Natural Language Inference: Attention

Christopher Potts

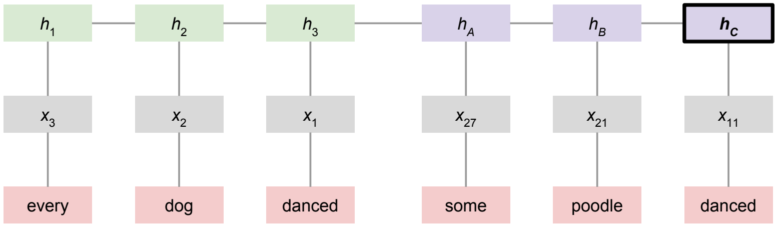
Stanford Linguistics

CS224u: Natural language understanding

Guiding ideas

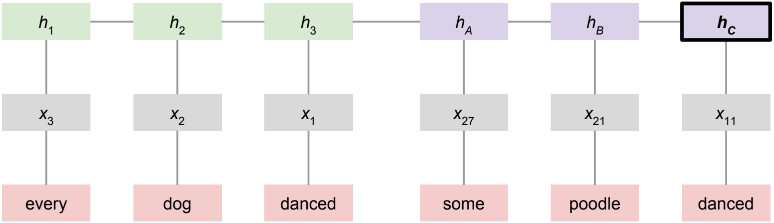
1. We need more connections between premise and hypothesis.
2. In processing the hypothesis, the model needs “reminders” of what the premise contained; the final premise hidden state isn’t enough.
3. Soft alignment between premise and hypothesis – a neural interpretation of an old idea in NLI.

Global attention



Global attention

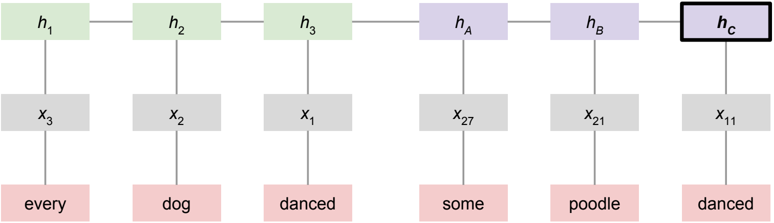
scores $\tilde{\alpha} = \left[\begin{array}{ccc} h_C^\top h_1 & h_C^\top h_2 & h_C^\top h_3 \end{array}\right]$



Global attention

attention weights $\alpha = \mathbf{softmax}(\tilde{\alpha})$

scores $\tilde{\alpha} = \left[\, h_C^\top h_1 \quad h_C^\top h_2 \quad h_C^\top h_3 \,\right]$

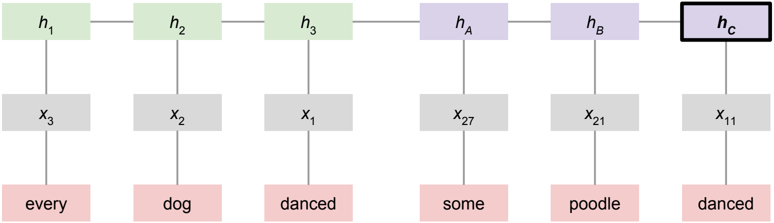


Global attention

context $$\kappa = \mathbf{mean}(\alpha_1 h_1, \alpha_2 h_2, \alpha_3 h_3)$$

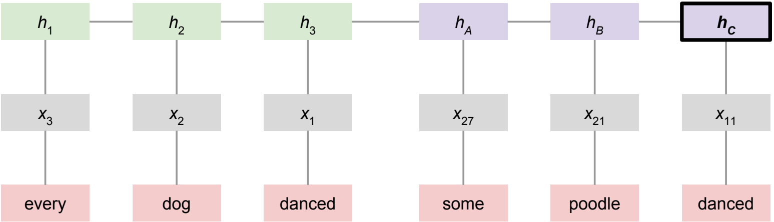
attention weights $$\alpha = \mathbf{softmax}(\tilde{\alpha})$$

scores $$\tilde{\alpha} = \begin{bmatrix} h_C^\top h_1 & h_C^\top h_2 & h_C^\top h_3 \end{bmatrix}$$



