Introduction to Attention Mechanism in Deep Learning

23.06.2020

Recap

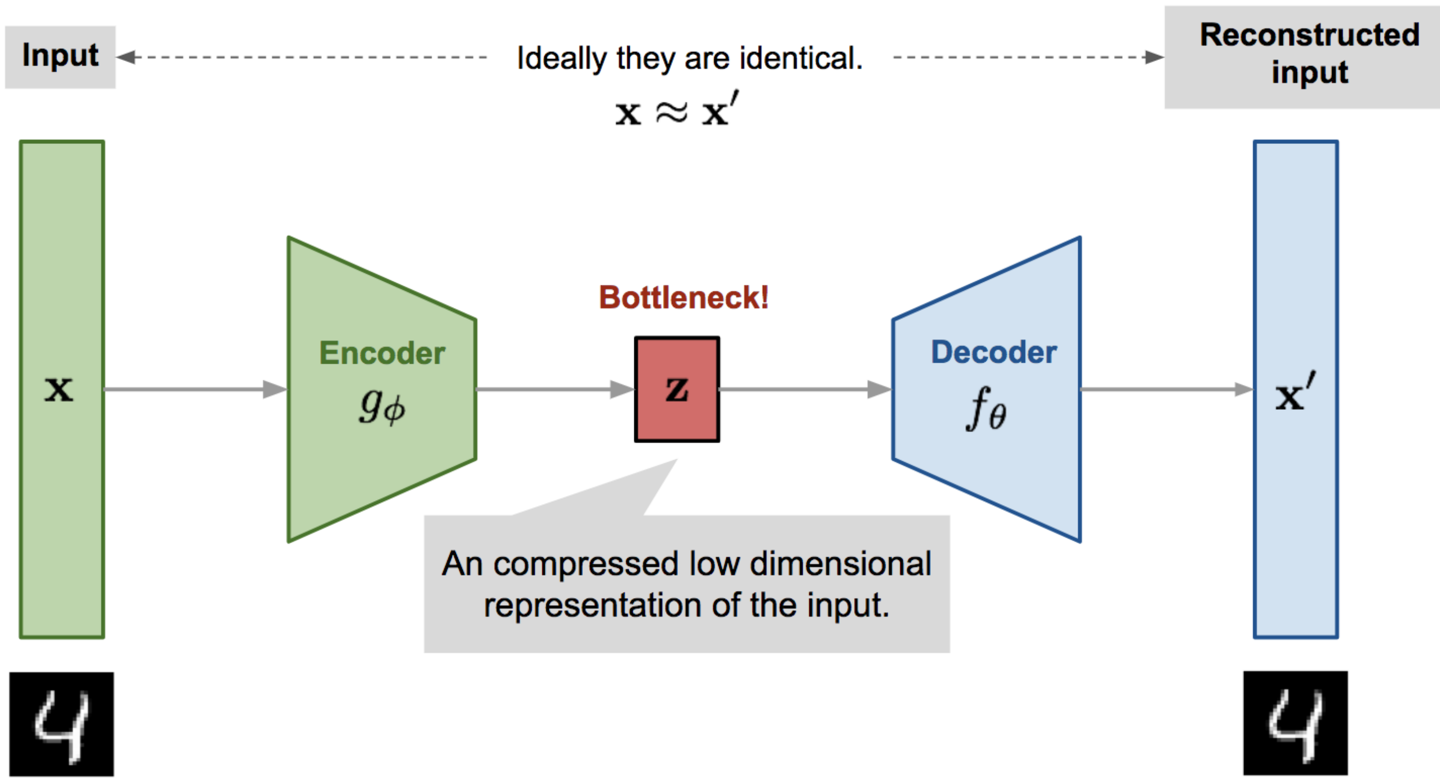


Figure 01. Autoencoder with MNIST dataset[1]

What about sequential data?

