Seq2seq, NMT, Transformer

Milan Straka

May 03, 2021

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research, Development and Education

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

Sequence-to-Sequence Architecture

Sequence-to-Sequence Architecture

Sequence-to-Sequence is a name for an architecture allowing to produce an arbitrary output sequence $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M$ from an input sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

Unlike CRF/CTC, no assumptions are necessary and we condition each output sequence element on all input sequence elements and all already generated output sequence elements:

$$P(\boldsymbol{y}_i | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{i-1}).$$

Sequence-to-Sequence Architecture

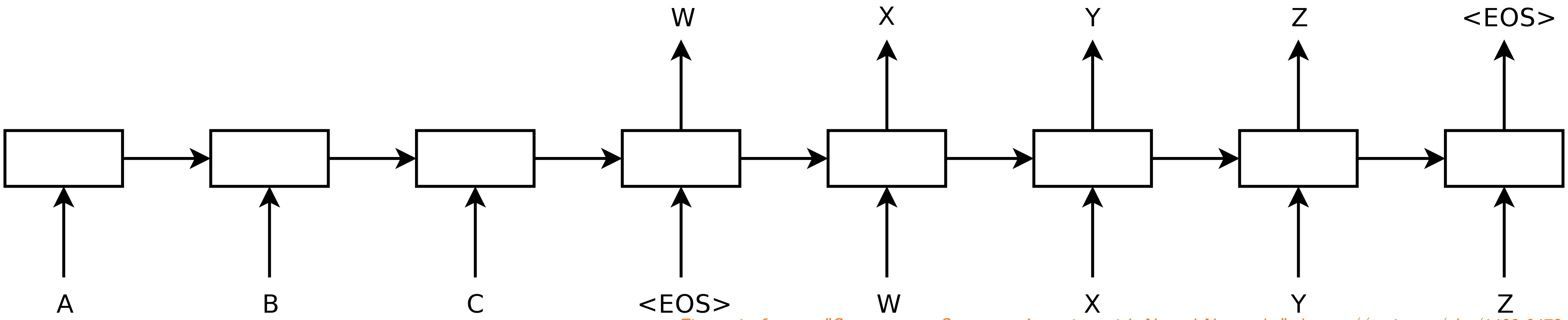


Figure 1 of paper "Sequence to Sequence Learning with Neural Networks", https://arxiv.org/abs/1409.0473.

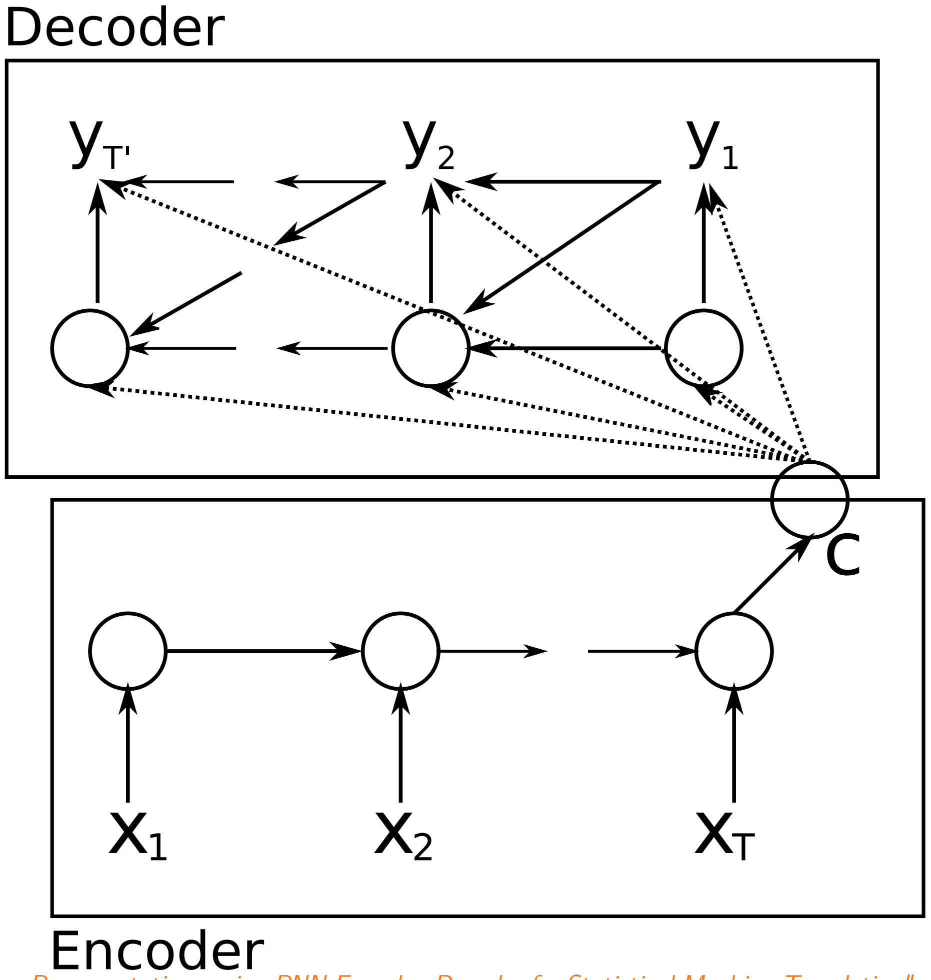
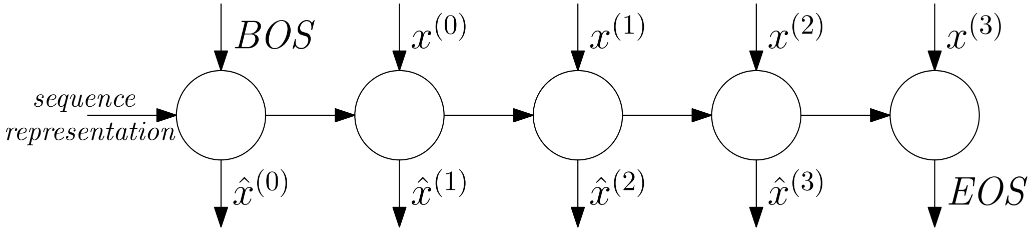


Figure 1 of paper "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation", https://arxiv.org/abs/1406.1078.

Training

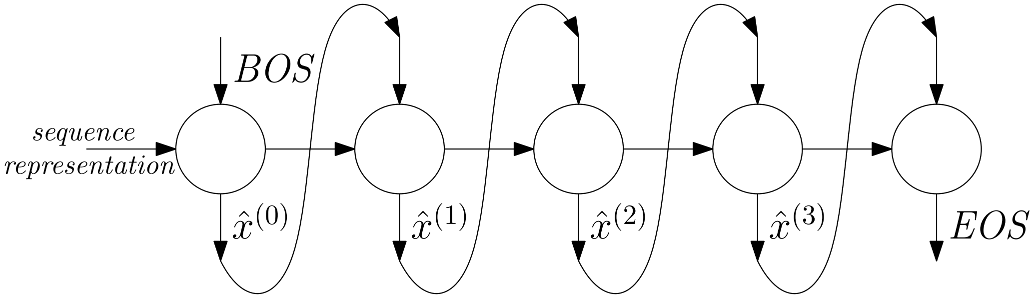
The so-called *teacher forcing* is used during training – the gold outputs are used as inputs during training.



Inference

During inference, the network processes its own predictions – such an approach is called **autoregressive decoding**.

Usually, the generated logits are processed by an $\arg\max$, the chosen word embedded and used as next input.