Global attention

attention combo	$\tilde{h} = \tanh([\kappa; h_C]W_\kappa)$
context	$\kappa = \textbf{mean}(\alpha_1 h_1, \alpha_2 h_2, \alpha_3 h_3)$
attention weights	$\alpha = \textbf{softmax}(\tilde{\alpha})$
scores	$\tilde{\alpha} = \left[\ h_C^\top h_1 \quad h_C^\top h_2 \quad h_C^\top h_3\ \right]$



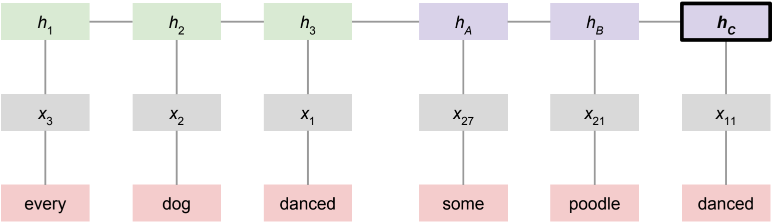
Global attention

attention combo $\tilde{h} = \tanh([\kappa; h_C]W_\kappa)$ or $\tilde{h} = \tanh(\kappa W_\kappa + h_C W_h)$

context $\kappa = \textbf{mean}(\alpha_1 h_1, \alpha_2 h_2, \alpha_3 h_3)$

attention weights $\alpha = \textbf{softmax}(\tilde{\alpha})$

scores $\tilde{\alpha} = \begin{bmatrix} h_C^\top h_1 & h_C^\top h_2 & h_C^\top h_3 \end{bmatrix}$



Global attention

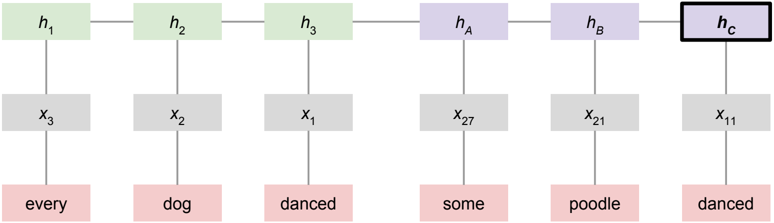
classifier $y = \textbf{softmax}(\tilde{h}W + b)$

attention combo $\tilde{h} = \tanh([\kappa; h_C]W_\kappa)$

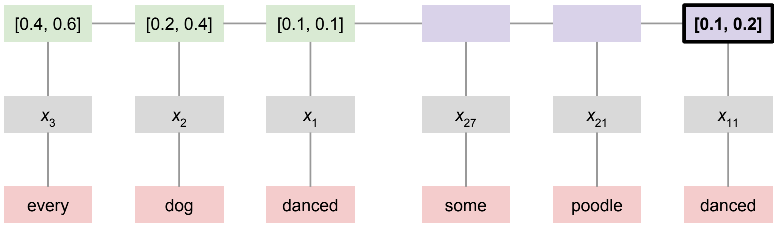
context $\kappa = \textbf{mean}(\alpha_1 h_1, \alpha_2 h_2, \alpha_3 h_3)$

attention weights $\alpha = \textbf{softmax}(\tilde{\alpha})$

scores $\tilde{\alpha} = \begin{bmatrix} h_C^\top h_1 & h_C^\top h_2 & h_C^\top h_3 \end{bmatrix}$

