MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Attention Is All You Need

Muhammad Umar Salman
umar.salman@mbzuai.ac.ae

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*,†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*,‡]
illia.polosukhin@gmail.com

NIPS 2017

Agenda

Motivation → **Summary** → **Strengths** → **Weaknesses** → **Methodology**

Motivation

- In S2S problems e.g. Neural Machine Translation, RNN with Encoder + Decoder architecture was the go-to solution.
- However, working with long sequences, the architecture loses its ability to retain information from first elements.
- Decoder has access to last hidden state of Encoder stack making it lose or forget early information, for this reason **Attention Mechanism** was introduced.
- The decoder now looks at all the states of the encoder, being able to access information about all the elements of the input sequence.

Summary

- **Transformers**, first sequence transduction model replacing recurrent layers used in Encoder-Decoder architecture to Multi-head self-attention.
- Its architecture makes use of stacked self-attention and point-wise, fully connected layers
- Experiments on WMT-2014 datasets (En-Fr & En-Ger) show that this approach achieves higher BLEU score compared to previous models or reported ensembles
- Approach has had significant impact on introducing new state-of-the-art language models and vision models such as BERT, GPT-2 and ViT.
- Observed Transformer generalized well when applied to other tasks such as English constituency parsing.

Strengths

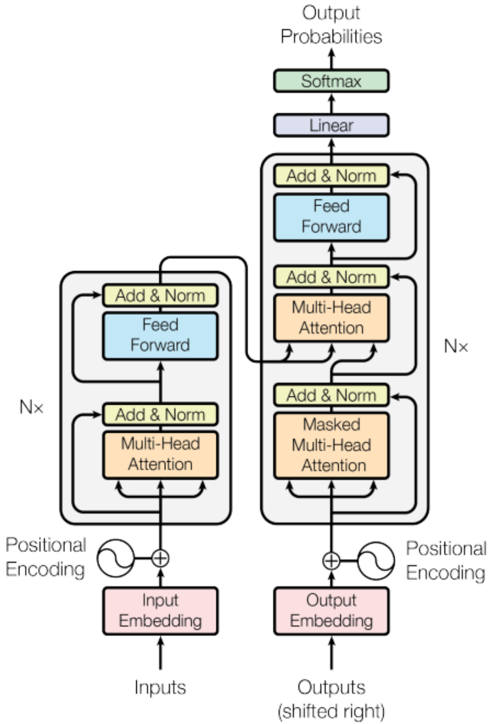
- Propose a **novel architecture** based on Attention Mechanism
- Established a **new state-of-the-art BLEU score** on machine translation tasks
- The architecture is **highly efficient** as architecture reduces model complexity which **reduces the training cost**
- Achieves **parallelizability** within the architecture to speed up training.
- Well-written, well-structured and experiments are clearly set-up for replication purposes.

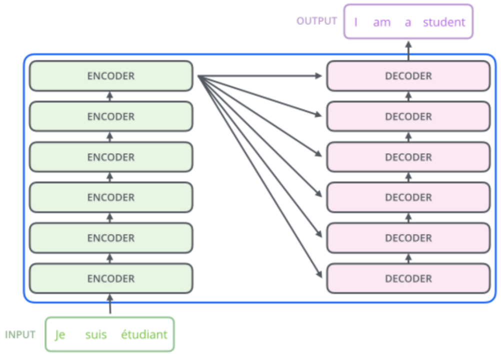
Weaknesses

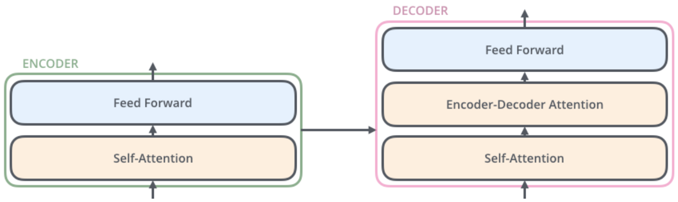
- O(n^2.d), larger the input the higher the time it will take to make inferences/train.
- Paper lacks in-depth mathematical and methodological details by giving a high overview. Reader has to go over code to fully understand the nitty-gritty details.
- Paper argues that retains information for long input sequences. It should have maybe specifically evaluated on longer sequences and its compare it with other models
- Although Tables are very well formatted table column names should be properly described in the footnote of the image for ease of the reader.

