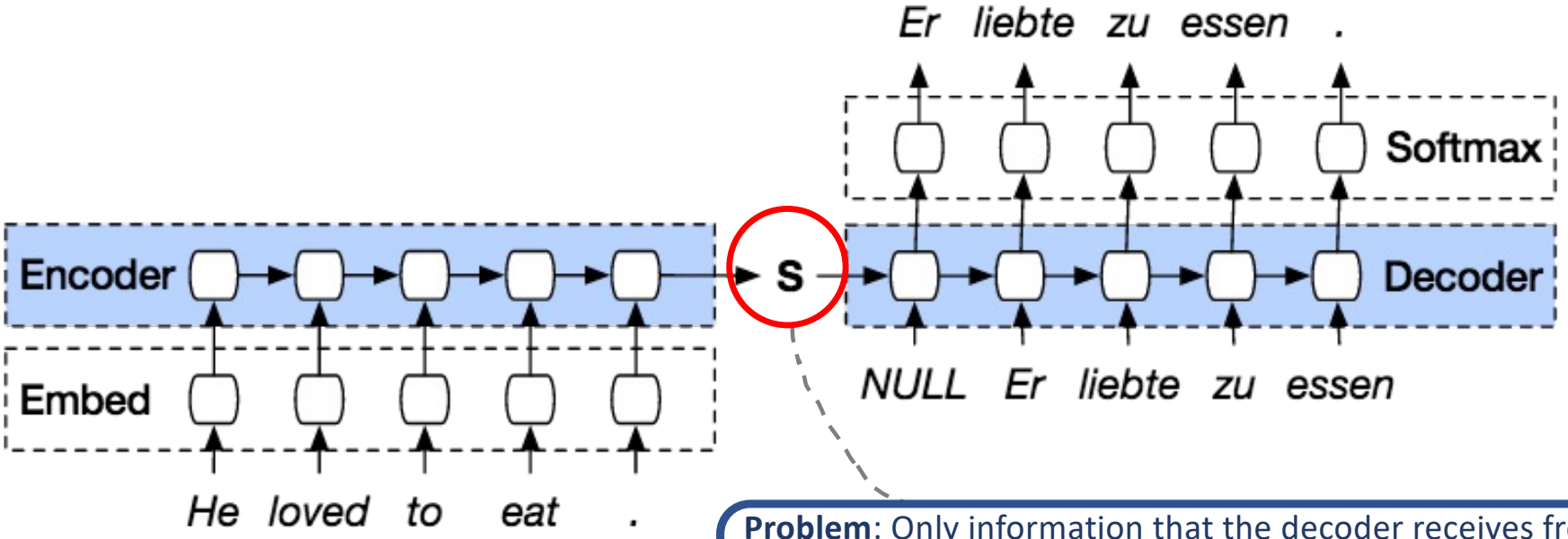
Week 05:
Attention Mechanism

Dr. Hongping Cai
Department of Computer Science
University of Bath

Topic 1: Attention Mechanism

Neural Machine Translation (NMT)

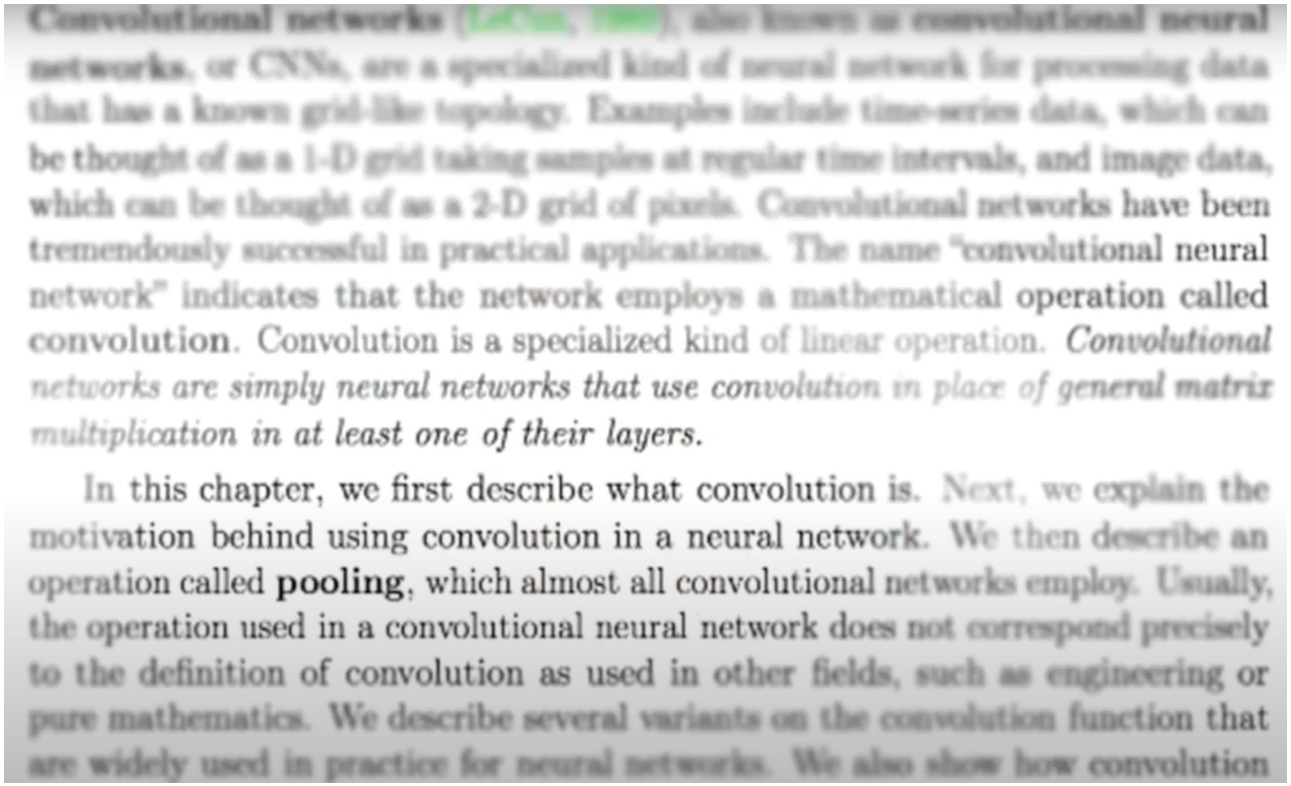
- Seq2seq [Sutskever et al, 2014][Cho et al, 2014]



Problem: Only information that the decoder receives from the encoder is the last encoder hidden state – **encoding bottleneck**. It doesn't work well when the sequence is long.
Solution: seq2seq + **attention**

