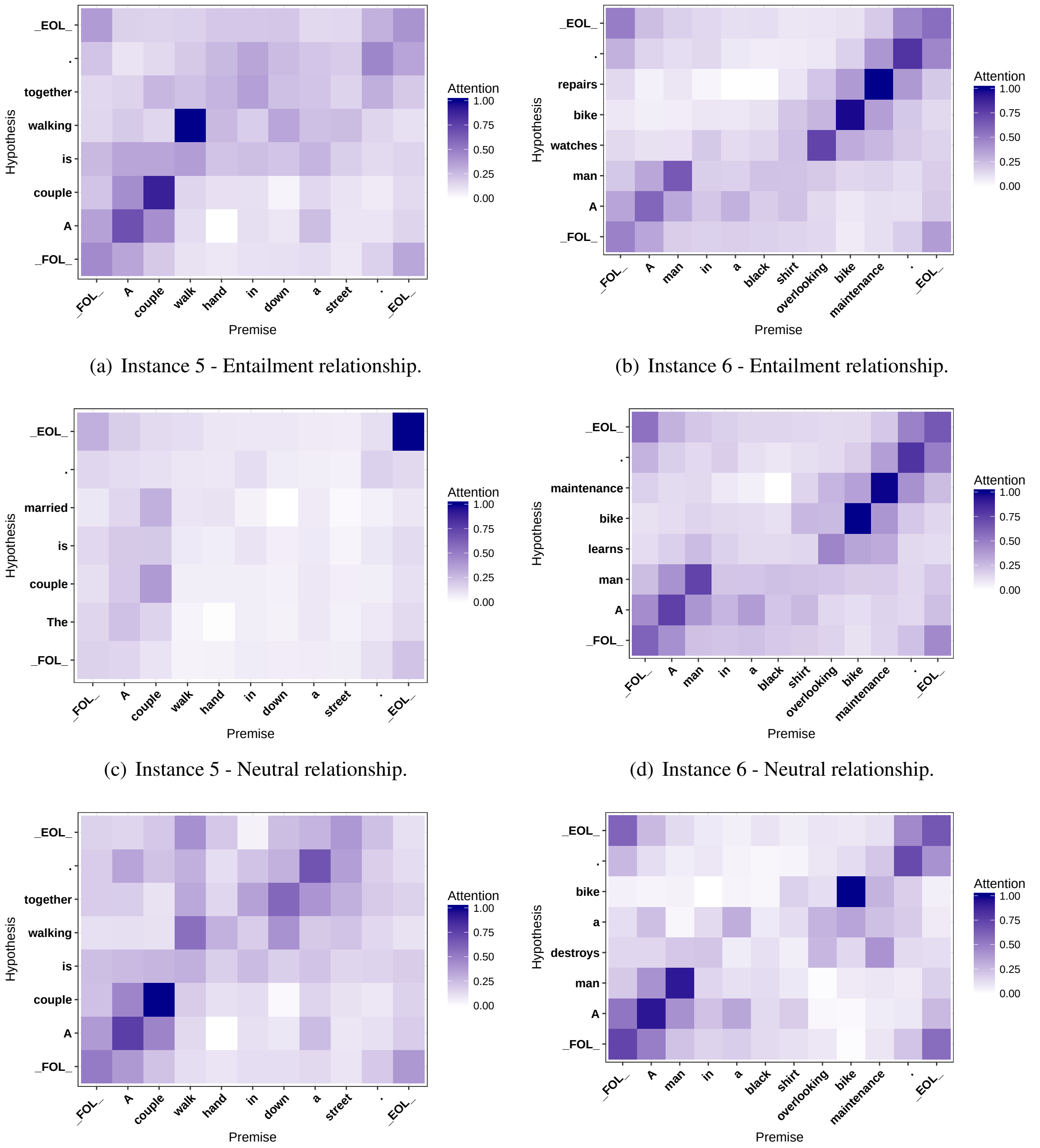
(c) Instance 3 - Neutral relationship.

(d) Instance 4 - Neutral relationship.

(e) Instance 3 - Contradiction relationship.

(f) Instance 4 - Contradiction relationship.

Figure 12: Normalized attention weights for 6 data samples from the test set of SNLI dataset. (a,c,e) and (b,d,f) represent the normalized attention weights for *Entailment*, *Neutral*, and *Contradiction* logical relationships of two premises (Instance 3 and 4) respectively. Darker color illustrates higher attention.



(a) Instance 5 - Entailment relationship.

(b) Instance 6 - Entailment relationship.

(c) Instance 5 - Neutral relationship.

(d) Instance 6 - Neutral relationship.

(e) Instance 5 - Contradiction relationship.

(f) Instance 6 - Contradiction relationship.

Figure 13: Normalized attention weights for 6 data samples from the test set of SNLI dataset. (a,c,e) and (b,d,f) represent the normalized attention weights for *Entailment*, *Neutral*, and *Contradiction* logical relationships of two premises (Instance 5 and 6) respectively. Darker color illustrates higher attention.