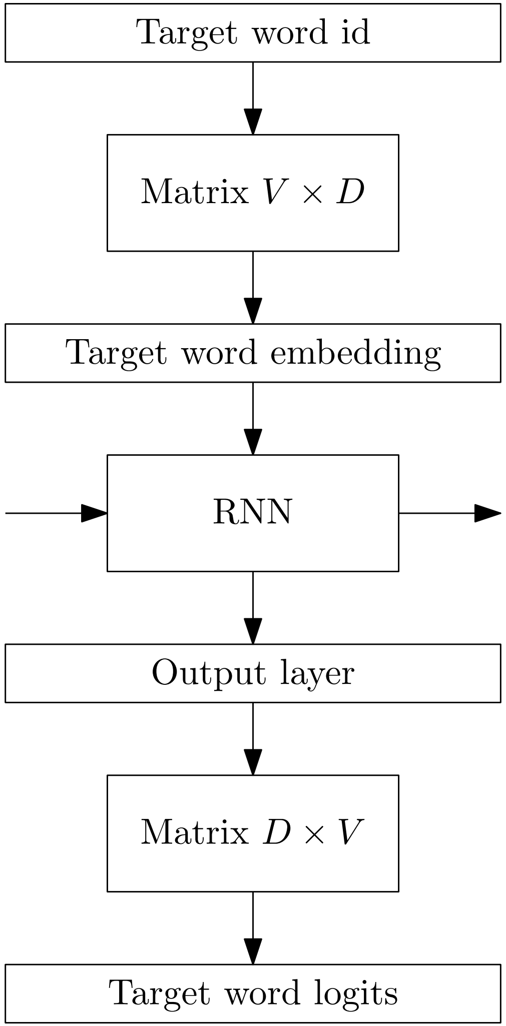
Tying Word Embeddings

In the decoder, we both:

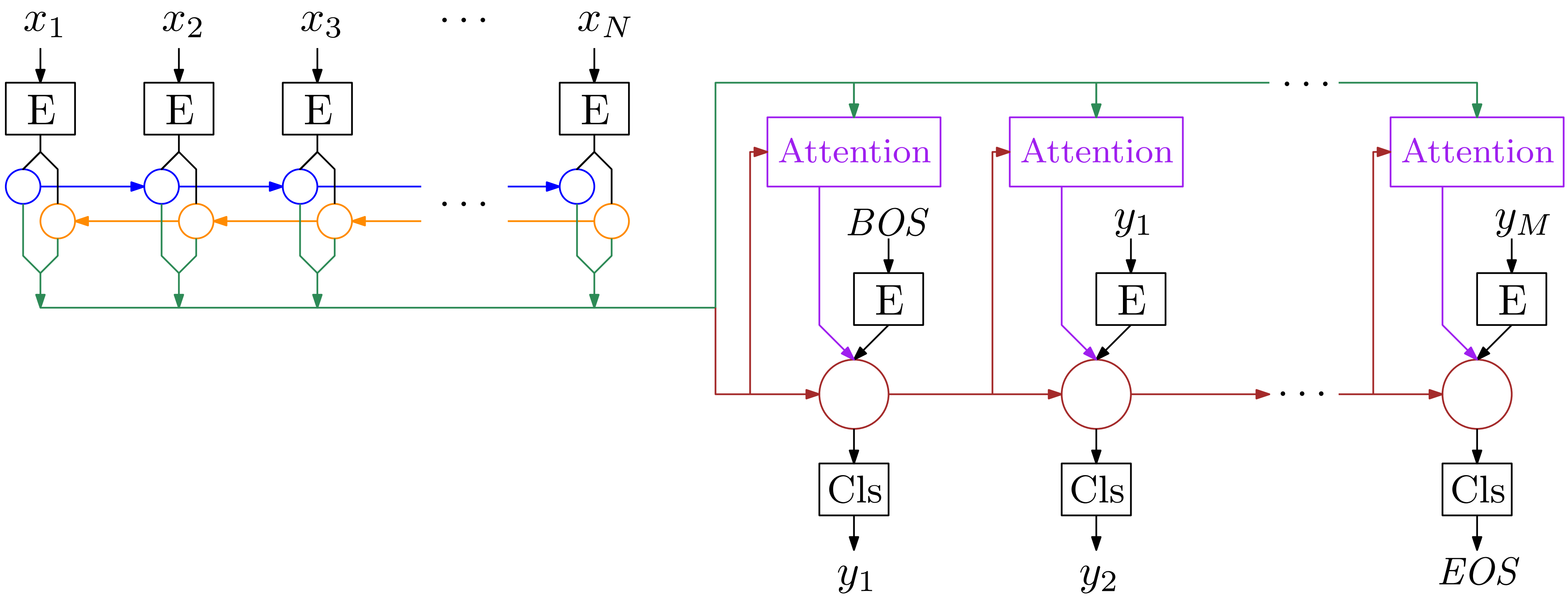
- embed the previous prediction, using a matrix of size $\mathbb{R}^{V \times D}$, where V is the vocabulary size and D is the embedding size;
- classify the hidden state into current prediction, using a matrix of size $\mathbb{R}^{D \times V}$.

Both these matrices have the same meaning – they represent the target-side words in the embedding space (the first explicitly represents the words by these embeddings, the second chooses the embedding in a sense "closest" to the produced hidden state).

Therefore, it makes sense to **tie** these matrices, i.e., to represent one of them as a transposition of the other.



Attention



x_1
x_2
x_3
$\cdots$
x_N
E
E
E
E
$\cdots$
Attention
Attention
Attention
$\cdots$
BOS
y_1
y_M
E
E
E
$\cdots$
Cls
Cls
Cls
y_1
y_2
EOS

As another input during decoding, we add *context vector* c_i:

$$\boldsymbol{s}_i = f(\boldsymbol{s}_{i-1}, \boldsymbol{y}_{i-1}, \boldsymbol{c}_i).$$

We compute the context vector as a weighted combination of source sentence encoded outputs:

$$\boldsymbol{c}_i = \sum_j \alpha_{ij} \boldsymbol{h}_j$$

The weights α_{ij} are softmax of e_{ij} over j,

$$\boldsymbol{\alpha}_i = \operatorname{softmax}(\boldsymbol{e}_i),$$

with e_{ij} being

$$e_{ij} = \boldsymbol{v}^\top \tanh(\boldsymbol{V}\boldsymbol{h}_j + \boldsymbol{W}\boldsymbol{s}_{i-1} + \boldsymbol{b}).$$

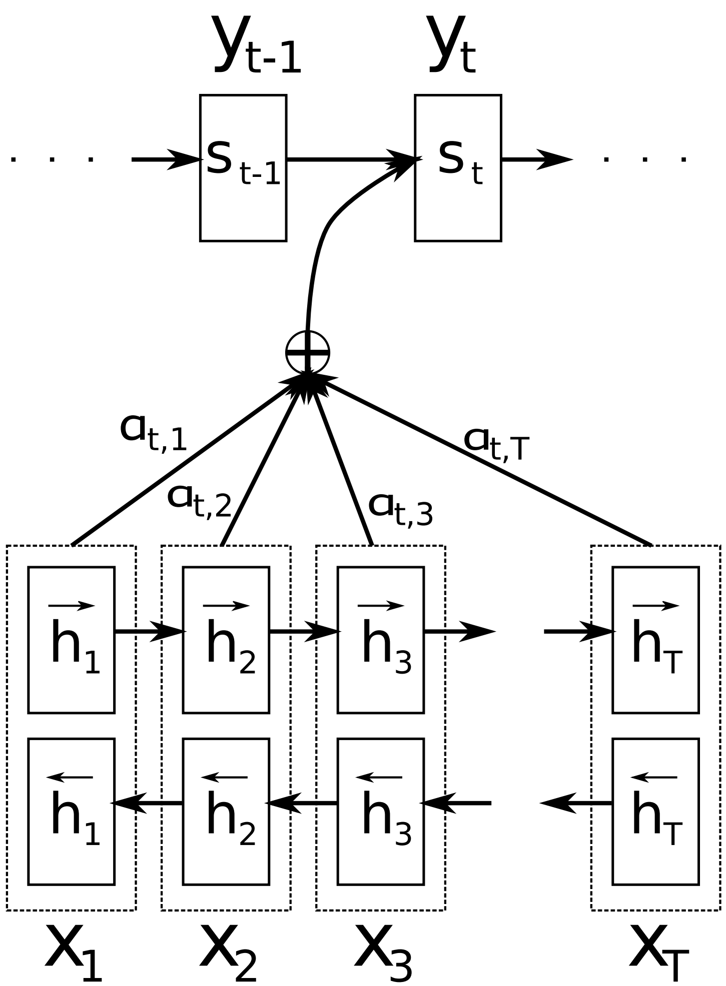
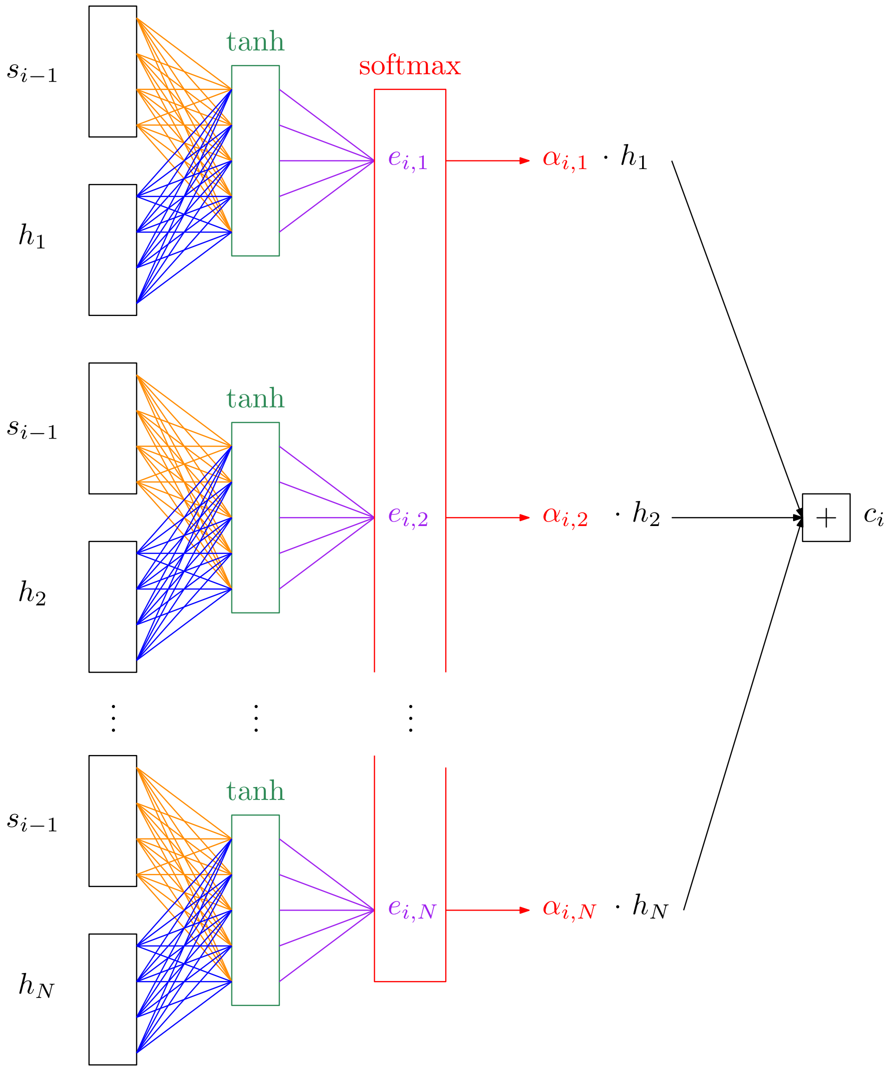