Image from: https://smerity.com/articles/2016/google_nmt_arch.html

Intuition of Attention Mechanism

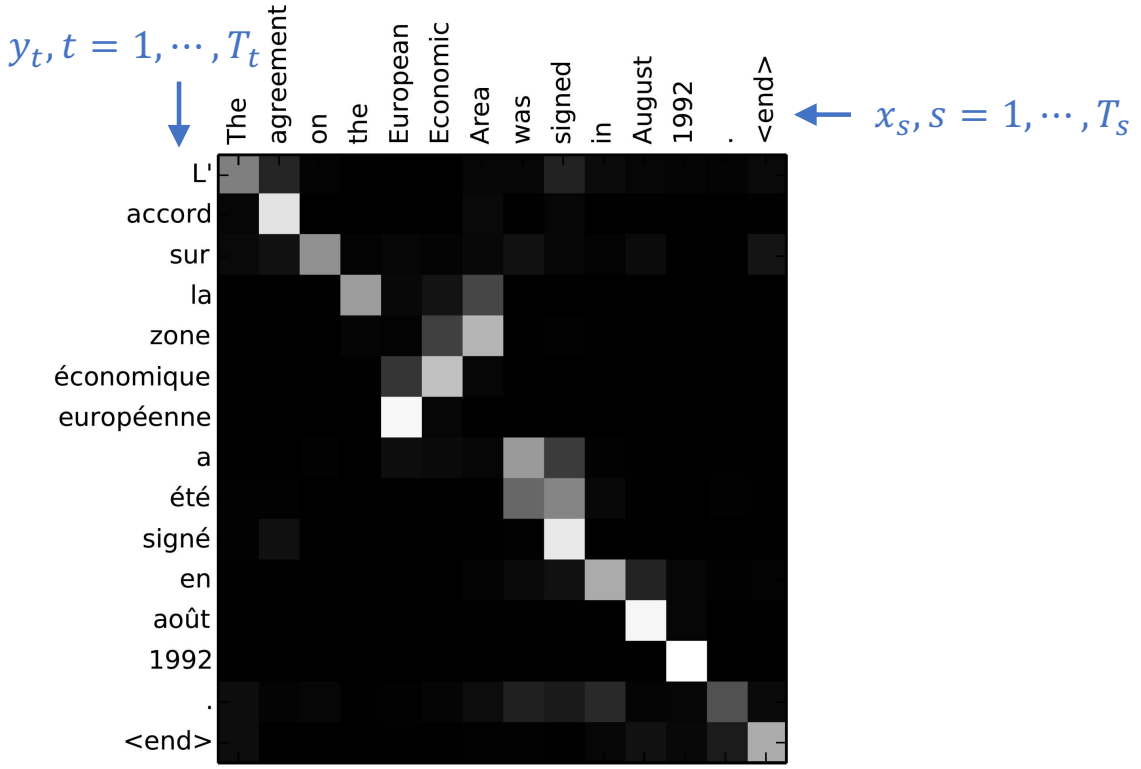


Convolutional networks ([LeCun, 1989]), also known as **convolutional neural networks**, or CNNs, are a specialized kind of neural network for processing data that has a known grid-like topology. Examples include time-series data, which can be thought of as a 1-D grid taking samples at regular time intervals, and image data, which can be thought of as a 2-D grid of pixels. Convolutional networks have been tremendously successful in practical applications. The name "convolutional neural network" indicates that the network employs a mathematical operation called **convolution**. Convolution is a specialized kind of linear operation. ***Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.***

In this chapter, we first describe what convolution is. Next, we explain the motivation behind using convolution in a neural network. We then describe an operation called **pooling**, which almost all convolutional networks employ. Usually, the operation used in a convolutional neural network does not correspond precisely to the definition of convolution as used in other fields, such as engineering or pure mathematics. We describe several variants on the convolution function that are widely used in practice for neural networks. We also show how convolution

Attention in translation: modelling the contextual relations by selectively focusing on parts of the source sentence.

Intuition of Attention Mechanism



α_{ts}: The amount of "**attention**" the target word y_t should pay to the source word x_s.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. ICLR 2015.

Seq2seq

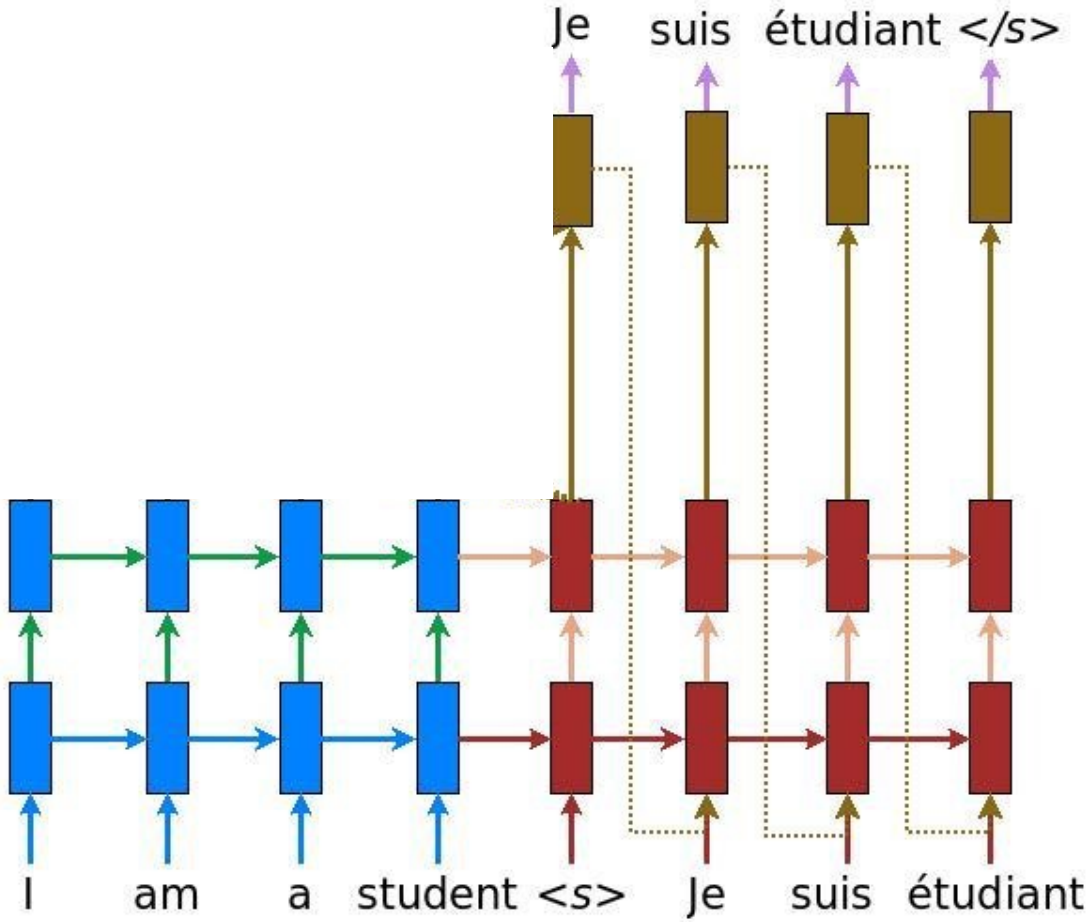


Image from: https://github.com/tensorflow/nmt

Seq2seq + Attention [Bahdanau et al, 2014]