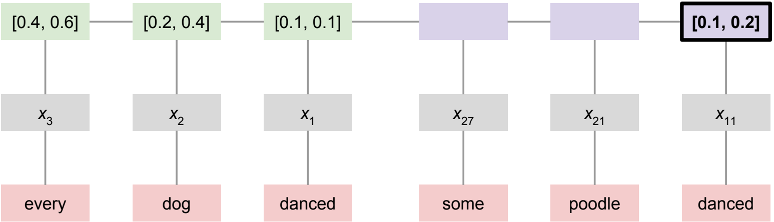


Global attention



Global attention

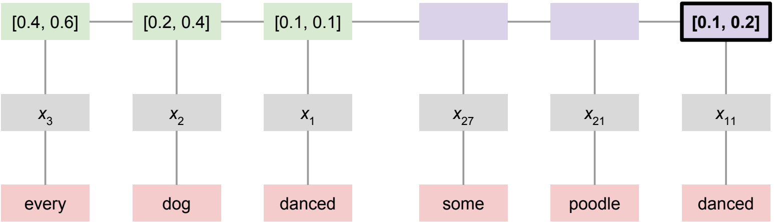
scores $\tilde{\alpha} = [0.16, 0.10, 0.03]$



Global attention

attention weights $\alpha = [0.35, 0.33, 0.31]$

scores $\tilde{\alpha} = [0.16, 0.10, 0.03]$

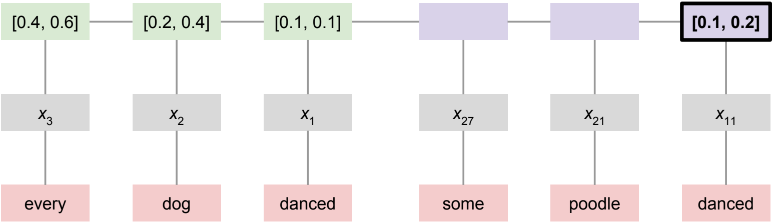


Global attention

context $\kappa = \textbf{mean}(.35 \cdot [.4, .6], .33 \cdot [.2, .4], .31 \cdot [.1, .1])$

attention weights $\alpha = [0.35, 0.33, 0.31]$

scores $\tilde{\alpha} = [0.16, 0.10, 0.03]$



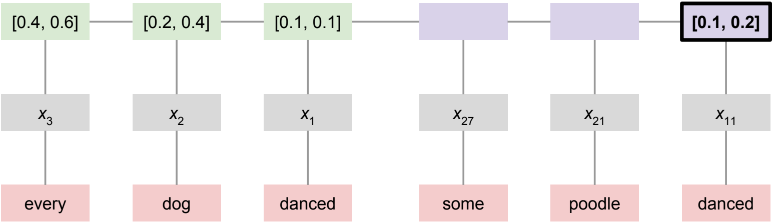
Global attention

attention combo $\tilde{h} = \tanh([0.07, 0.11, 0.1, 0.2]W_\kappa)$

context $\kappa = \textbf{mean}(.35 \cdot [.4, .6], .33 \cdot [.2, .4], .31 \cdot [.1, .1])$

attention weights $\alpha = [0.35, 0.33, 0.31]$

scores $\tilde{\alpha} = [0.16, 0.10, 0.03]$



Global attention

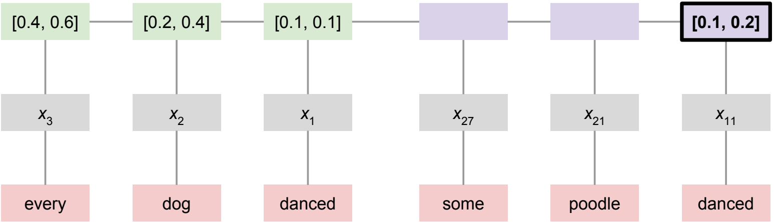
classifier $y = \mathbf{softmax}(\tilde{h}W + b)$

attention combo $\tilde{h} = \tanh([0.07, 0.11, 0.1, 0.2]W_\kappa)$

context $\kappa = \mathbf{mean}(.35 \cdot [.4, .6], .33 \cdot [.2, .4], .31 \cdot [.1, .1])$

