Attention Strategies for Multi-Source Sequence-to-Sequence Learning

Jindřich Libovický, Jindřich Helcl

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University

August 2, 2017

Introduction

- Attention over multiple source sequences relatively unexplored.
- This work proposes two techniques:
  - *Flat* attention combination
  - *Hierarchical* attention combination
- Applied to tasks of multimodal translation and automatic post-editing.

Motivation

No universal method that models explicitly the importance of each input.

Multi-Source Sequence-to-Sequence Learning

Any number of input sequences with possibly different modalities.

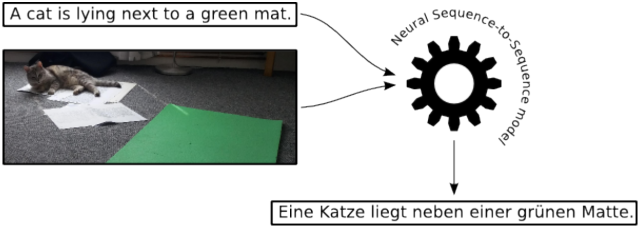


Figure 1: Multimodal translation example.

Examples

Multimodal translation, automatic post-editing, multi-source machine translation, ...

Attentive Sequence Learning

In each decoder step i

- compute distribution over encoder states given the decoder state
- the decoder gets a context vector to decide about its output

$$e_{ij} = v_a^\top \tanh(W_a s_i + U_a h_j) \quad (1)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (2)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad (3)$$

Attentive Sequence Learning

In each decoder step i

- compute distribution over encoder states given the decoder state
- the decoder gets a context vector to decide about its output

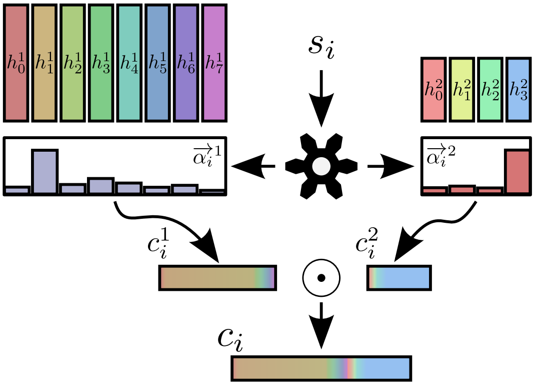
$$e_{ij} = v_a^\top \tanh(W_a s_i + U_a h_j) \quad (1)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (2)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad (3)$$

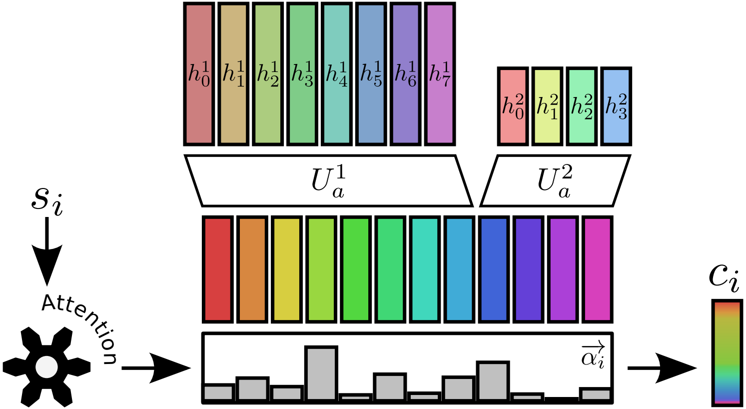
What about multiple inputs?

Context Vector Concatenation



- Widely used technique [Firat et al., 2016, Zoph and Knight, 2016].
- Attention over input sequences computed independently.
- Combination resolved later on in the network

Flat Attention Combination



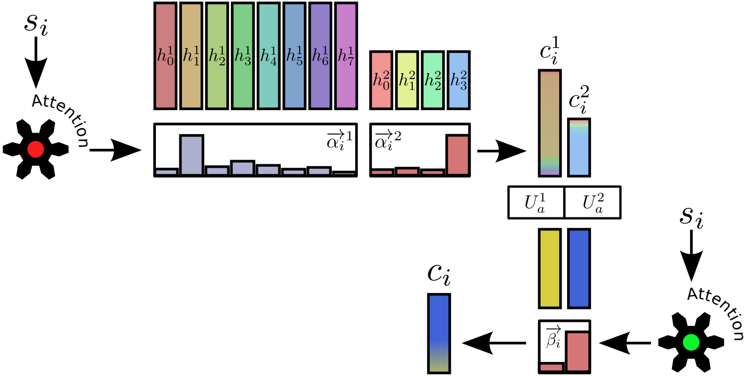
Importance of different inputs reflected in the **joint** attention distribution.

Flat Attention Combination

$$
\begin{array}{rcl}
\textbf{one source} & \rightarrow & N\ \textbf{sources} \\
e_{ij} = v_a^\top \tanh(W_a s_i + U_a h_j) & \rightarrow & e_{ij}^{(k)} = v_a^\top \tanh(W_a s_i + {U_a}^{(k)} h_j) \\
\alpha_{ij} = \dfrac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} & \rightarrow & \alpha_{ij}^{(k)} = \dfrac{\exp(e_{ij}^{(k)})}{\sum_{n=1}^{N} \sum_{m=1}^{T_x^{(n)}} \exp\left(e_{im}^{(n)}\right)} \\
c_i = \displaystyle\sum_{j=1}^{T_x} \alpha_{ij} h_j & \rightarrow & c_i = \displaystyle\sum_{k=1}^{N} \sum_{j=1}^{T_x^{(k)}} \alpha_{ij}^{(k)} {U_c}^{(k)} h_j^{(k)}
\end{array}
$$

- $U_a^{(k)}$, $U_c^{(k)}$ project states to a common space
- Question: Should $U_a^{(k)} = U_c^{(k)}$? (i.e. should the projection parameters be shared?)