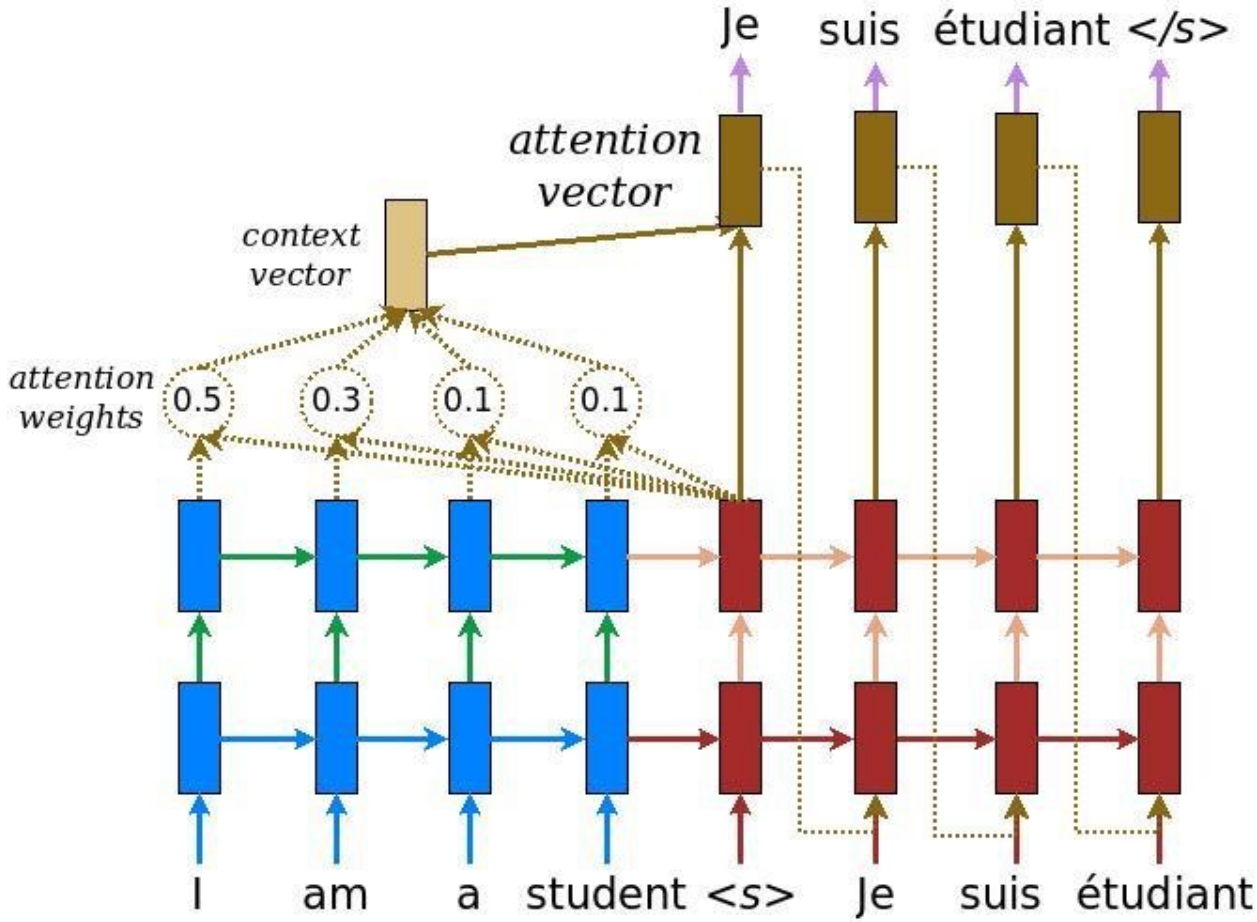
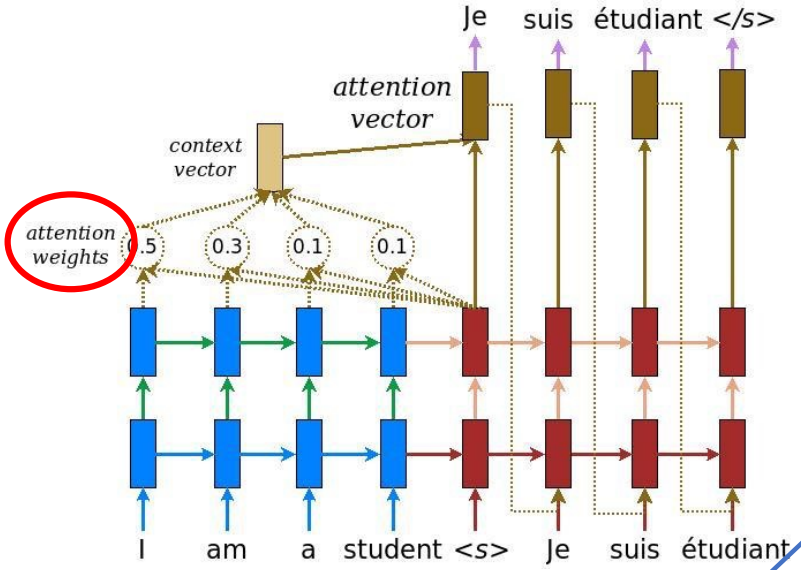


Image from: https://github.com/tensorflow/nmt

[Bahdanau et al, 2014] Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint* arXiv:1409.0473, 2014.

STEP1: Attention Weight



- **Attention weight** (α_{ts}) measures how much relevance between each source state ($\bar{h}_s$) with the target state (h_t).

$$\text{score}(h_t, \bar{h}_s) = h_t^T \cdot \bar{h}_s$$

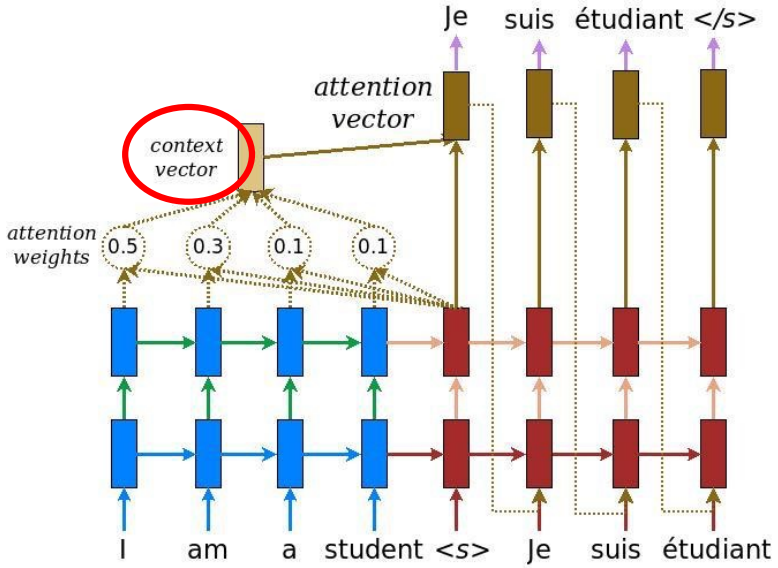
$$\alpha_{ts} = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \bar{h}_{s'}))}$$

There are various choices of scoring function, e.g., dot-product, additive, general/multiplicative, etc.

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \boldsymbol{W} \bar{\boldsymbol{h}}_s & \text{[Luong's multiplicative style]} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_1} \boldsymbol{h}_t + \boldsymbol{W_2} \bar{\boldsymbol{h}}_s\right) & \text{[Bahdanau's additive style]} \end{cases}$$

Image from: https://github.com/tensorflow/nmt

STEP 2: Context vector

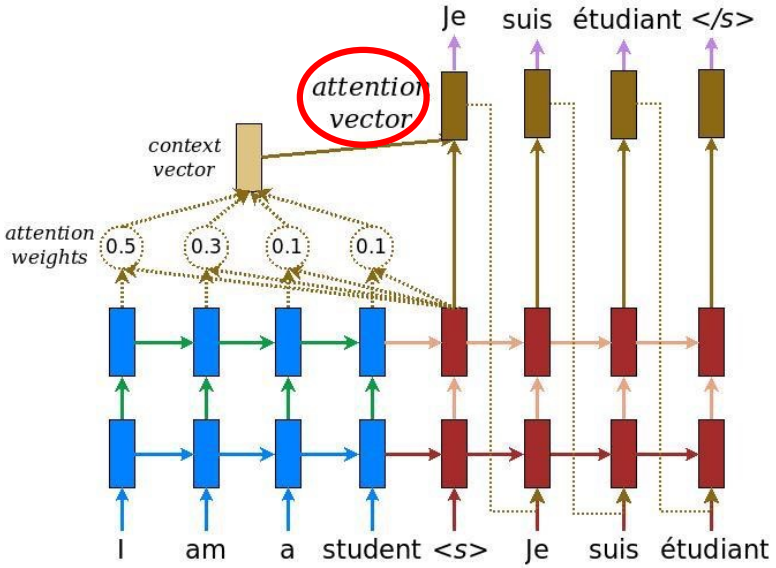


- **Context vector** (c_t) is the weighted average of the source states

$$c_t = \sum_s \alpha_{ts} \bar{h}_s$$

Image from: https://github.com/tensorflow/nmt

STEP 3: Attention vector



- **Attention vector** (a_t) is yielded by combining the context vector with the current target hidden state.