Figure 1 of paper "Neural Machine Translation by Jointly Learning to Align and Translate", https://arxiv.org/abs/1409.0473.

Attention Implementation



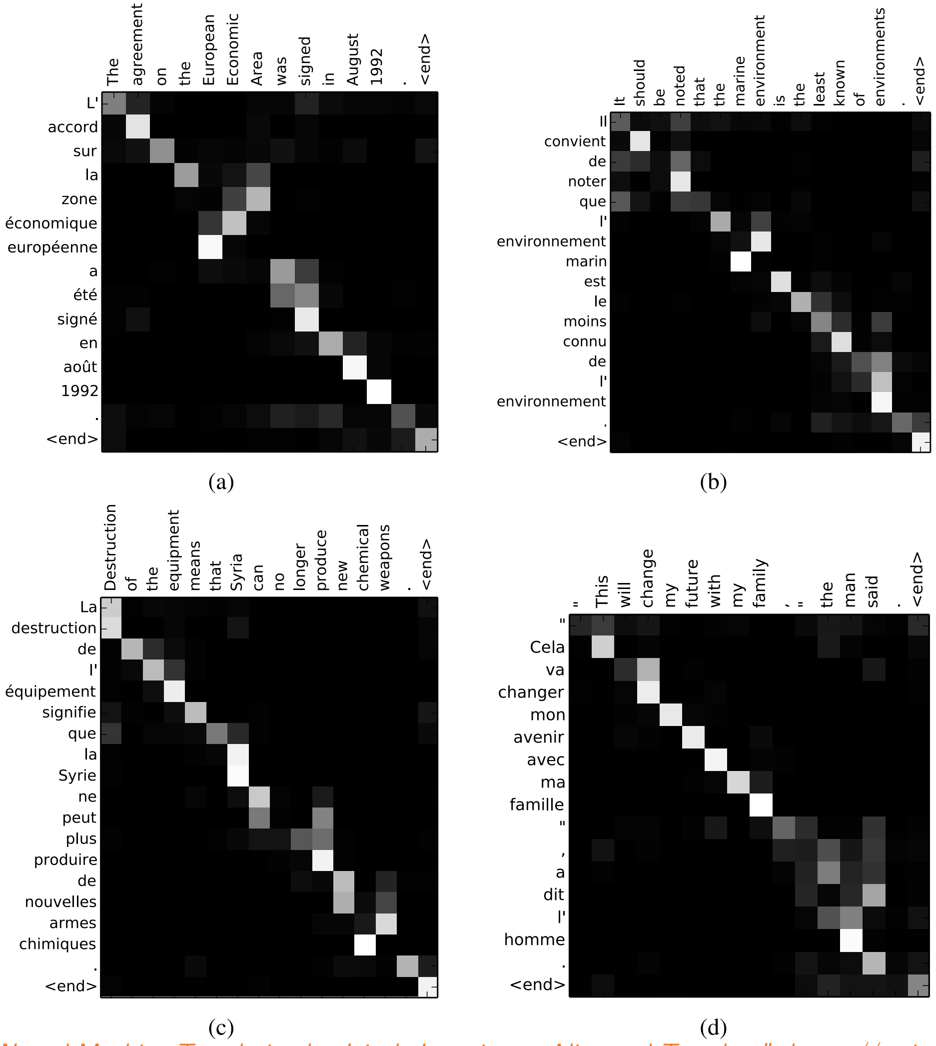


Figure 3 of paper "Neural Machine Translation by Jointly Learning to Align and Translate", https://arxiv.org/abs/1409.0473.

Subword Units

Translate *subword units* instead of words. The subword units can be generated in several ways, the most commonly used are:

- **BPE**: Using the *byte pair encoding* algorithm. Start with individual characters plus a special end-of-word symbol $\cdot$. Then, merge the most occurring symbol pair A, B by a new symbol AB, with the symbol pair never crossing word boundary (so that the end-of-word symbol cannot be inside a subword).

  Considering a dictionary with words *low, lowest, newer, wider*, a possible sequence of merges:

$$
\begin{aligned}
r \;\; \cdot &\rightarrow r\cdot \\
l \;\; o &\rightarrow lo \\
lo \;\; w &\rightarrow low \\
e \;\; r\cdot &\rightarrow er\cdot
\end{aligned}
$$

Subword Units

- **Wordpieces**: Given a text divided into subwords, we can compute unigram probability of every subword, and then get the likelihood of the text under a unigram language model by multiplying the probabilities of the subwords in the text.

  When we have only a text and a subword dictionary, we divide the text in a greedy fashion, iteratively choosing the longest existing subword.

  When constructing the subwords, we again start with individual characters, and then repeatedly join such a pair of subwords, which increases the unigram language model likelihood the most.

Both approaches give very similar results; a biggest difference is that during the inference:

- for BPE, the sequence of merges must be performed in the same order as during the construction of the BPE;
- for Wordpieces, it is enough to find longest matches from the subword dictionary.

Usually quite little subword units are used (32k-64k), often generated on the union of the two vocabularies (the so-called *joint BPE* or *shared wordpieces*).