[1] Image source: https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

Seq2seq Model

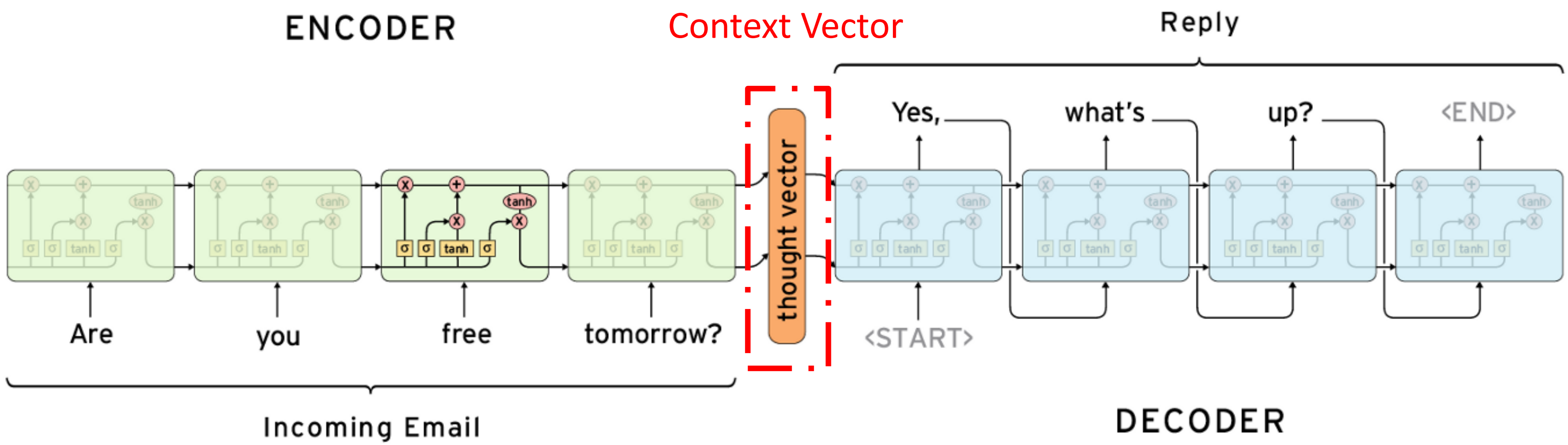


Figure 02. Seq2seq LSTMs[1]

Seq2seq was designed to encode sequential data (e.g. text, audio, speech) and decode sequential data of arbitrary length respectively.

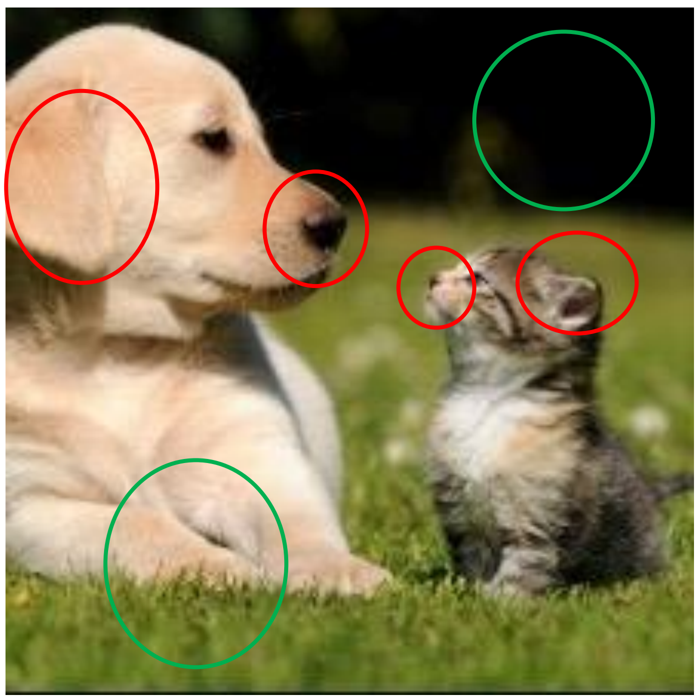
Application: Machine Translation, Speech Recognition, Parsing Sentence into Grammar Tree

Problem: Context Vector of fixed length!

