[20] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., bernardi, R., and Zamparelli, R. (2014). A sick cure for the evaluation of compositional distributional semantic models.

[21] Nie, Y. and Bansal, M. (2017). Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45. Association for Computational Linguistics.

[22] Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933.

[23] Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., and Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1802.08129.

[24] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

[25] Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kociský, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664.

[26] Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

[27] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

Appendices

A List of templates to filter uninformative explanations

General templates

"<premise>"

"<hypothesis>"

"<hypothesis> <premise>"

"<premise> <hypothesis>"

"Sentence 1 states <premise>. Sentence 2 is stating <hypothesis>"

"Sentence 2 states <hypothesis>. Sentence 1 is stating <premise>"

"There is <hypothesis>"

"There is <premise>"

Entailment templates

"<premise> implies <hypothesis>"

"If <premise> then <hypothesis>"

"<premise> would imply <hypothesis>"

"<hypothesis> is a rephrasing of <premise>"

"<premise> is a rephrasing of <hypothesis>"

"In both sentences <hypothesis>"

"<premise> would be <hypothesis>"

"<premise> can also be said as <hypothesis>"

"<hypothesis> can also be said as <premise>"

"<hypothesis> is a less specific rephrasing of <premise>"

"This clarifies that <hypothesis>"

"If <premise> it means <hypothesis>"

"<hypothesis> in both sentences"

"<hypothesis> in both"

"<hypothesis> is same as <premise>"

"<premise> is same as <hypothesis>"

"<premise> is a synonym of <hypothesis>"

"<hypothesis> is a synonym of <premise>".

Neutral templates

"Just because <premise> doesn't mean <hypothesis>"

"Cannot infer the <hypothesis>"

"One cannot assume <hypothesis>"

"One cannot infer that <hypothesis>"

"Cannot assume <hypothesis>"

"<premise> does not mean <hypothesis>"

"We don't know that <hypothesis>"

"The fact that <premise> doesn't mean <hypothesis>"

"The fact that <premise> does not imply <hypothesis>"

"The fact that <premise> does not always mean <hypothesis>"

"The fact that <premise> doesn't always imply<hypothesis>".

Contradiction templates

"In sentence 1 <premise> while in sentence 2 <hypothesis>"

"It can either be <premise> or <hypothesis>"

"It cannot be <hypothesis> if <premise>"

"Either <premise> or <hypothesis>"

"Either <hypothesis> or <premise>"

"<premise> and other <hypothesis>"

"<hypothesis> and other <premise>"

"<hypothesis> after <premise>"

"<premise> is not the same as <hypothesis>"

"<hypothesis> is not the same as <premise>"

"<premise> is contradictory to <hypothesis>"

"<hypothesis> is contradictory to <premise>"

"<premise> contradicts <hypothesis>"

"<hypothesis> contradicts <premise>"

"<premise> cannot also be <hypothesis>"

"<hypothesis> cannot also be <premise>"

"either <premise> or <hypothesis>"

"either <premise> or <hypothesis> not both at the same time"

"<premise> or <hypothesis> not both at the same time".

B Architecture of EXPLAINTHENPREDICTATTENTION

Our attention model EXPLAINTHENPREDICTATTENTION is composed of two identical but separate modules for premise and hypothesis. We fix the number of attended tokens at 84, the maximum length of a sentence in SNLI. We denote by h_t^p and h_t^h the bidirectional embeddings of the premise and hypothesis at timestep t. We denote by h_τ^{dec} the decoder hidden state at timestep τ, which we refer to as the context of the attention.

We use 3 couples of linear projections followed by $tanh$ non-linearities as follows:

We project each timestep of the encoder for premise and hypothesis:

$$proj1_t^p = tanh(W_p^1 h_t^p + b_p^1)$$

$$proj1_t^h = tanh(W_h^1 h_t^h + b_h^1).$$

We separately project the context vector, that is, the hidden vector of the decoder at each timestep, before doing its dot product with the tokens of the premise and hypothesis:

$$proj_\tau^{c,p} = tanh(W_p^c h_\tau^{dec} + b_p^c)$$

$$proj_\tau^{c,h} = tanh(W_h^c h_\tau^{dec} + b_h^c).$$

At each decoding timestep τ, we do the dot product between the projections of the context with all the timesteps of the premise and hypothesis, respectively:

$$\widetilde{w_t}^{p,\tau} =< proj_\tau^{c,p}, proj1_t^p >$$

$$\widetilde{w_t}^{h,\tau} =< proj_\tau^{c,h}, proj1_t^h > .$$

The final attention weights are computed from a softmax over the non-normalized weights:

$$w_t^{p,\tau} = Softmax(\widetilde{w_t}^{p,\tau})$$

$$w_t^{h,\tau} = Softmax(\widetilde{w_t}^{h,\tau}).$$

We use another couple of projections for the embeddings of the tokens of premise and hypothesis, before we apply the weighted sum.

$$proj2_t^p = tanh(W_p^2 h_t^p + b_p^2)$$

$$proj2_t^h = tanh(W_h^2 h_t^h + b_h^2).$$

Finally, we compute the weighted sums for premise and hypothesis:

$$p_\tau = \sum_t w_t^{p,\tau} proj2_t^p$$

$$h_\tau = \sum_t w_t^{h,\tau} proj2_t^h.$$

At each timestept τ, we concatenate p_τ and h_τ with the word embedding from the previous timestep $\tau - 1$ and give as input to our decoder.