$$a_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t])$$

The output vector (with attention): $\hat{y}_t = g(W_a a_t)$

The output vector (without attention): $\hat{y}_t = g(W_{hy} h_t)$

Image from: https://github.com/tensorflow/nmt

		German-English translations
	src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
Seq2seq	ref	However , in an interview , Bloom has said that he and *Kerr* still love each other .
attention →	*best*	In an interview , however , Bloom said that he and *Kerr* still love .
Seq2seq →	base	However , in an interview , Bloom said that he and **Tina** were still <unk> .
	src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
	ref	The *austerity imposed by Berlin and the European Central Bank , coupled with the straitjacket* imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
Seq2seq		
attention →	*best*	Because of the strict *austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket* in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
Seq2seq →	base	Because of the pressure **imposed by the European Central Bank and the Federal Central Bank with the strict austerity** imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

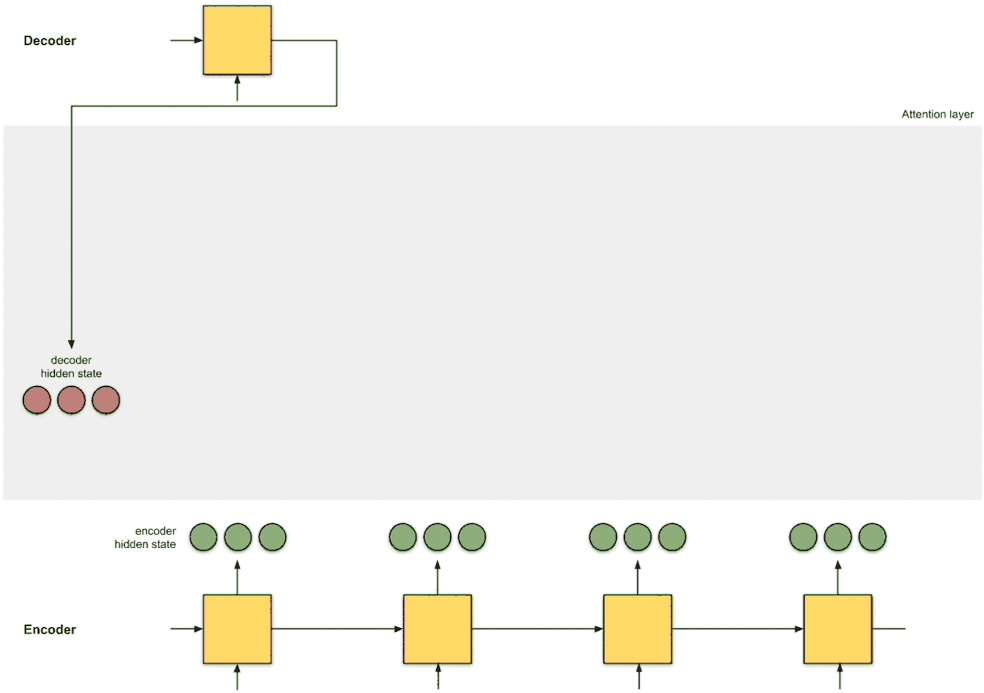
Luong et al. Effective approaches to Attention-based Neural Machine Translation, EMNLP 2015.

References

- Thang Luong. Neural Machine Translation (seq2seq) Tutorial. https://github.com/tensorflow/nmt
- Blog by Raimi Karim. Attn: Illustrated Attention.2022. https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Topic 2: An Example for Attention-based Seq2Seq

Example



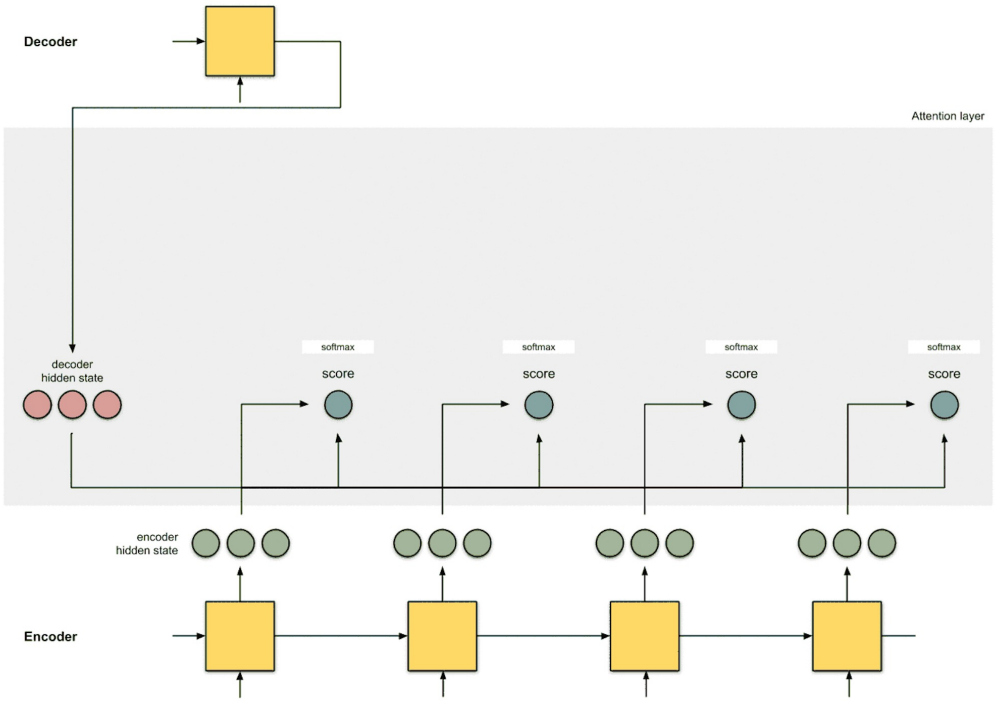
dec_h = [10 5 10]

enc_h

```
---------
[0 1 1]
[5 0 1]
[1 1 0]
[0 5 1]
```

https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

STEP 1: The current target hidden state is compared with all source states to derive attention weights.



Attention weight

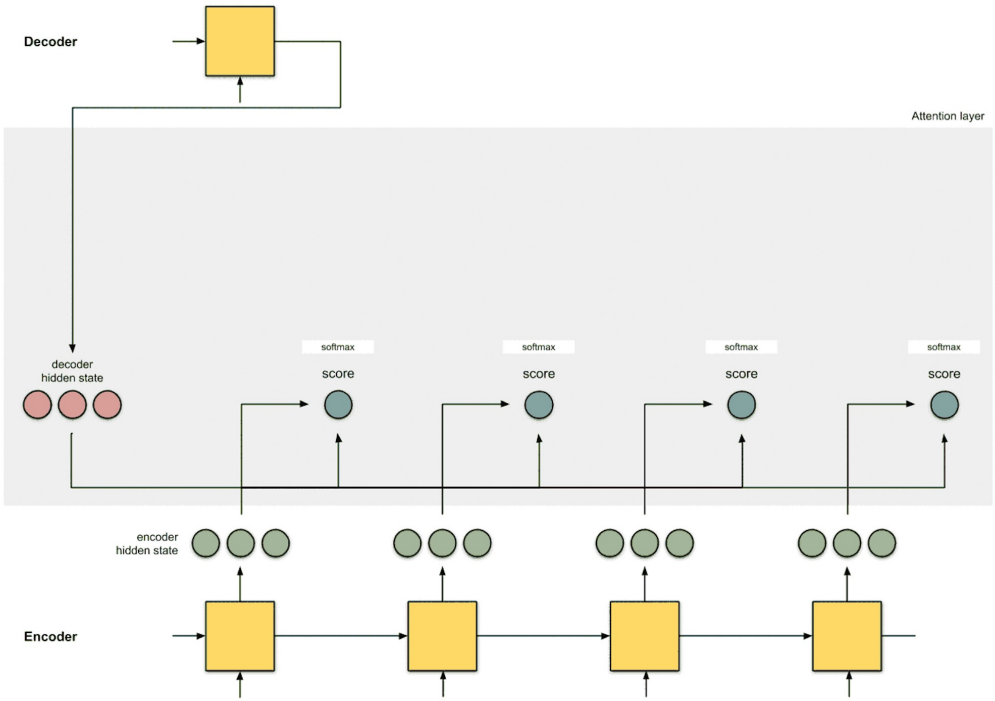
$$\text{score}(h_t, \bar{h}_s) = h_t^T \cdot \bar{h}_s$$