Methodology

General Architecture

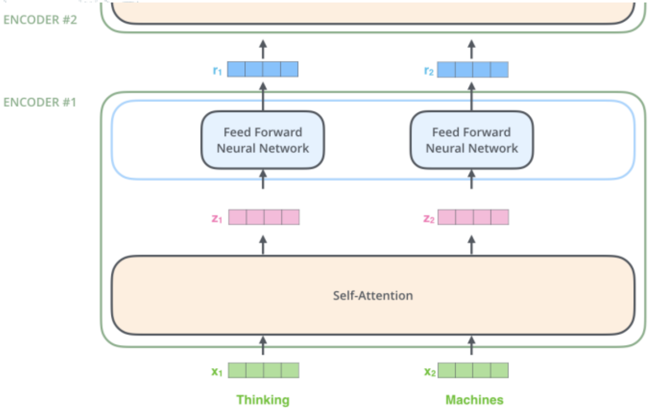






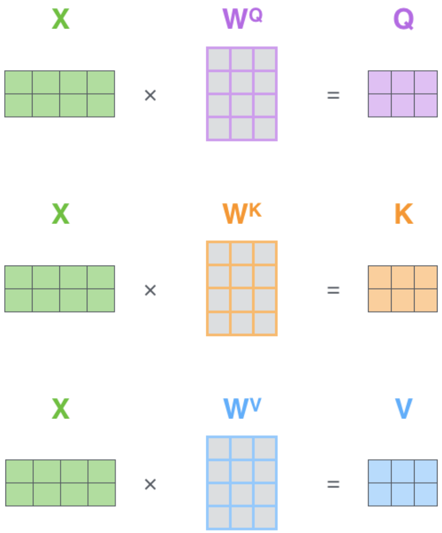
Methodology

Embedding + Positional Encoding



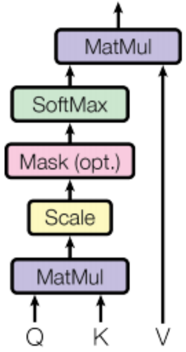
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Key, Value & Query

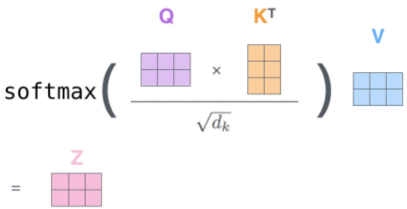


Self Attention / Attention Score / Attention Head

Scaled Dot-Product Attention

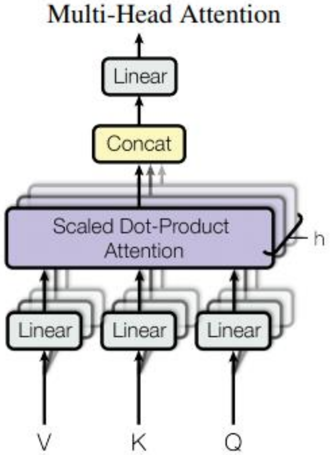


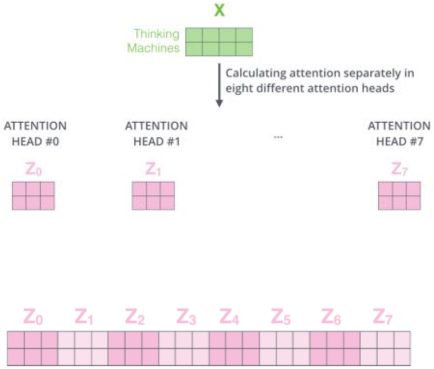
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