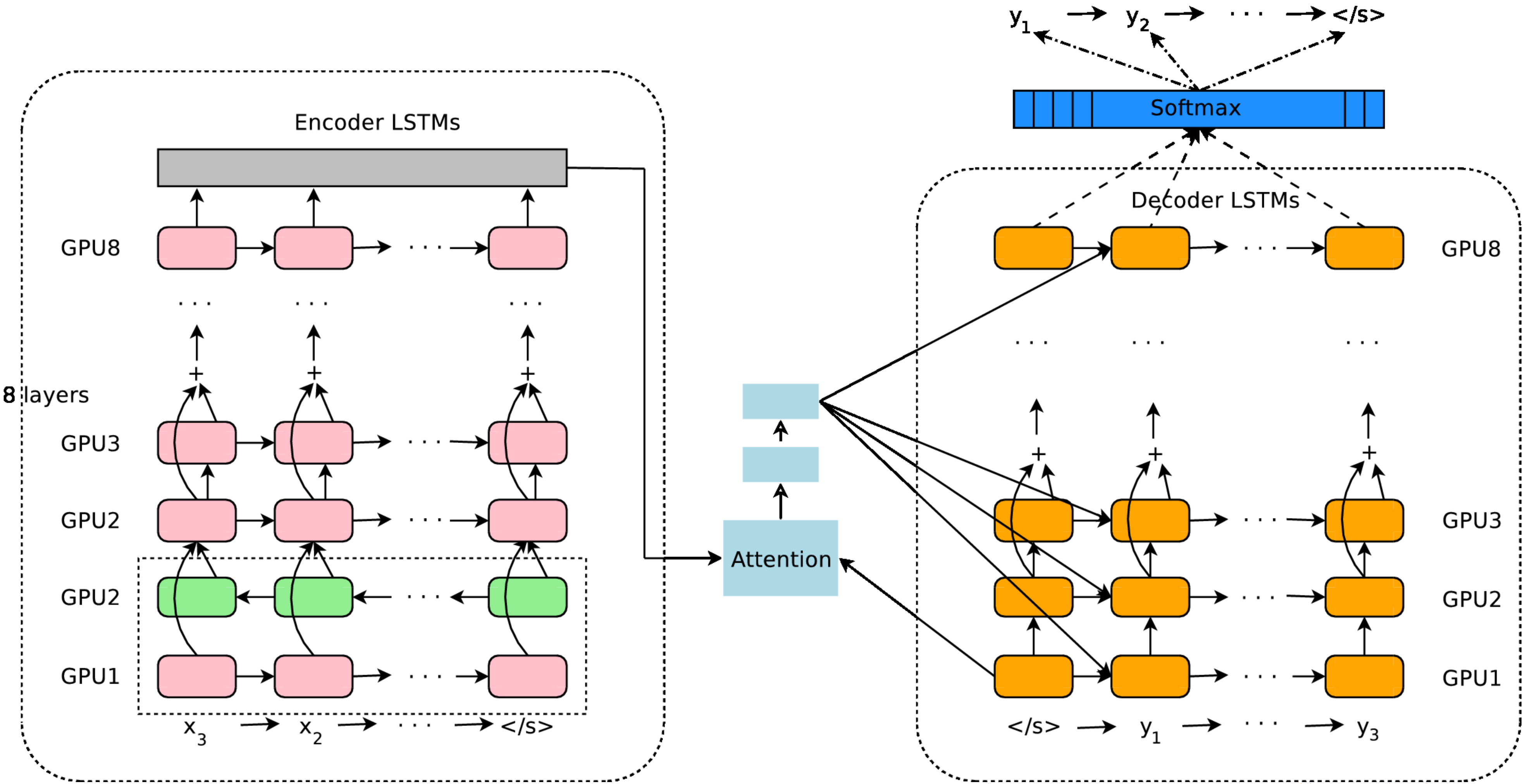


Figure 1 of paper "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", https://arxiv.org/abs/1609.08144.

Google NMT

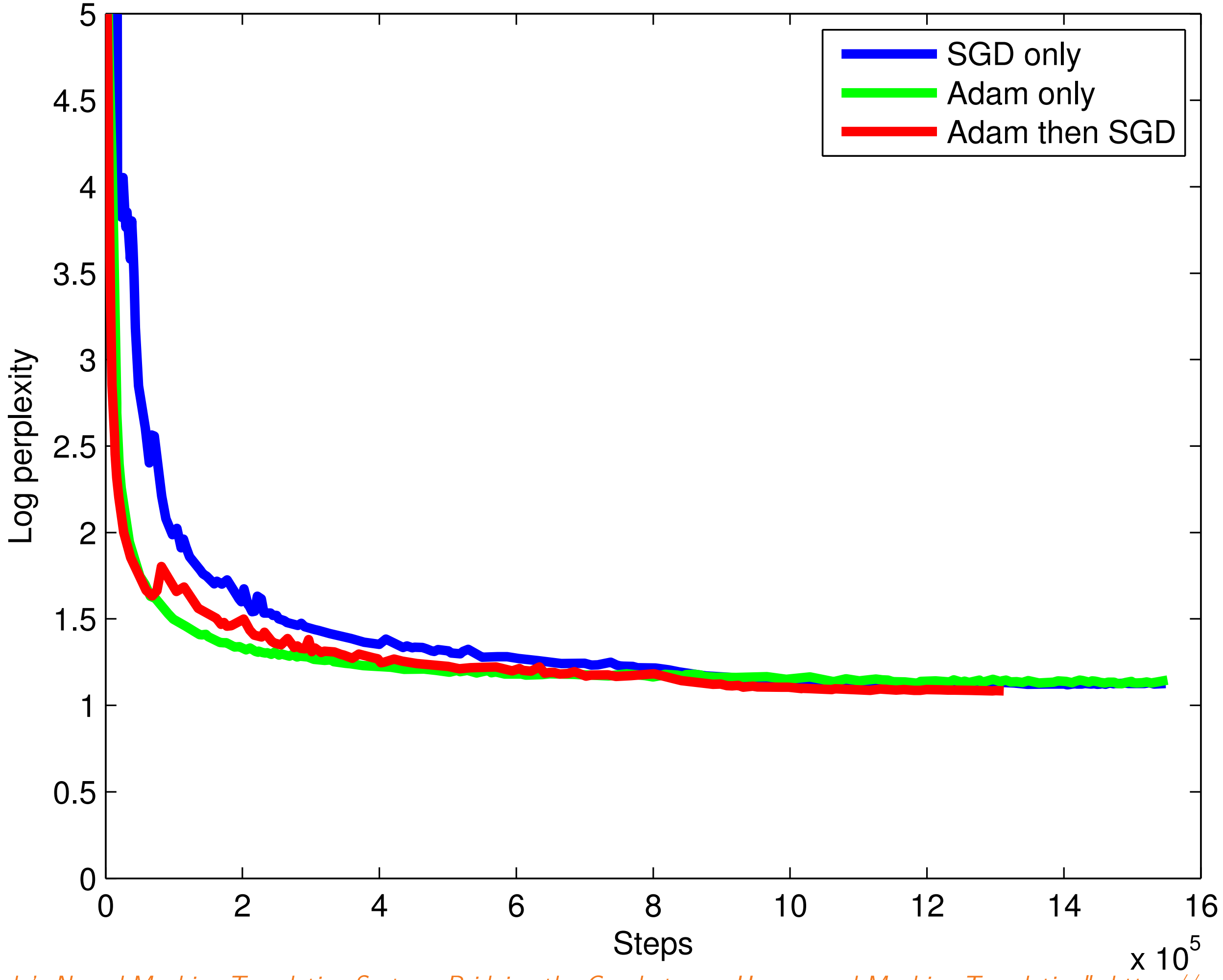


Figure 5 of paper "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", https://arxiv.org/abs/1609.08144.