$$\alpha_{ts} = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \bar{h}_{s'}))}$$

```
dec_h = [10 5 10]

enc_h  score attn_w
------------------------------
[0 1 1]
[5 0 1]
[1 1 0]
[0 5 1]
```

https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

STEP 1: The current target hidden state is compared with all source states to derive attention weights.



Attention weight

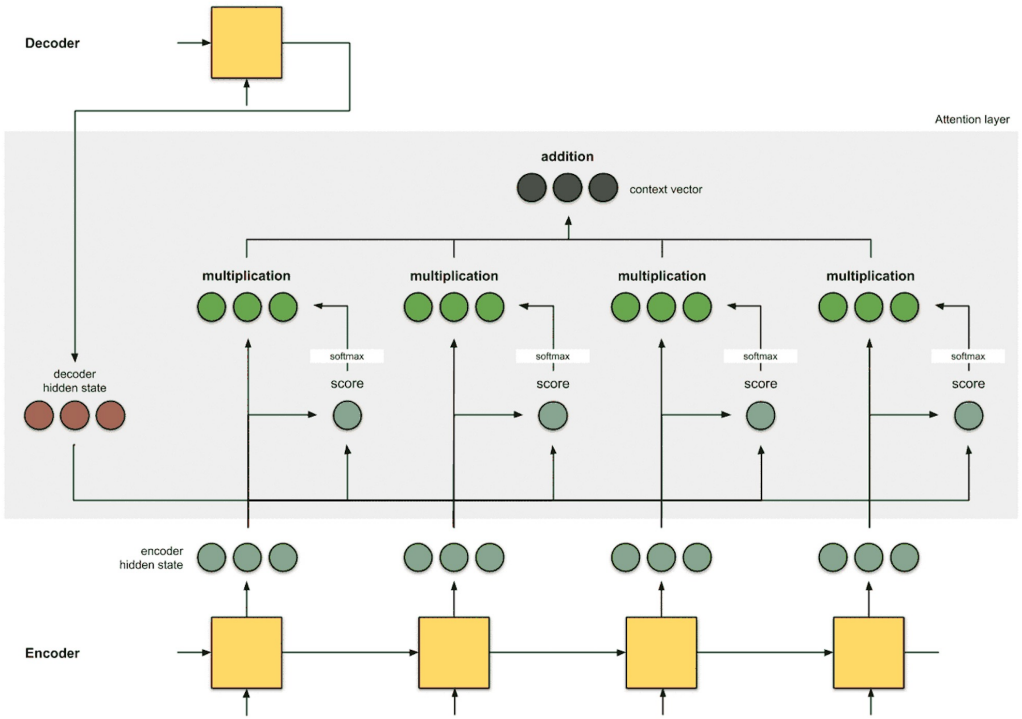
$$\text{score}(h_t, \bar{h}_s) = h_t^T \cdot \bar{h}_s$$

$$\alpha_{ts} = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \bar{h}_{s'}))}$$

```
dec_h = [10 5 10]

enc_h  score attn_w
------------------------------
[0 1 1]  15    0
[5 0 1]  60    1
[1 1 0]  15    0
[0 5 1]  35    0
```

https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

STEP 2: Multiply each encoder hidden state by its attention weight, then sum these vectors up.



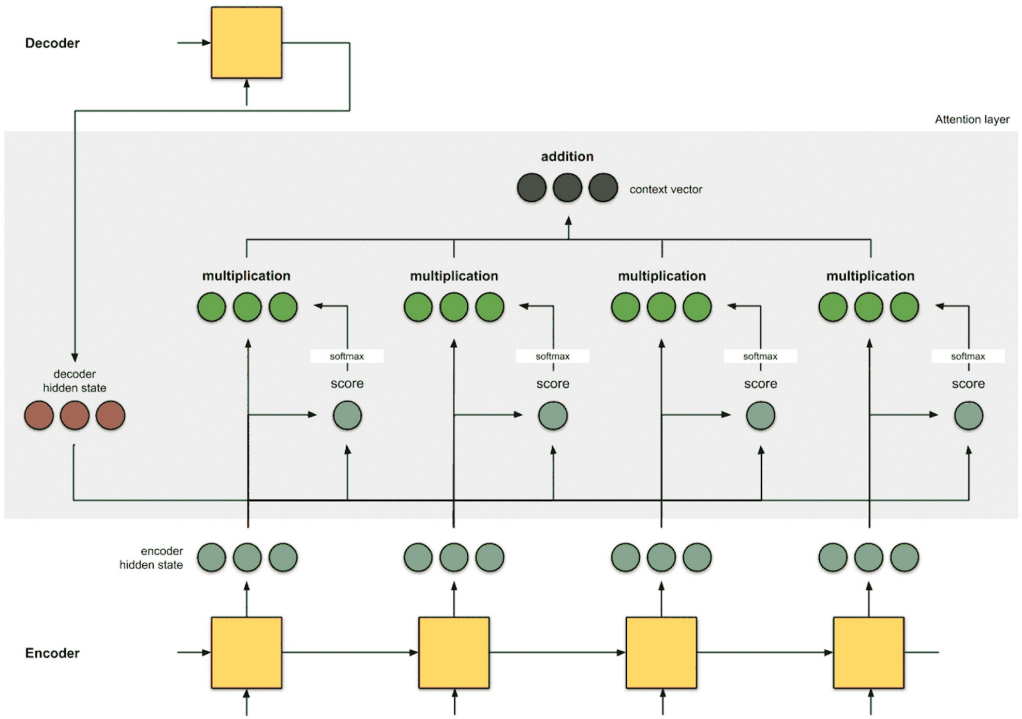
Context vector

$$c_t = \sum_s \alpha_{ts} \bar{h}_s$$

```
dec_h = [10 5 10]

enc_h  score attn_w alignment
------------------------------
[0 1 1]  15    0       [0 0 0]
[5 0 1]  60    1       [5 0 1]
[1 1 0]  15    0       [0 0 0]
[0 5 1]  35    0       [0 0 0]

Context =
```

https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

STEP 2: Multiply each encoder hidden state by its attention weight, then sum these vectors up.