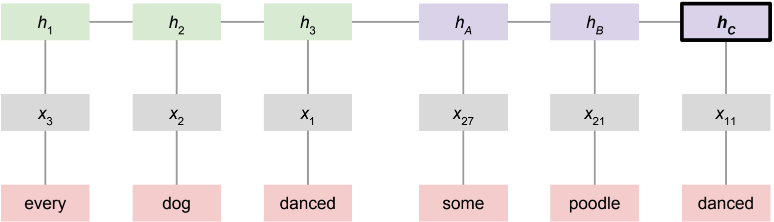
attention weights $\alpha = [0.35, 0.33, 0.31]$

scores $\tilde{\alpha} = [0.16, 0.10, 0.03]$

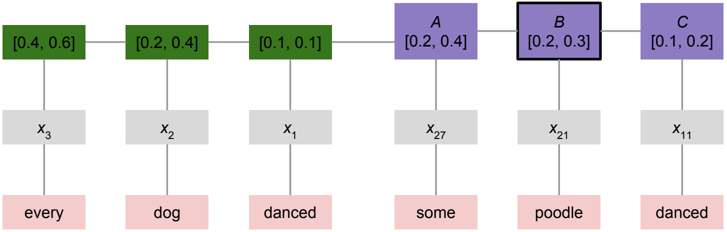


Other scoring functions (Luong et al. 2015)

$$\textbf{score}(h_C, h_i) = \begin{cases} h_C^\top h_i & \text{dot} \\ h_C^\top W_\alpha h_i & \text{general} \\ W_\alpha[h_C; h_i] & \text{concat} \end{cases}$$

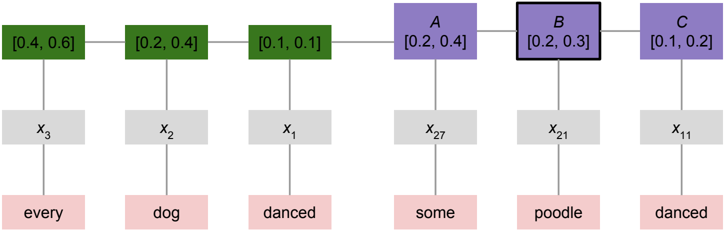


Word-by-word attention



Word-by-word attention

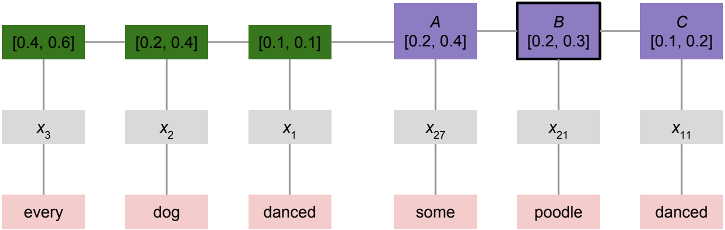
$$M = \tanh\left(\begin{bmatrix} 0.4 & 0.6 \\ 0.2 & 0.4 \\ 0.1 & 0.1 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 & \kappa_A \\ 0.2 & 0.3 & \kappa_A \\ 0.2 & 0.3 & \kappa_A \end{bmatrix} W_h\right)$$



Word-by-word attention

weights at B $\qquad \alpha_B = \textbf{softmax}(Mw)$

$$M = \tanh\left(\begin{bmatrix} 0.4 & 0.6 \\ 0.2 & 0.4 \\ 0.1 & 0.1 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 & \kappa_A \\ 0.2 & 0.3 & \kappa_A \\ 0.2 & 0.3 & \kappa_A \end{bmatrix} W_h\right)$$