Hierarchical Attention Combination



Attention distribution is **factored** by input.

Hierarchical Attention Combination

1. $\forall$ $k = 1 \dots N$ inputs

Compute the context vector:

$$c_i^{(k)} = \sum_{j=1}^{T_x^{(k)}} \alpha_{ij}^{(k)} h_j^{(k)}, \text{ where } \alpha_{ij}^{(k)} = \ \dots$$

...using the vanilla attention

2.

Compute another attention distribution over the intermediate context vectors $c_i^{(k)}$ and get the resulting context vector c_i.

$$\begin{aligned} e_i^{(k)} &= v_b^\top \tanh(W_b s_i + U_b^{(k)} c_i^{(k)}) \\ \beta_i^{(k)} &= \frac{\exp(e_i^{(k)})}{\sum_{n=1}^{N} \exp(e_i^{(n)})} \\ c_i &= \sum_{k=1}^{N} \beta_i^{(k)} U_c^{(k)} c_i^{(k)} \end{aligned}$$

- As in the flat scenario, the context vectors have to be projected to a shared space.
- Same question arises – should $U_b^{(k)} = U_c^{(k)}$?

Experiments and Results

- Experiments conducted on multimodal translation (MMT) and automatic post-editing (APE)
- In both flat and hierarchical scenarios, we tried both sharing and not sharing the projection matrices.
- Additionally, we tried using the sentinel gate [Lu et al., 2016], which enables the decoder to decide whether or not to attend to any encoder.

Experiments conducted using Neural Monkey, code available here:
`https://github.com/ufal/neuralmonkey`.

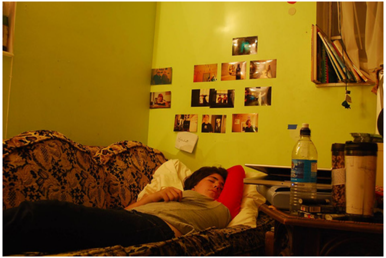
Experiments and Results

	share	sent.	MMT BLEU	MMT METEOR	APE BLEU	APE HTER
concat.			31.4 $\pm$ .8	48.0 $\pm$ .7	62.3 $\pm$ .5	24.4 $\pm$ .4
flat	$\times$	$\times$	30.2 $\pm$ .8	46.5 $\pm$ .7	62.6 $\pm$ .5	24.2 $\pm$ .4
	$\times$	$\checkmark$	29.3 $\pm$ .8	45.4 $\pm$ .7	62.3 $\pm$ .5	24.3 $\pm$ .4
	$\checkmark$	$\times$	30.9 $\pm$ .8	47.1 $\pm$ .7	62.4 $\pm$ .6	24.4 $\pm$ .4
	$\checkmark$	$\checkmark$	29.4 $\pm$ .8	46.9 $\pm$ .7	62.5 $\pm$ .6	24.2 $\pm$ .4
hierarchical	$\times$	$\times$	**32.1 $\pm$ .8**	**49.1 $\pm$ .7**	62.3 $\pm$ .5	24.1 $\pm$ .4
	$\times$	$\checkmark$	28.1 $\pm$ .8	45.5 $\pm$ .7	62.6 $\pm$ .6	24.1 $\pm$ .4
	$\checkmark$	$\times$	26.1 $\pm$ .7	42.4 $\pm$ .7	62.4 $\pm$ .5	24.3 $\pm$ .4
	$\checkmark$	$\checkmark$	22.0 $\pm$ .7	38.5 $\pm$ .6	62.5 $\pm$ .5	24.1 $\pm$ .4

Results on the Multi30k dataset and the APE dataset. The column ‘share’ denotes whether the projection matrix is shared for energies and context vector computation, ‘sent.’ indicates whether the sentinel vector has been used or not.

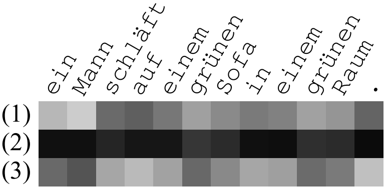
Example

Source:



A man sleeping in a green room on a couch.

Output with attention:



(1) source, (2) image, (3) sentinel

Reference:

ein Mann schläft in einem grünen Raum auf einem Sofa .

Conclusions

- The results show both methods achieve comparable results to the existing approach (concatenation of the context vectors).
- Hierarchical attention combination achieved best results on MMT, and is faster to train.
- Both methods provide a trivial way to inspect the attention distribution w.r.t. the individual inputs.

Conclusions

- The results show both methods achieve comparable results to the existing approach (concatenation of the context vectors).
- Hierarchical attention combination achieved best results on MMT, and is faster to train.
- Both methods provide a trivial way to inspect the attention distribution w.r.t. the individual inputs.

Thank you for your *attention!*

References I

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, CA, USA, pages 866–875. http://www.aclweb.org/anthology/N16-1101.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *CoRR* abs/1612.01887. http://arxiv.org/abs/1612.01887.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, CA, USA, pages 30–34. http://www.aclweb.org/anthology/N16-1004.