Context vector

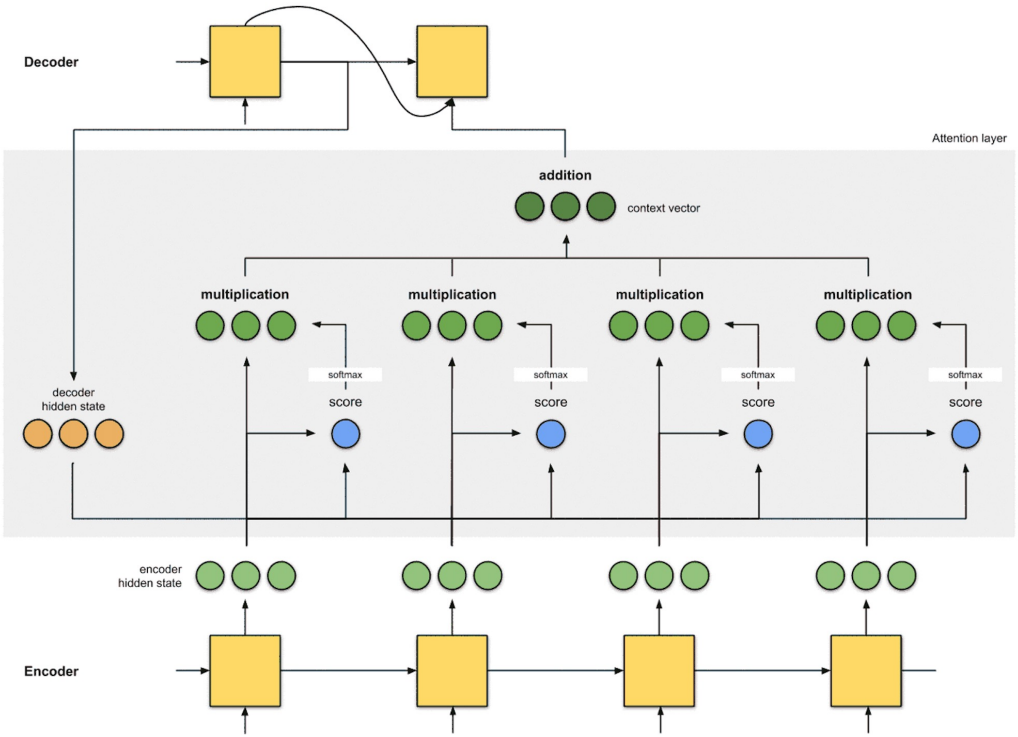
$$c_t = \sum_s \alpha_{ts} \bar{h}_s$$

```
dec_h = [10 5 10]

enc_h  score attn_w alignment
------------------------------
[0 1 1]  15    0       [0 0 0]
[5 0 1]  60    1       [5 0 1]
[1 1 0]  15    0       [0 0 0]
[0 5 1]  35    0       [0 0 0]

Context = [5 0 1]
```

https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

STEP 3: Feed the context vector into the decoder.

Attention vector $a_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t])$




```
dec_h = [10 5 10]

enc_h  score attn_w alignment
-------------------------------
[0 1 1]  15    0       [0 0 0]
[5 0 1]  60    1       [5 0 1]
[1 1 0]  15    0       [0 0 0]
[0 5 1]  35    0       [0 0 0]

Context = [5 0 1]

dec_h_context = [5 0 1 10 5 10]
```

https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

References

- Thang Luong. Neural Machine Translation (seq2seq) Tutorial. https://github.com/tensorflow/nmt
- Blog by Raimi Karim. Attn: Illustrated Attention.2022. https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Topic 3: Three Attention-Based Seq2Seq Models

UNIVERSITY OF BATH

Bahdanau et al. (2014)

- **Encoder**:
- **Decoder**:
- **Score function**:
- **Context vector integrated to decoder**:

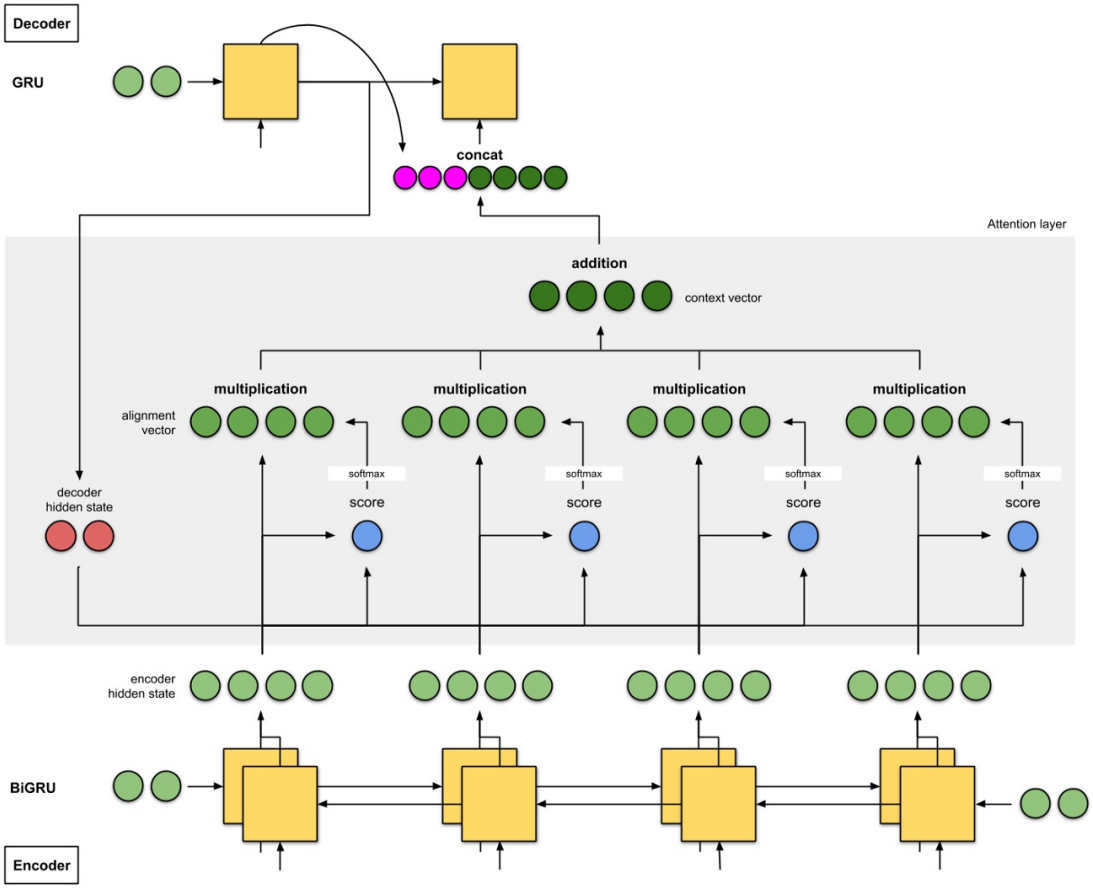
Luong et al. (2015)

Google's NMT (2016)

Bahdanau et al. (2014)

- **Encoder**: Bidirectional GRU
- **Decoder**: GRU
- **Score function**: concat
- **Context vector integrated to decoder**: concat

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_a [\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$