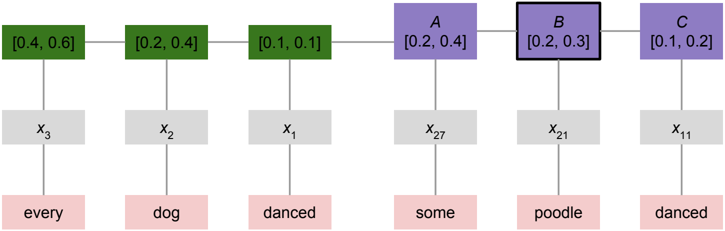


Word-by-word attention

context at B $\qquad \kappa_B = \begin{bmatrix} 0.4 & 0.6 \\ 0.2 & 0.4 \\ 0.1 & 0.1 \end{bmatrix} \alpha_B + \tanh\left(\kappa_A W_\alpha\right)$

weights at B $\qquad \alpha_B = \mathbf{softmax}(Mw)$

$$M = \tanh\left(\begin{bmatrix} 0.4 & 0.6 \\ 0.2 & 0.4 \\ 0.1 & 0.1 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 & \kappa_A \\ 0.2 & 0.3 & \kappa_A \\ 0.2 & 0.3 & \kappa_A \end{bmatrix} W_h \right)$$



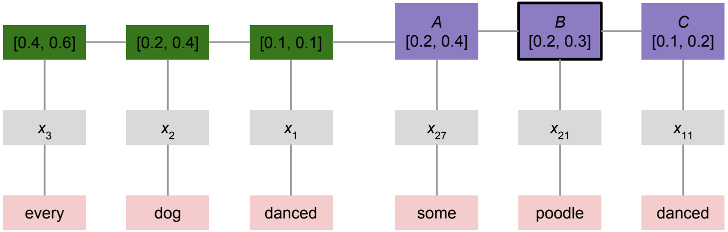
Word-by-word attention

classifier input $\tilde{h} = \tanh([\kappa_C; h_C]W_\kappa)$

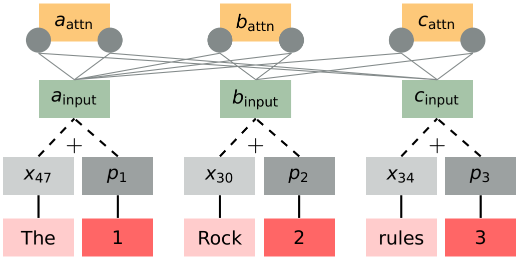
context at B $\kappa_B = \begin{bmatrix} 0.4 & 0.6 \\ 0.2 & 0.4 \\ 0.1 & 0.1 \end{bmatrix} \alpha_B + \tanh\left(\kappa_A W_\alpha\right)$

weights at B $\alpha_B = \textbf{softmax}(Mw)$

$$M = \tanh\left(\begin{bmatrix} 0.4 & 0.6 \\ 0.2 & 0.4 \\ 0.1 & 0.1 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 & \kappa_A \\ 0.2 & 0.3 & \kappa_A \\ 0.2 & 0.3 & \kappa_A \end{bmatrix} W_h \right)$$



Connection with the Transformer



$$c_{\text{attn}} = \textbf{sum}\left(\left[\alpha_1 a_{\text{input}}, \alpha_2 b_{\text{input}}\right]\right)$$

$$\alpha = \textbf{softmax}(\tilde{\alpha})$$

$$\tilde{\alpha} = \left[\frac{{c_{\text{input}}}^{\top} a_{\text{input}}}{\sqrt{d_k}}, \frac{{c_{\text{input}}}^{\top} b_{\text{input}}}{\sqrt{d_k}}\right]$$

$$c_{\text{input}} = x_{34} + p_3$$

Vaswani et al. 2017

Other variants

- Local attention (Luong et al. 2015) builds connections between selected points in the premise and hypothesis.
- Word-by-word attention can be set up in many ways, with many more learned parameters than my simple example. A pioneering instance for NLI is Rocktäschel et al. 2016.
- The attention representation at time t could be appended to the hidden representation at $t+1$ (Luong et al. 2015).
- Memory networks (Weston et al. 2015) can be used to address similar issues related to properly recalling past experiences.

References I

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kočiský, and Phil Plunsom. 2016. Reasoning about entailment with neural attention. ArXiv:1509.06664.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *Proceedings of ICLR 2015*.