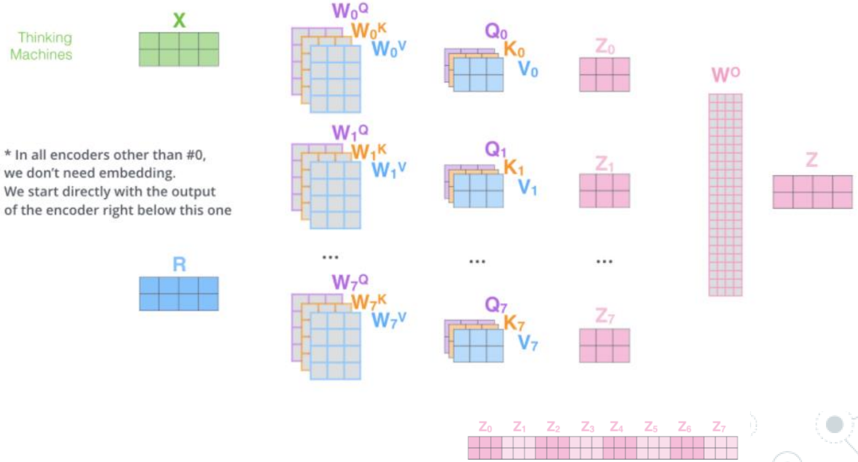


Methodology

Multi-Head Attention



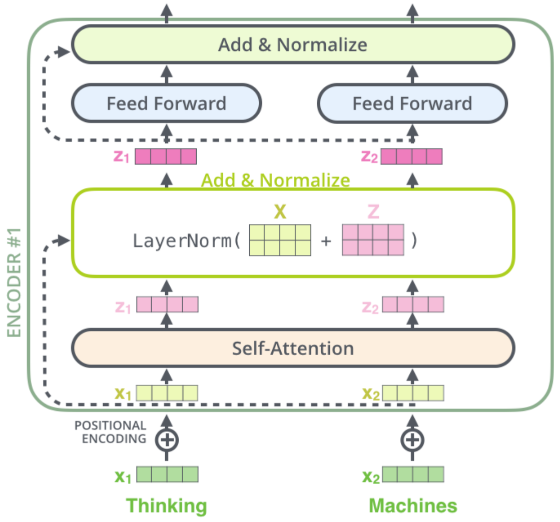


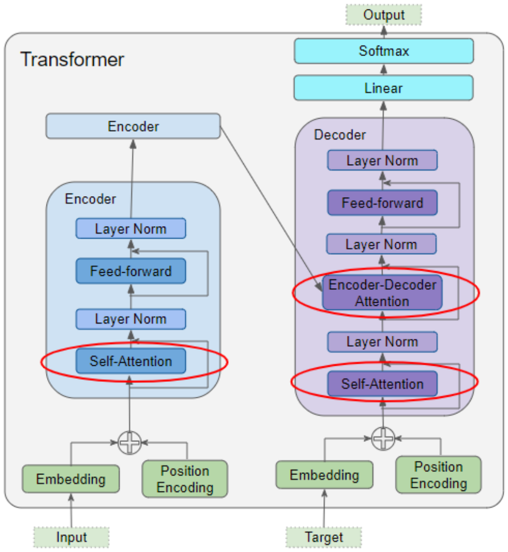


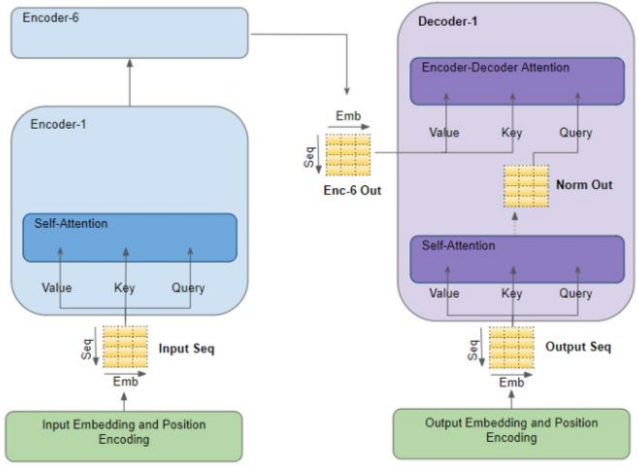
$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O$$
$$\text{where } \mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Methodology