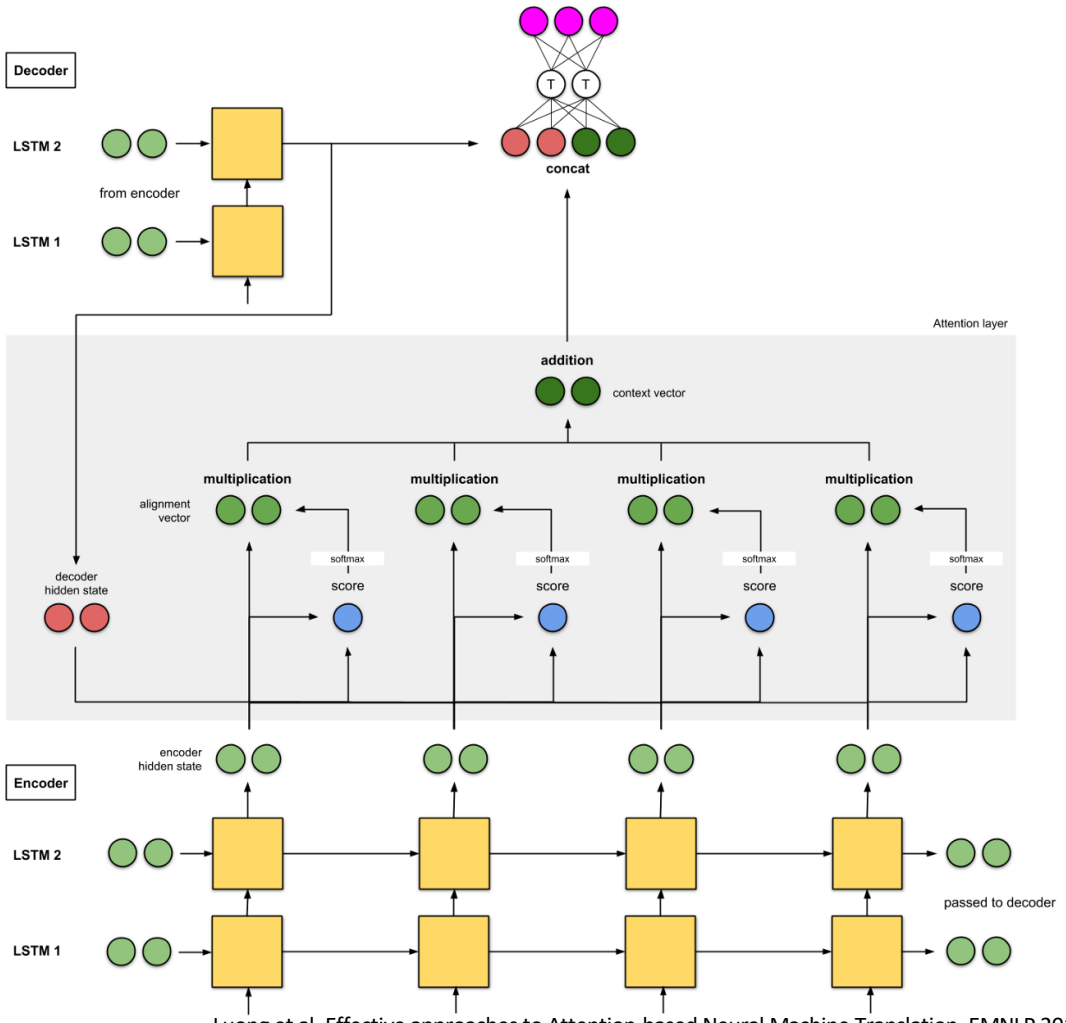
Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint* arXiv:1409.0473, 2014.

Image from: https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Luong et al. (2015)

- **Encoder**: two-stacked LSTM
- **Decoder**: two-stacked LSTM
- **Score function**: concat; dot; location based; general
- **Context vector integrated to decoder**: concat + dense layer

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_a [\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

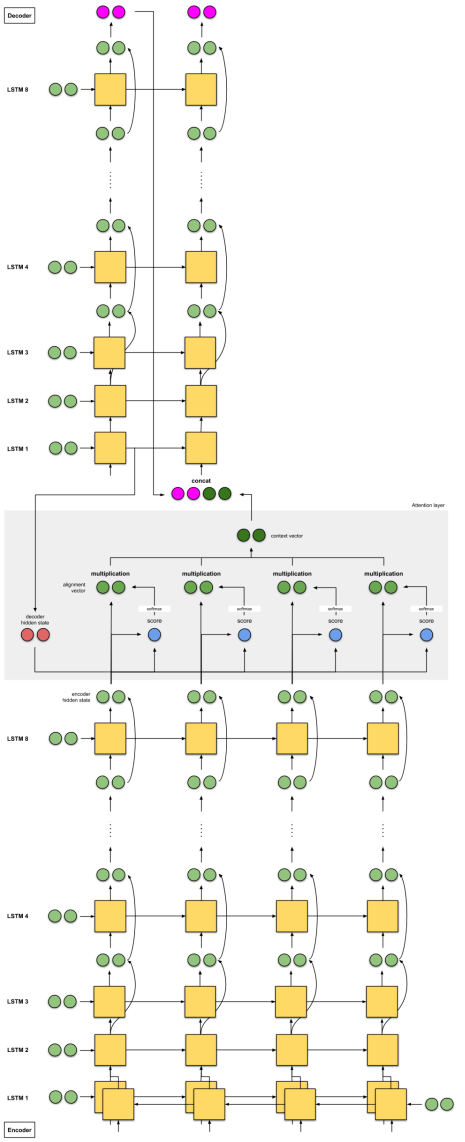


Luong et al. Effective approaches to Attention-based Neural Machine Translation, EMNLP 2015.
Image from: https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Google's NMT (2016)

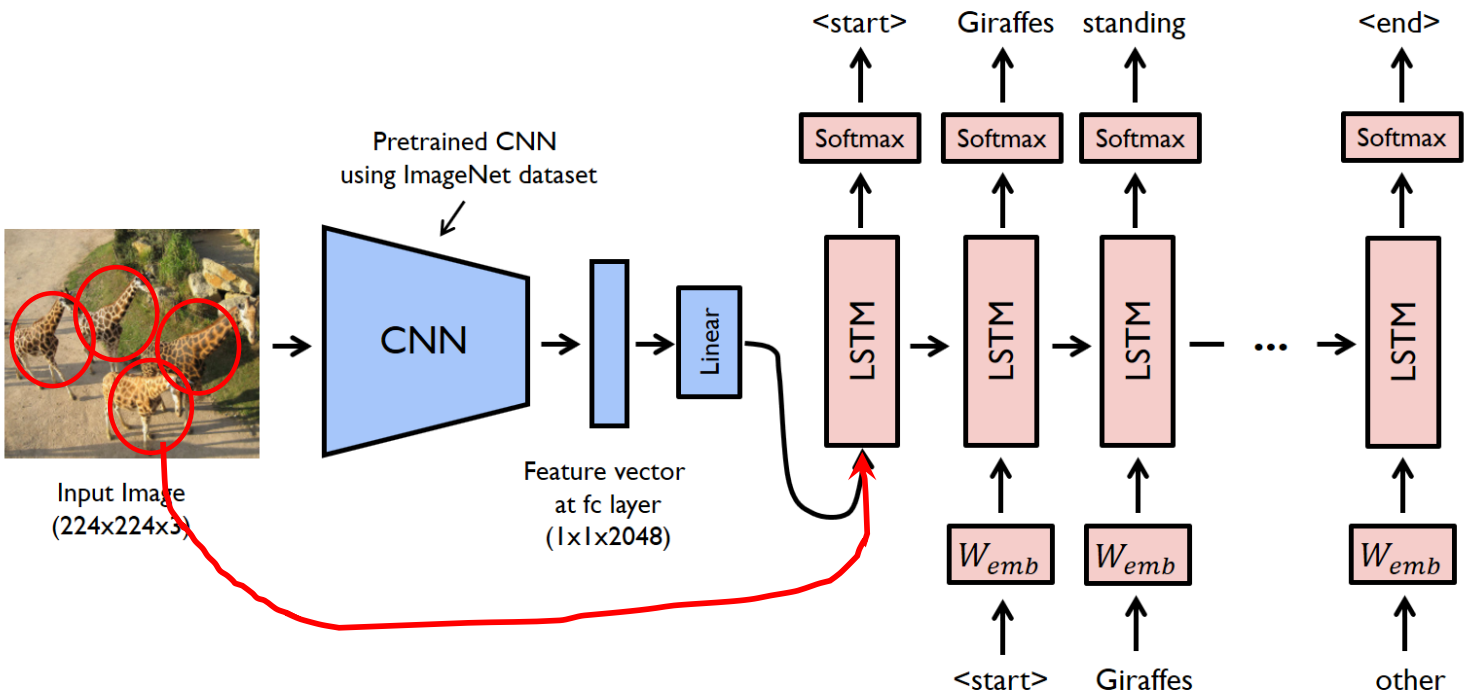
- **Encoder**: 8-stacked LSTM, bidirectional (first layer), residual connections
- **Decoder**: 8-stacked LSTM
- **Score function**: concat
- **Context vector integrated to decoder**: concat