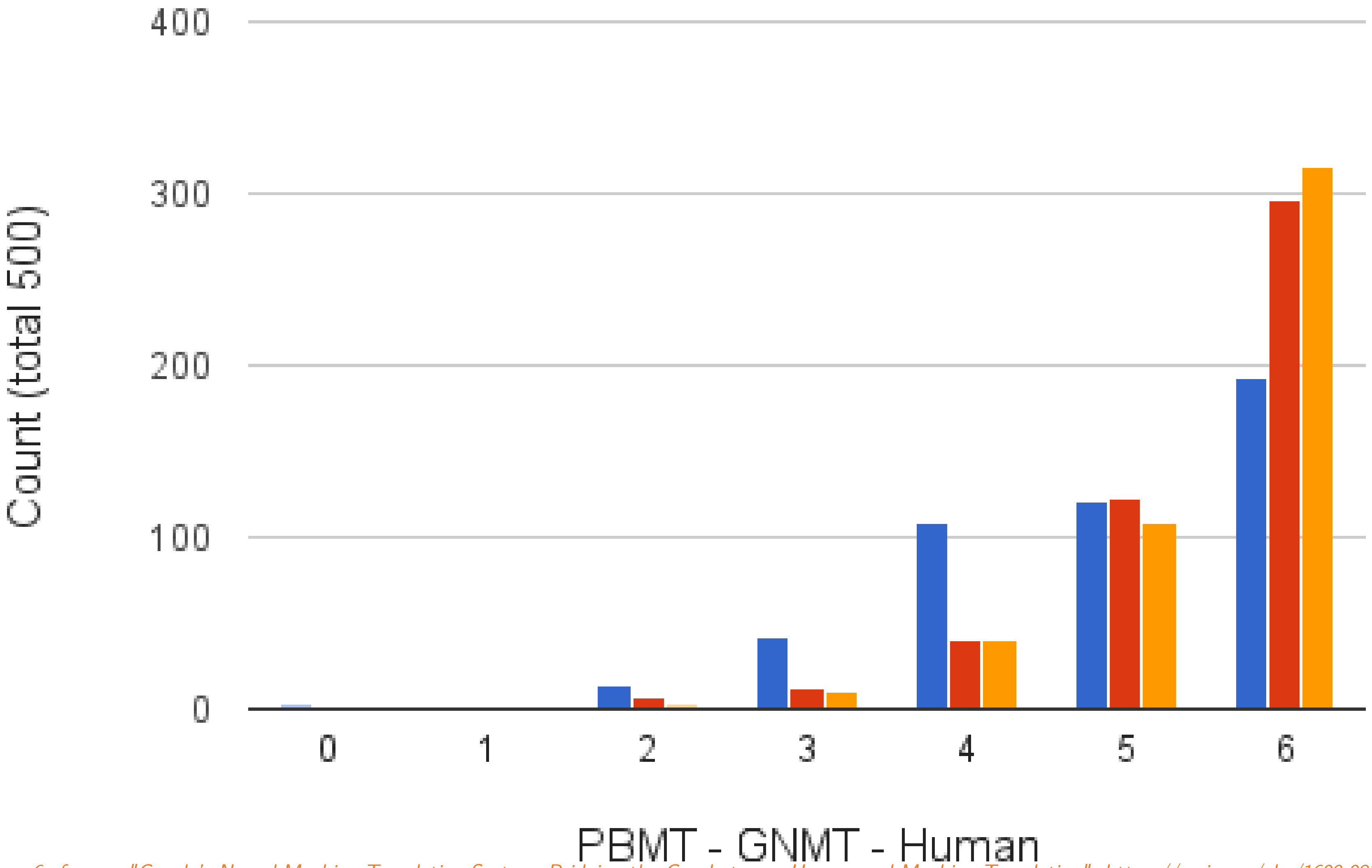


Figure 6 of paper "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", https://arxiv.org/abs/1609.08144.

Beyond one Language Pair

Fig. 5. A selection of evaluation results, grouped by human rating.

Figure 5 of "Show and Tell: Lessons learned from the 2015 MSCOCO...", https://arxiv.org/abs/1609.06647.

Beyond one Language Pair

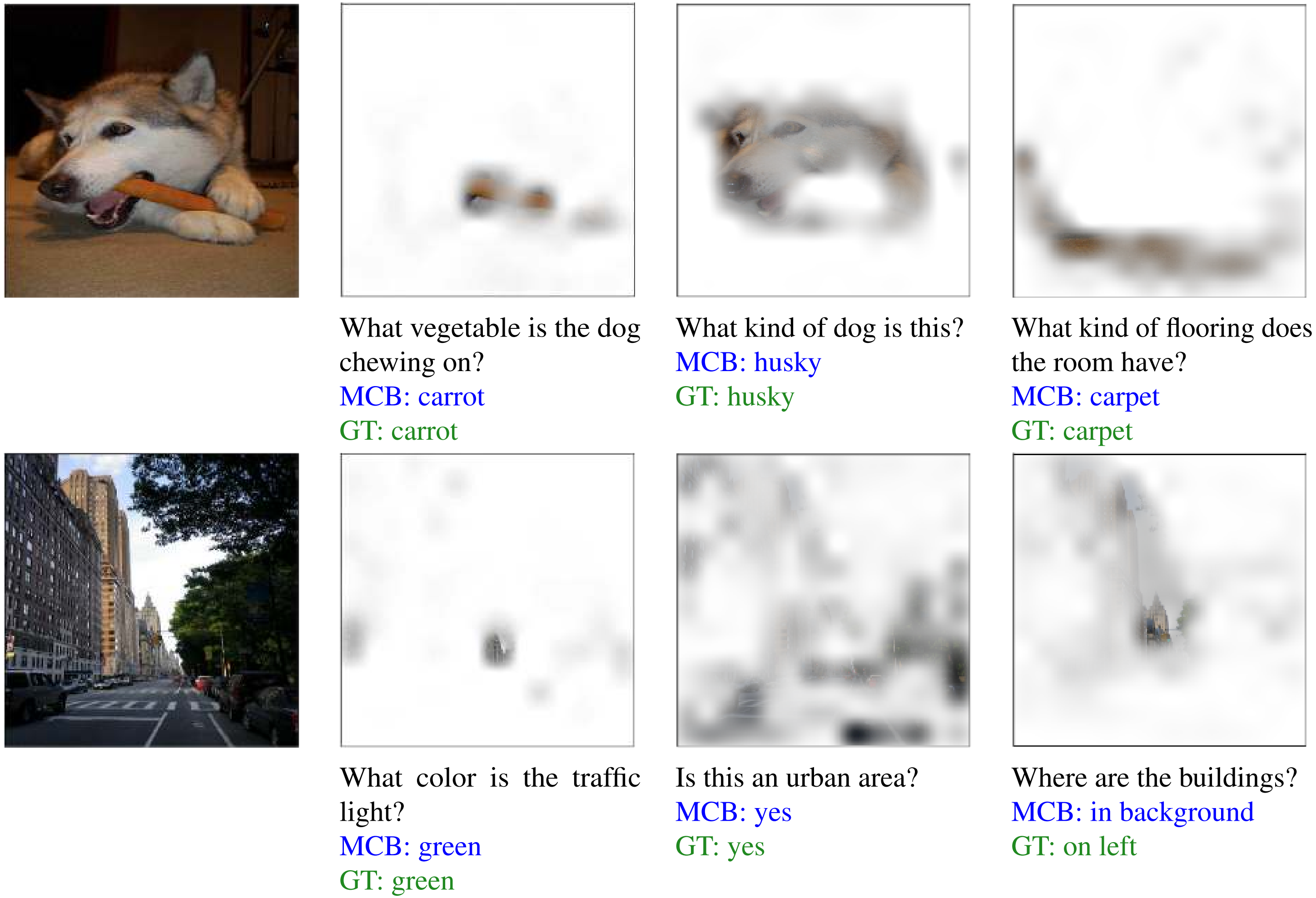


Figure 6 of "Multimodal Compact Bilinear Pooling for VQA and Visual Grounding", https://arxiv.org/abs/1606.01847.

Multilingual and Unsupervised Translation

Many attempts at multilingual translation.

- Individual encoders and decoders, shared attention.
- Shared encoders and decoders.

Surprisingly, even unsupervised translation is attempted lately. By unsupervised we understand settings where we have access to large monolingual corpora, but no parallel data.

In 2019, the best unsupervised systems were on par with the best 2014 supervised systems.

		WMT-14			
		fr-en	en-fr	de-en	en-de
Unsupervised	Proposed system	33.5	36.2	27.0	22.5
	*detok. SacreBLEU**	33.2	33.6	26.4	21.2
Supervised	WMT best*	35.0	35.8	29.0	20.6†
	Vaswani et al. (2017)	-	41.0	-	28.4
	Edunov et al. (2018)	-	45.6	-	35.0

Table 3: Results of the proposed method in comparison to different supervised systems (BLEU).

Table 3 of paper "An Effective Approach to Unsupervised Machine Translation", https://arxiv.org/abs/1902.01313.

Attention is All You Need

For some sequence processing tasks, *sequential* processing (as performed by recurrent neural networks) of its elements might be too restrictive.

Instead, we may want to be able to combine sequence elements independently on their distance.

Such processing is allowed in the *Transformer* architecture, originally proposed for neural machine translation in 2017 in *Attention is All You Need* paper.