Encoder Decoder



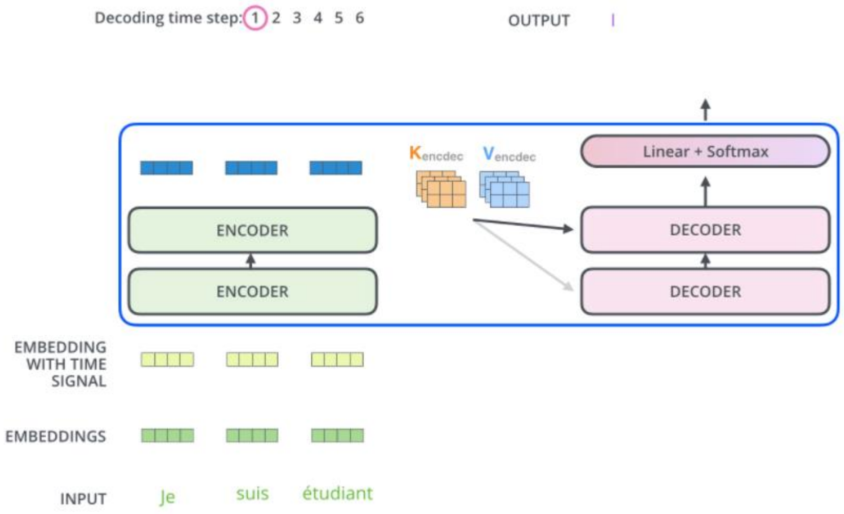


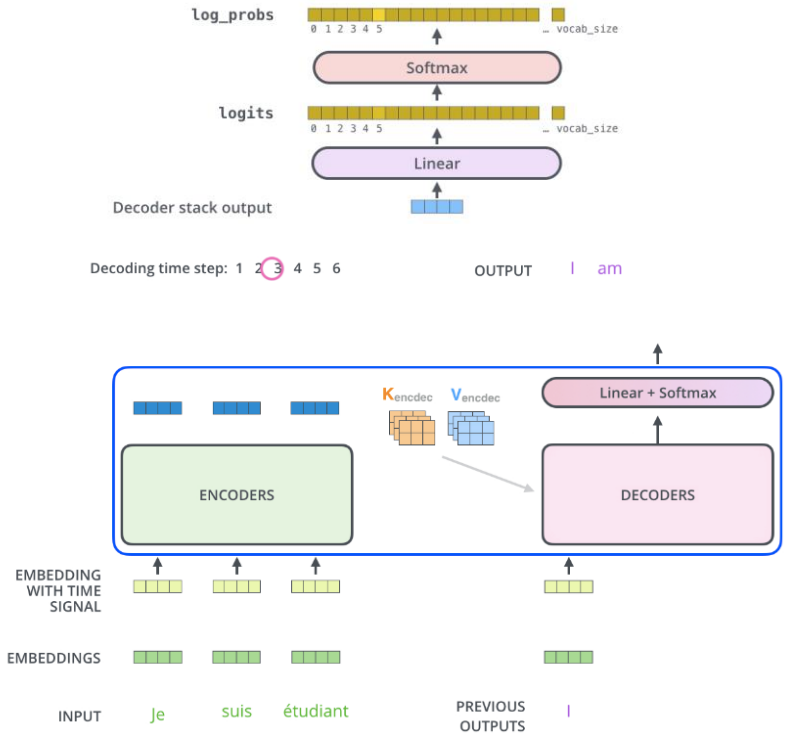


$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Methodology

Inference





Thank you

Muhammad Umar Salman
umar.salman@mbzaui.ac.ae