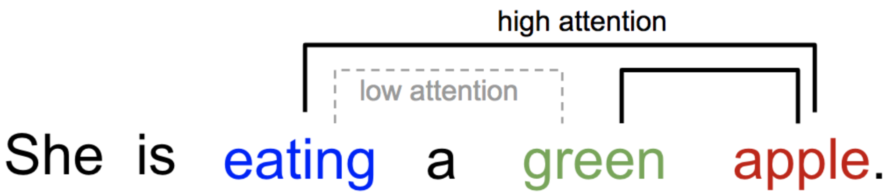
[1] Image source: https://towardsdatascience.com/sequence-to-sequence-tutorial-4fde3ee798d8

Attention

- Biologically, attention existed is the visual system of a human being.
- High Attention: "High Resolution"
- Low Attention: "Low Resolution"



(a)



(b)

Figure 03. (a) Perceiving attention visually [1] and in (b) text [2]

[1] Image source: https://distill.pub/2018/building-blocks/
[2] Image source: https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Attention

- Attention mechanism was design to tackle fixed context vector
- Attention create shortcuts between the context vector and the input source
- Focus more on certain regions and less on others
- Attention fells what to keep during reduction to minimize information loss salient features
- Attention provide more “fine grain” of region to be attended
- Classification: Self- Attention, Soft-Attention, Hard-Attention

Soft vs Hard Attention

Soft Attention	Hard Attention
Different Parts, Different Subregions	Only ONE subregion
Deterministic	Stochastic
Differentiable	Non-differentiable

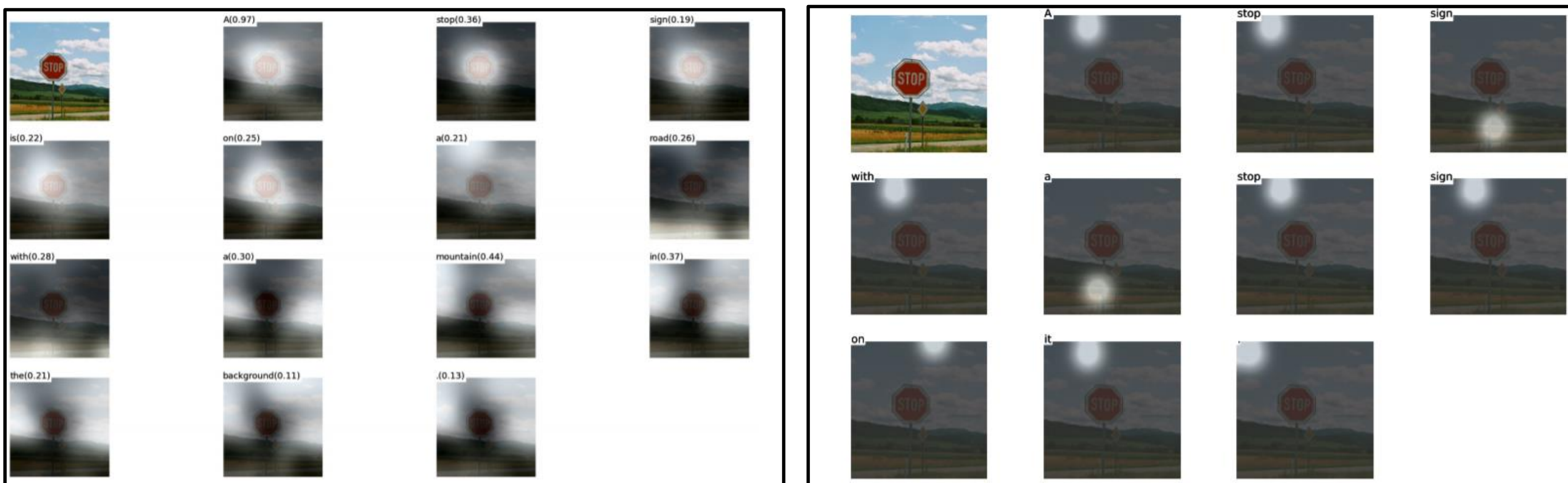
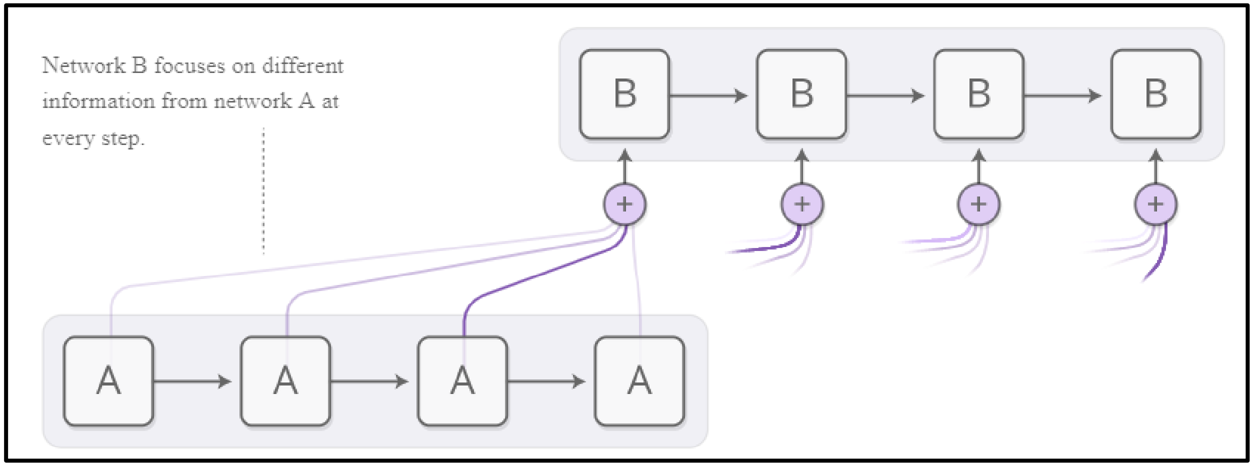