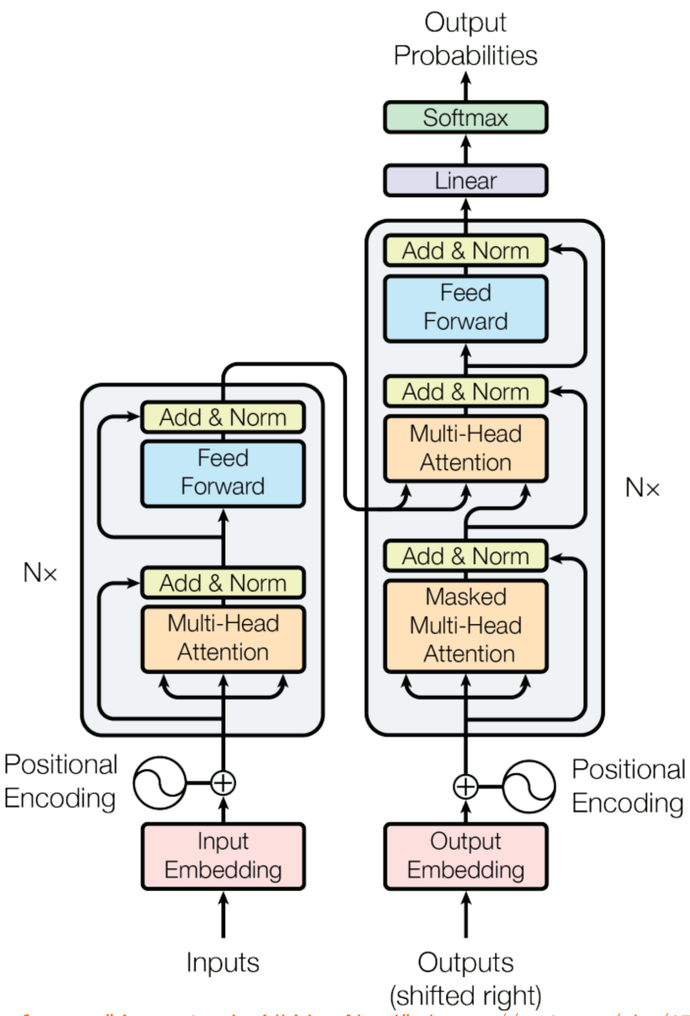
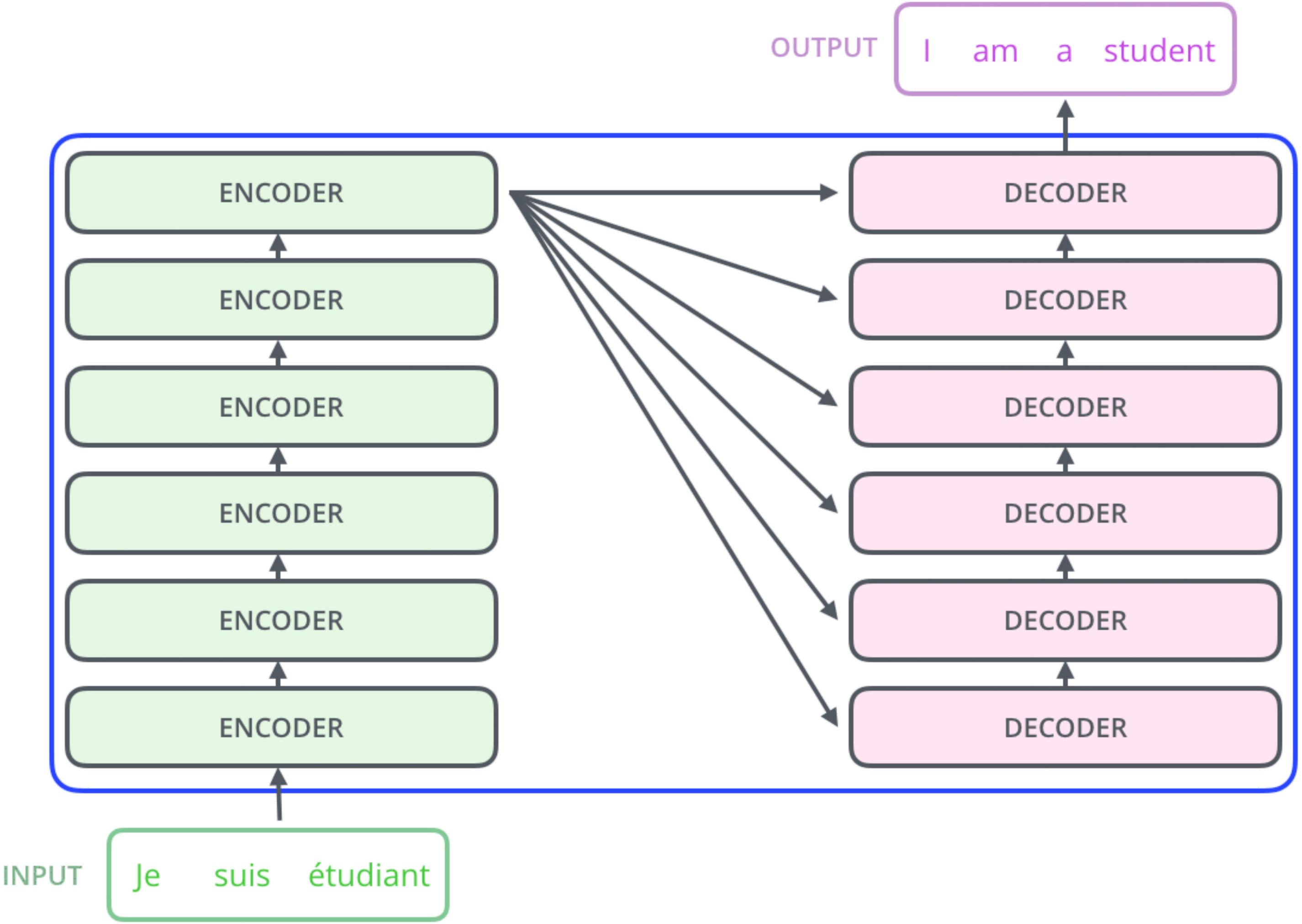


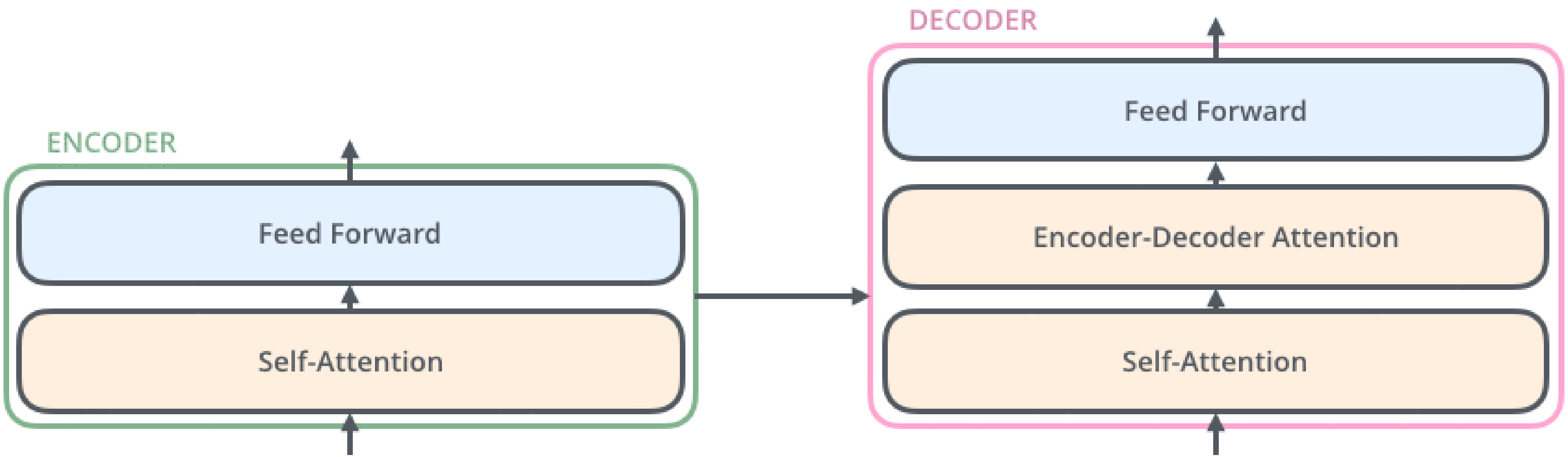
Figure 1 of paper "Attention Is All You Need", https://arxiv.org/abs/1706.03762