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_a [\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$



Google's neural machine translation system: Bridge the gap between human and machine translation, 2016.
Image from: https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

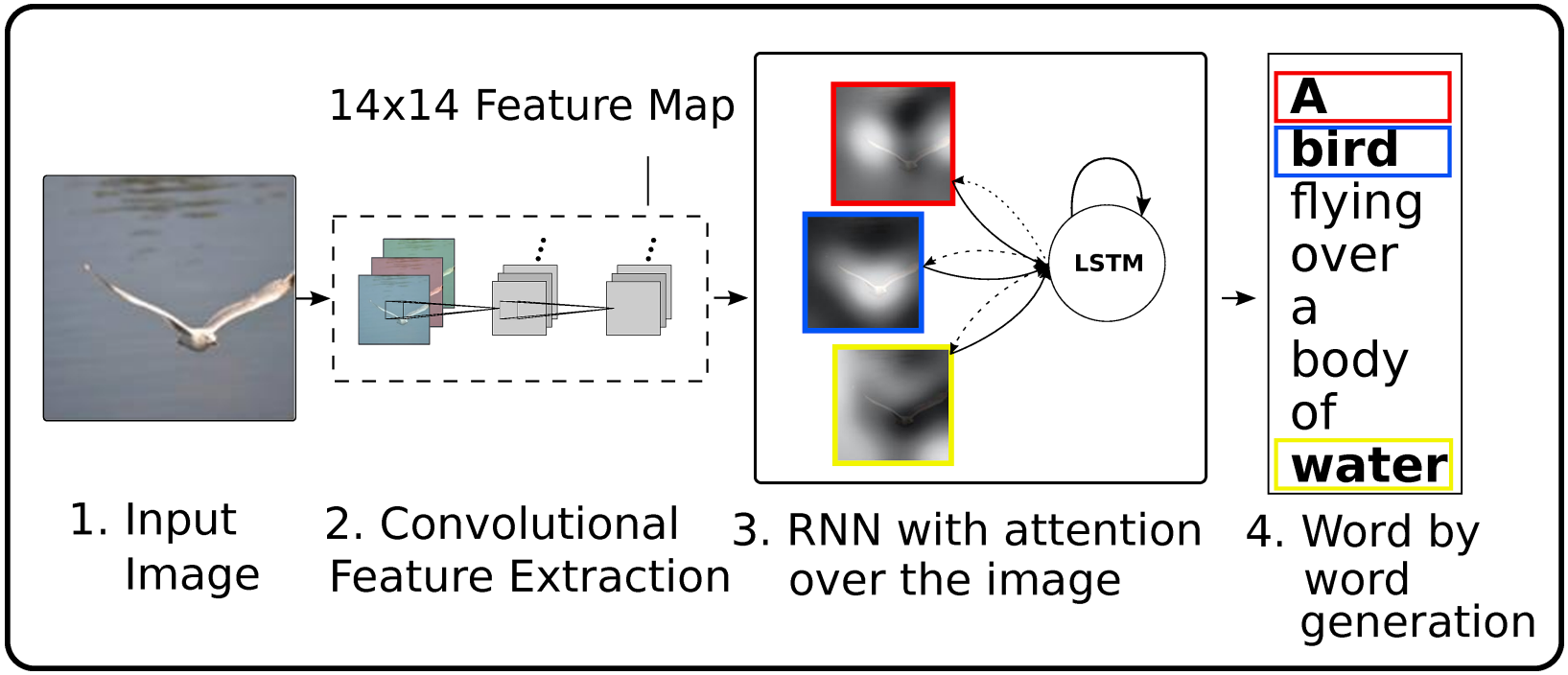
Attention is not limited for NMT …



The decoder focuses its **attention** at curtain spatial location when generating each word.

Image from: https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/

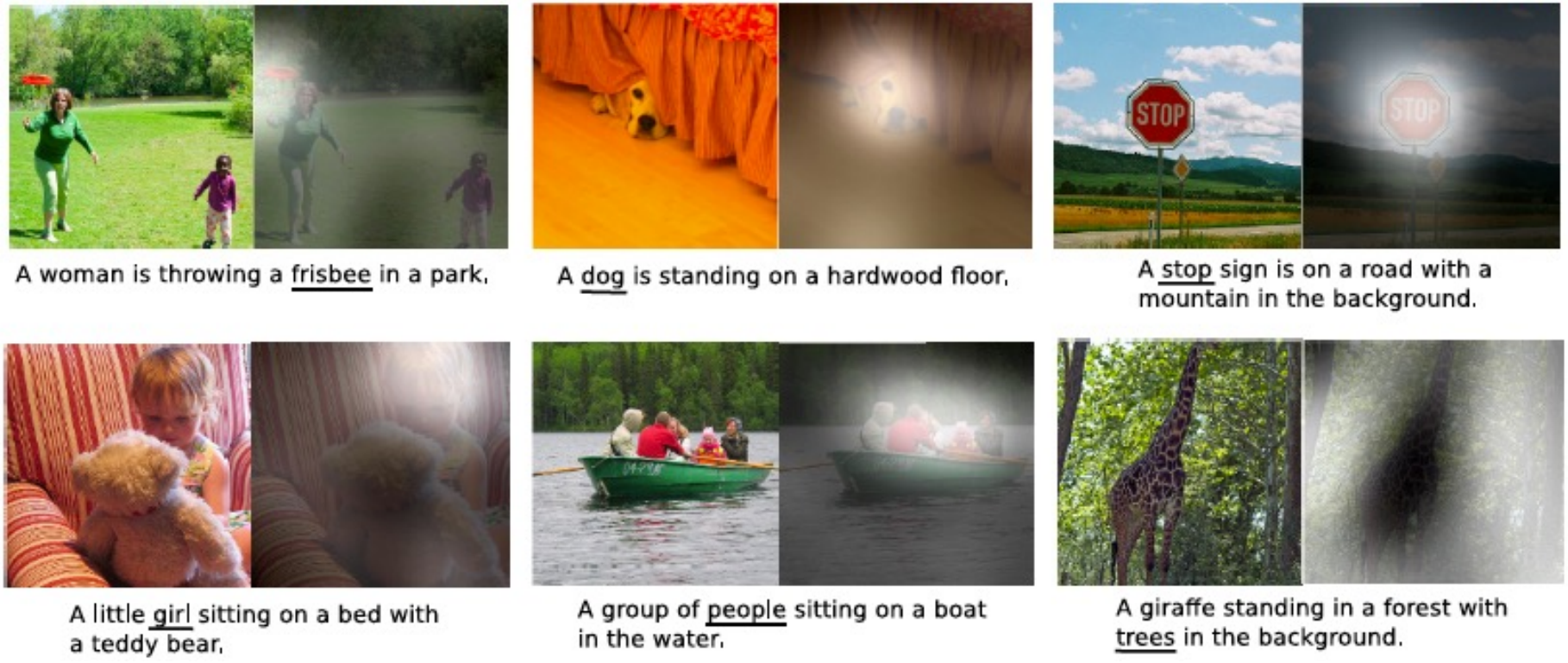
Attention-based Image Captioning [Xu-ICML15]



Xu et al, "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

Attention-based Image Captioning [Xu-ICML15]

Figure 4. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



Xu et al, "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

Attention-based Image Captioning [Xu-ICML15]

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

Xu et al, "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

References

- Thang Luong. Neural Machine Translation (seq2seq) Tutorial. https://github.com/tensorflow/nmt
- Blog by Raimi Karim. Attn: Illustrated Attention.2022. https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3