Figure 04. (*Left*) Soft subregions vs (*Right*) hard subregion [1]

[1] Image source: K. Xu et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", Fig.10

Soft Attention-Additive Attention (1)



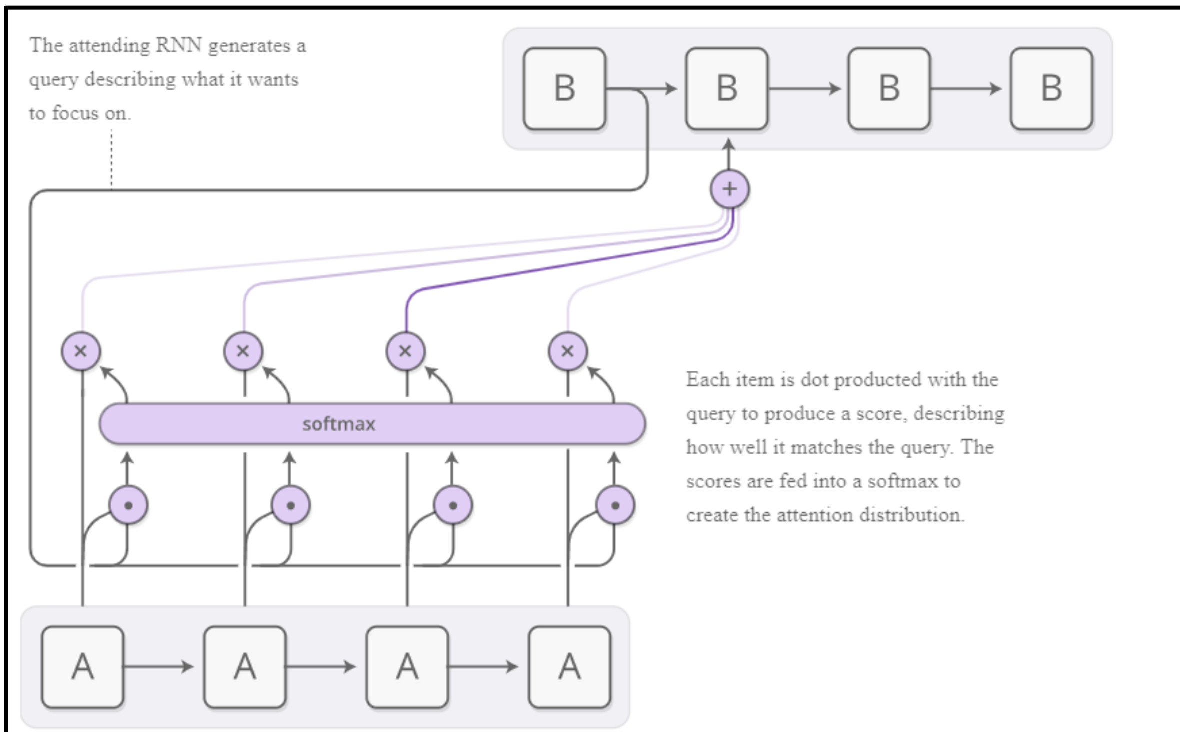


Figure 05. (*Top*) Aggregated additive attention mechanism. (*Bottom*) Unfold additive attention mechanism [1]

[1] Image source: Olah & Carter, "Attention and Augmented Recurrent Neural Networks", Distil, 2016.