Transformer

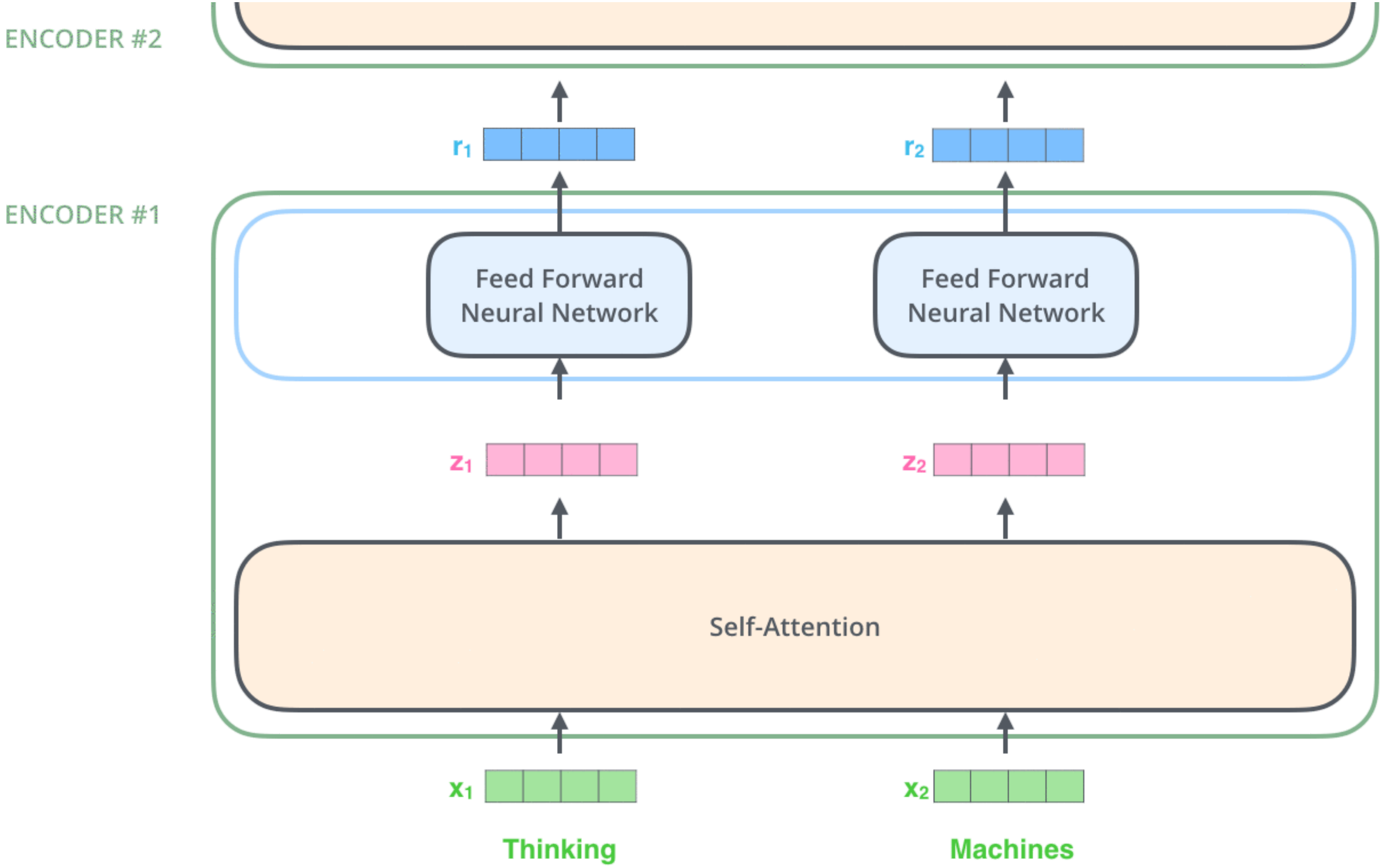


http://jalammar.github.io/images/t/The_transformer_encoder_decoder_stack.png

Transformer



http://jalammar.github.io/images/t/Transformer_decoder.png



http://jalammar.github.io/images/t/encoder_with_tensors_2.png

Transformer – Self-Attention

Assume that we have a sequence of n words represented using a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$.

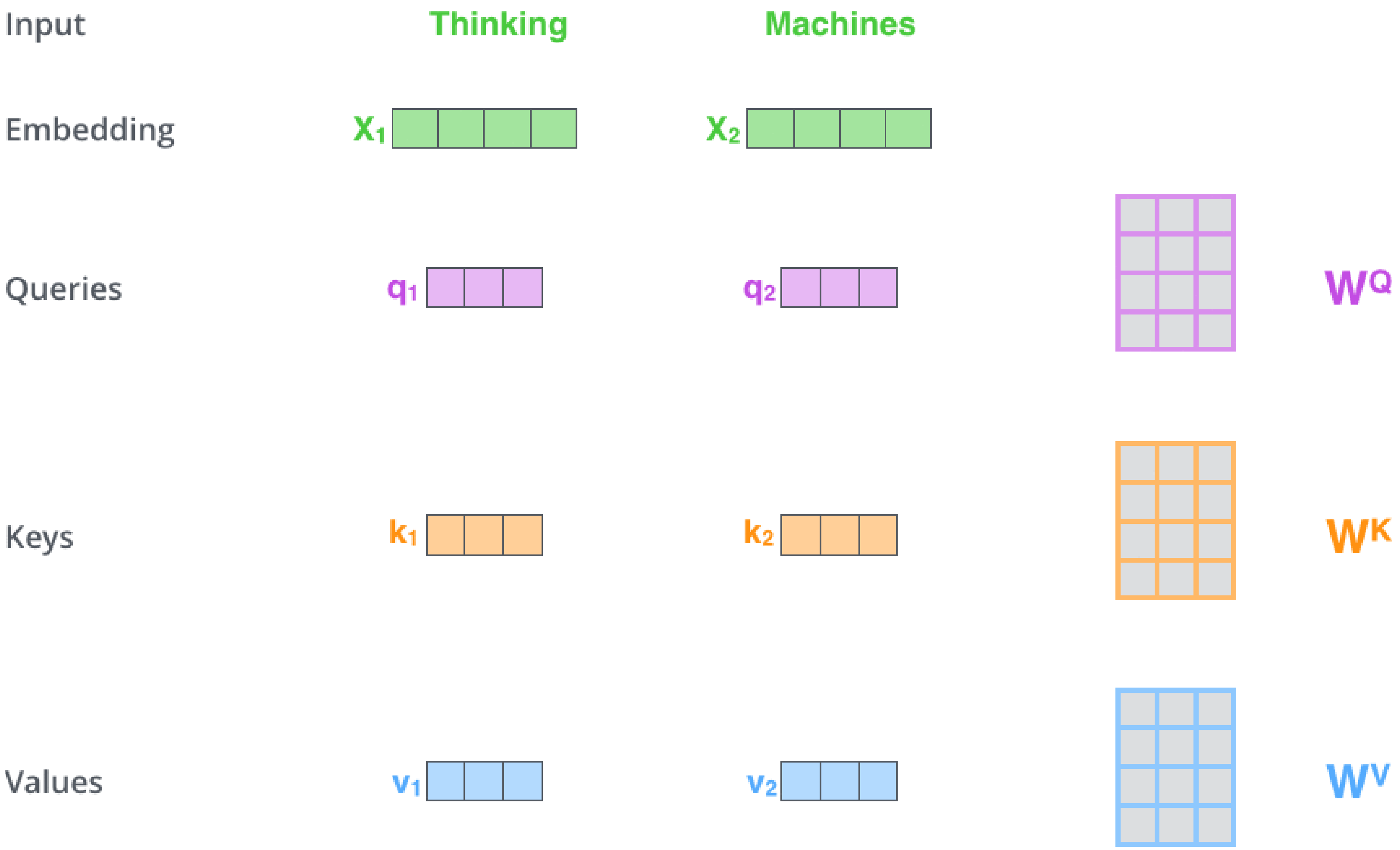
The attention module for a queries $\boldsymbol{Q} \in \mathbb{R}^{n\times d_k}$, keys $\boldsymbol{K} \in \mathbb{R}^{n\times d_k}$ and values $\boldsymbol{V} \in \mathbb{R}^{n\times d_v}$ is defined as:

$$\operatorname{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \operatorname{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}.$$

The queries, keys and values are computed from the input word representations $\boldsymbol{X}$ using a linear transformation as

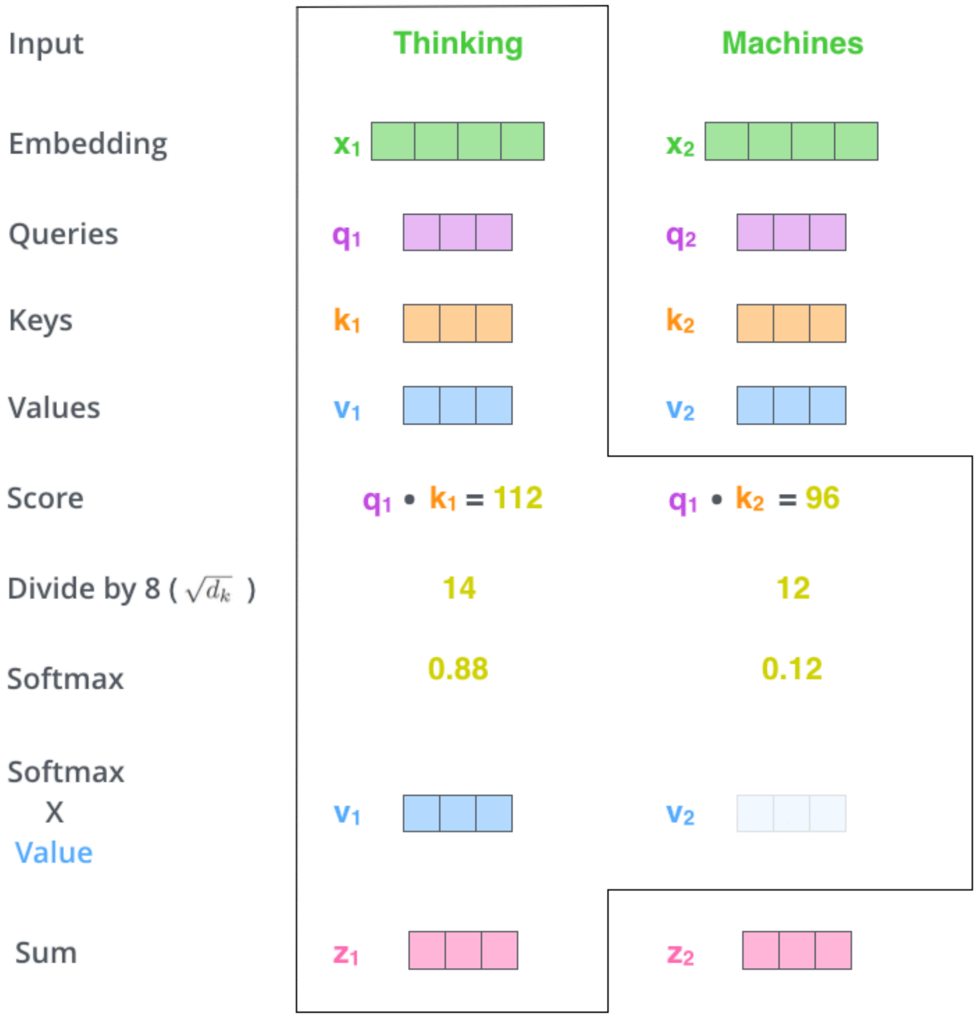
$$\boldsymbol{Q} = \boldsymbol{W}^Q \cdot \boldsymbol{X}$$
$$\boldsymbol{K} = \boldsymbol{W}^K \cdot \boldsymbol{X}$$
$$\boldsymbol{V} = \boldsymbol{W}^V \cdot \boldsymbol{X}$$

Transformer – Self-Attention

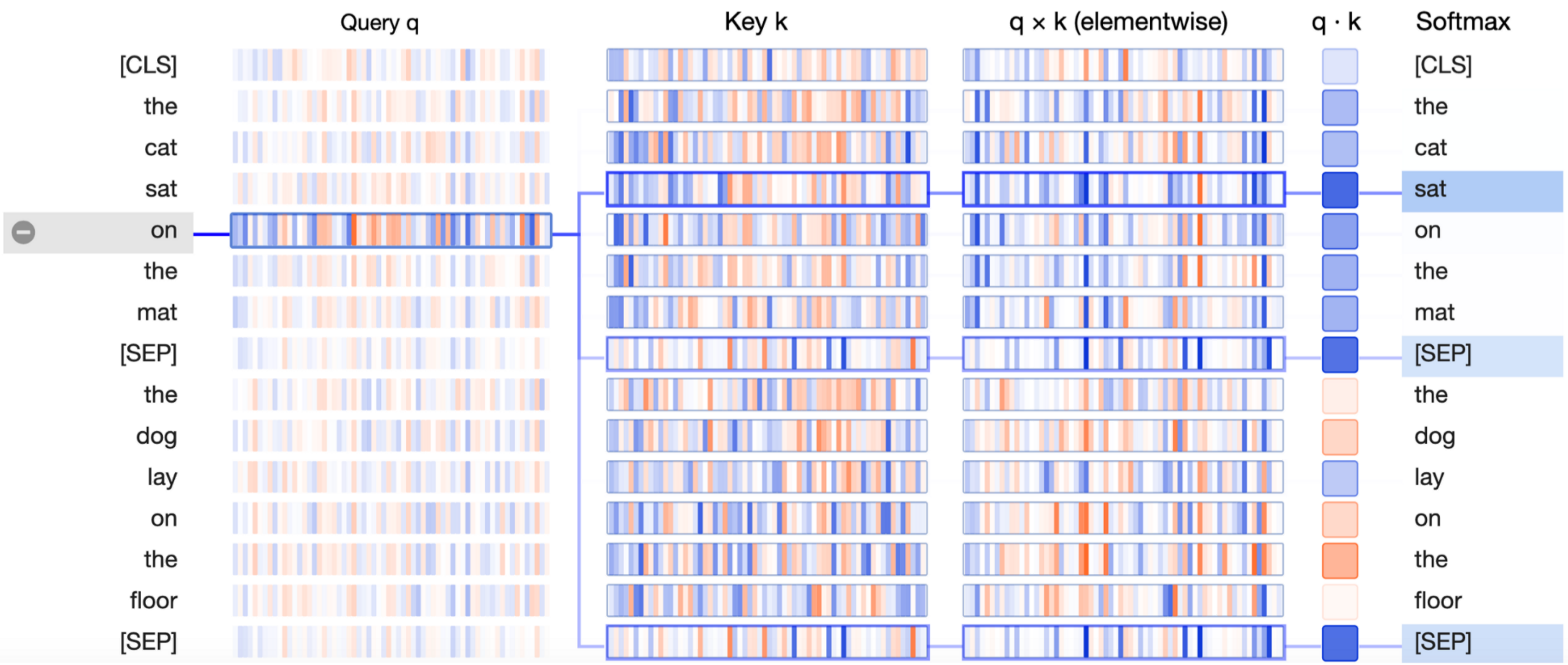


http://jalammar.github.io/images/t/transformer_self_attention_vectors.png

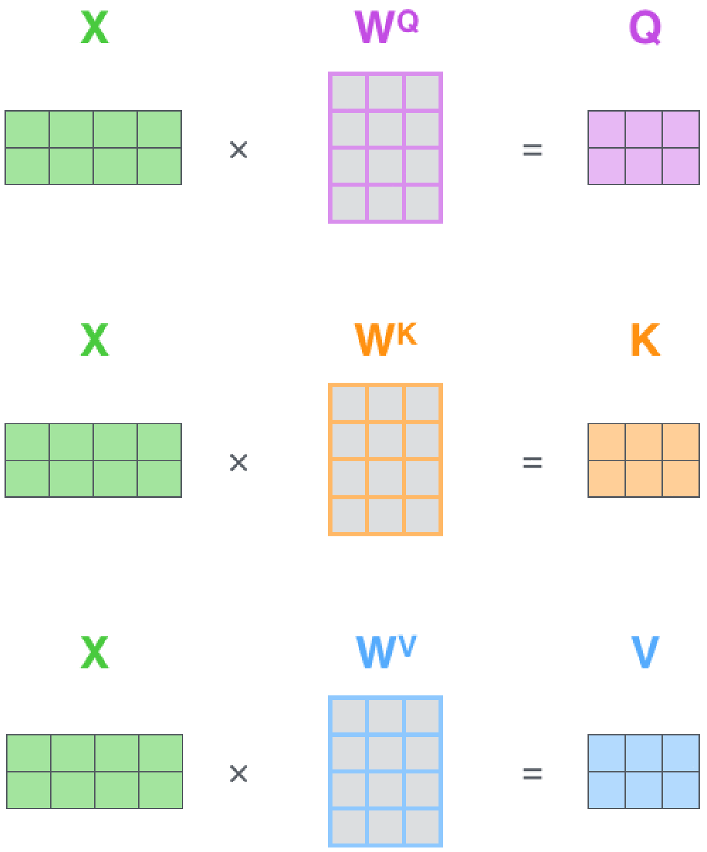
Transformer – Self-Attention



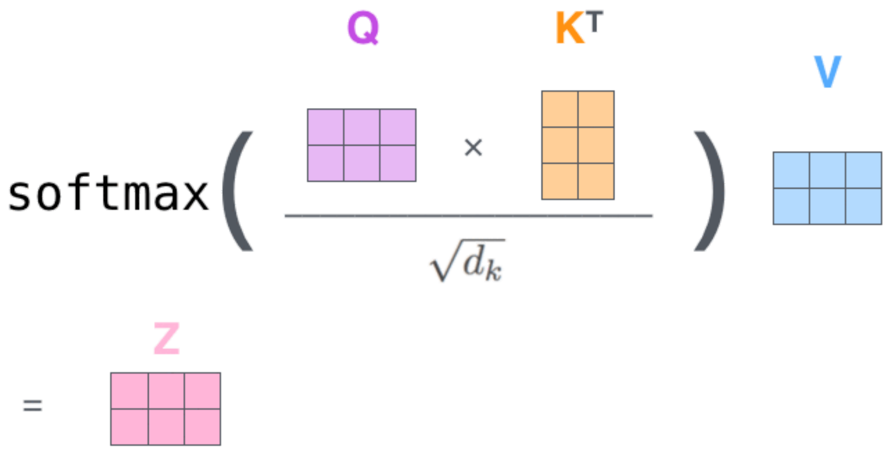
http://jalammar.github.io/images/t/self-attention-output.png



https://miro.medium.com/max/2000/1*jBsfVNOOcJ-I3tsLVgni_w.png



http://jalammar.github.io/images/t/self-attention-matrix-calculation.png



http://jalammar.github.io/images/t/self-attention-matrix-calculation-2.png