Soft Attention-Additive Attention(2)

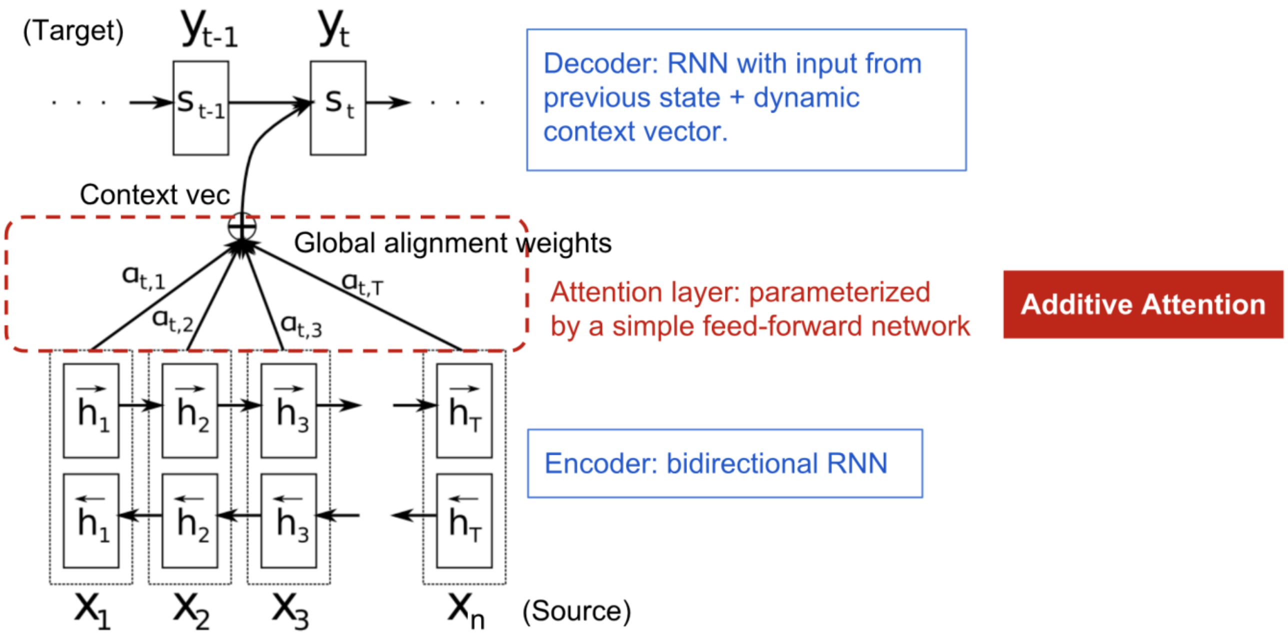


Figure 06. Diagram derived from Bahdanau et al., 2015 with addition information from [1]

$$\mathbf{c}_t = \sum_{i=1}^{n} \alpha_{t,i} \boldsymbol{h}_i \quad \text{; Context vector for output } y_t$$

$$\alpha_{t,i} = \text{align}(y_t, x_i) \quad \text{; How well two words } y_t \text{ and } x_i \text{ are aligned.}$$

$$= \frac{\exp(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_i))}{\sum_{i'=1}^{n} \exp(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_{i'}))} \quad \text{; Softmax of some predefined alignment score..}$$

$$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$$

[1] Image source: https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Soft Attention-Additive Attention(3)

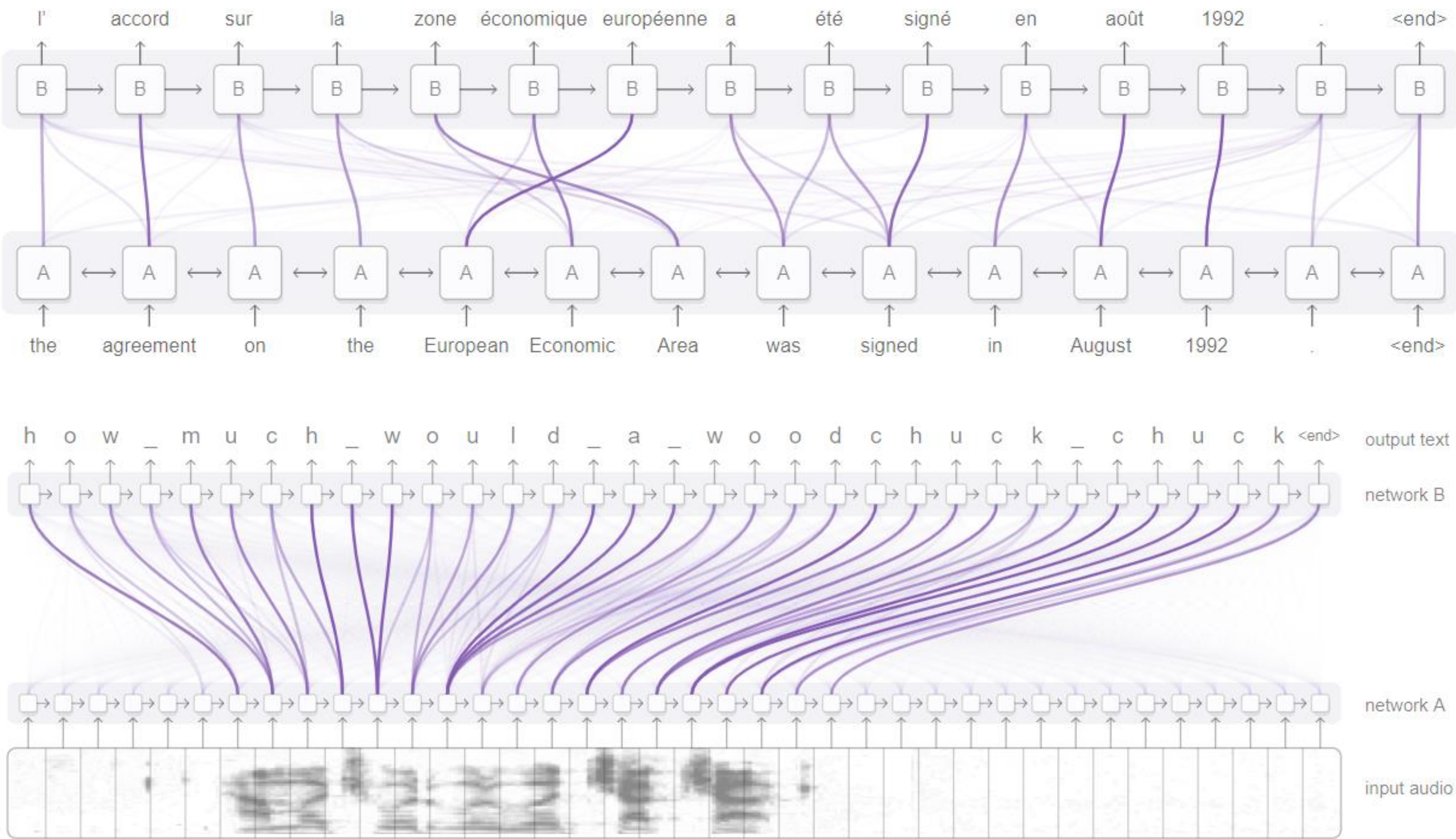


Figure 06. (*Top*) Text2text machine translation, (*Bottom*) Speech2text machine translation [1]

[1] Image source: Olah & Carter, "Attention and Augmented Recurrent Neural Networks“, Distill, 2016.

Summary

Name	Definition	Citation
Self-Attention(&)	Relating different positions of the same input sequence. Theoretically the self-attention can adopt any score functions above, but just replace the target sequence with the same input sequence.	Cheng2016
Global/Soft	Attending to the entire input state space.	Xu2015
Local/Hard	Attending to the part of input state space; i.e. a patch of the input image.	Xu2015; Luong2015

Name	Alignment score function	Citation
Content-base attention	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$	Graves2014
Additive(*)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$	Bahdanau2015
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ Note: This simplifies the softmax alignment to only depend on the target position.	Luong2015
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.	Luong2015
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$	Luong2015
Scaled Dot-Product(^)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	Vaswani2017

[1] Image source: https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Thank you for your kind attention!