Master in Computer Vision Barcelona

[http://pagines.uab.cat/mcv/]

Xavier Giro-i-Nieto

@DocXavi

xavier.giro@upc.edu

Associate Professor

Universitat Politècnica de Catalunya

Module 6 - Day 8 - Lecture 1

The Transformer

31th March 2022

Video-lecture

Acknowledgments

Marta R. Costa-jussà

Associate Professor
Universitat Politècnica de Catalunya

Carlos Escolano

PhD Candidate
Universitat Politècnica de Catalunya

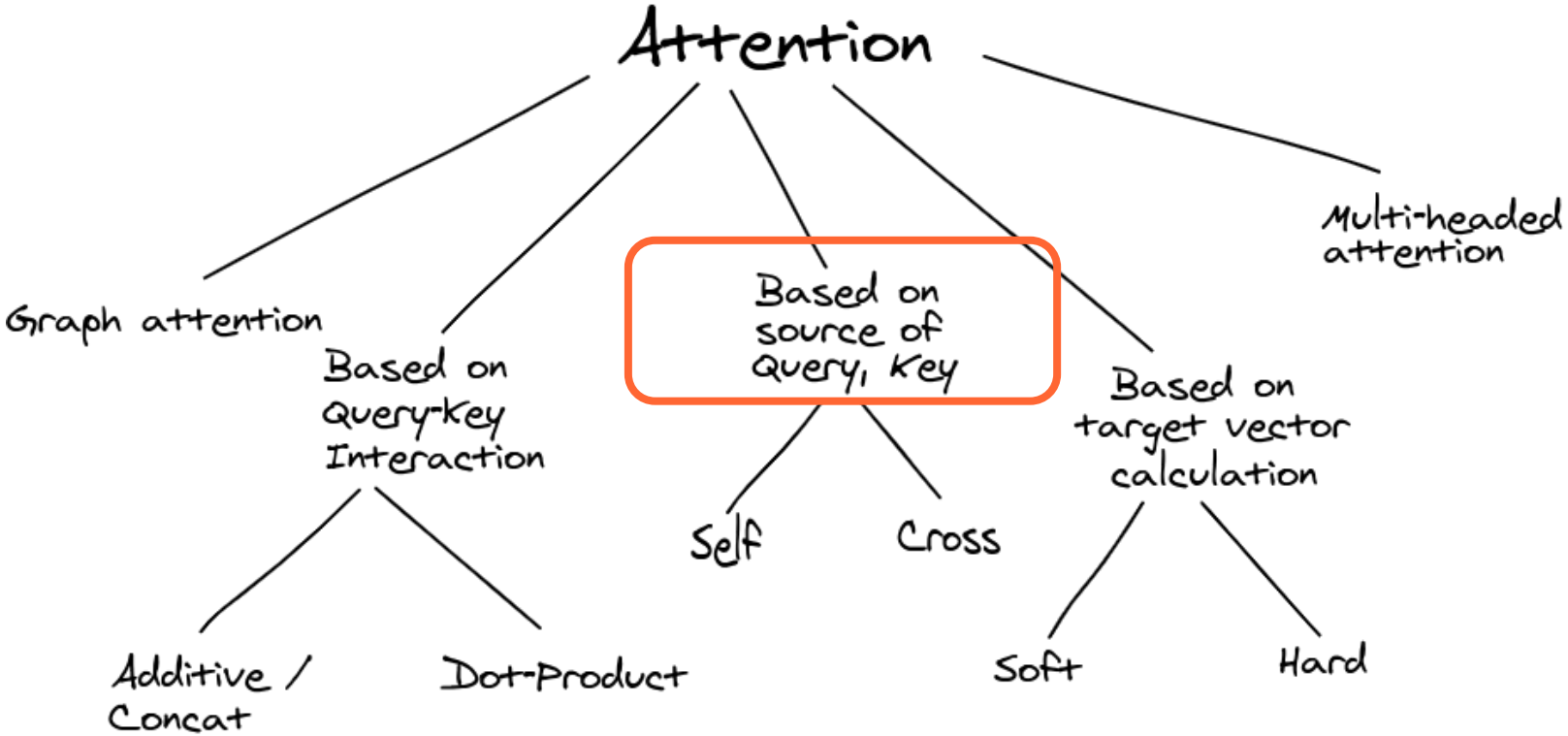
Gerard I. Gállego
PhD Student
Universitat Politècnica de Catalunya
gerard.ion.gallego@upc.edu
@geiongallego

UPC

Outline

1. **Reminders**

Reminder



Nikhil Sha, "Attention ? An other Perspective!". 2020.

Reminder

Attention is a mechanism to compute a context vector (c) for a **query (Q)** as a weighted sum of **values (V).**

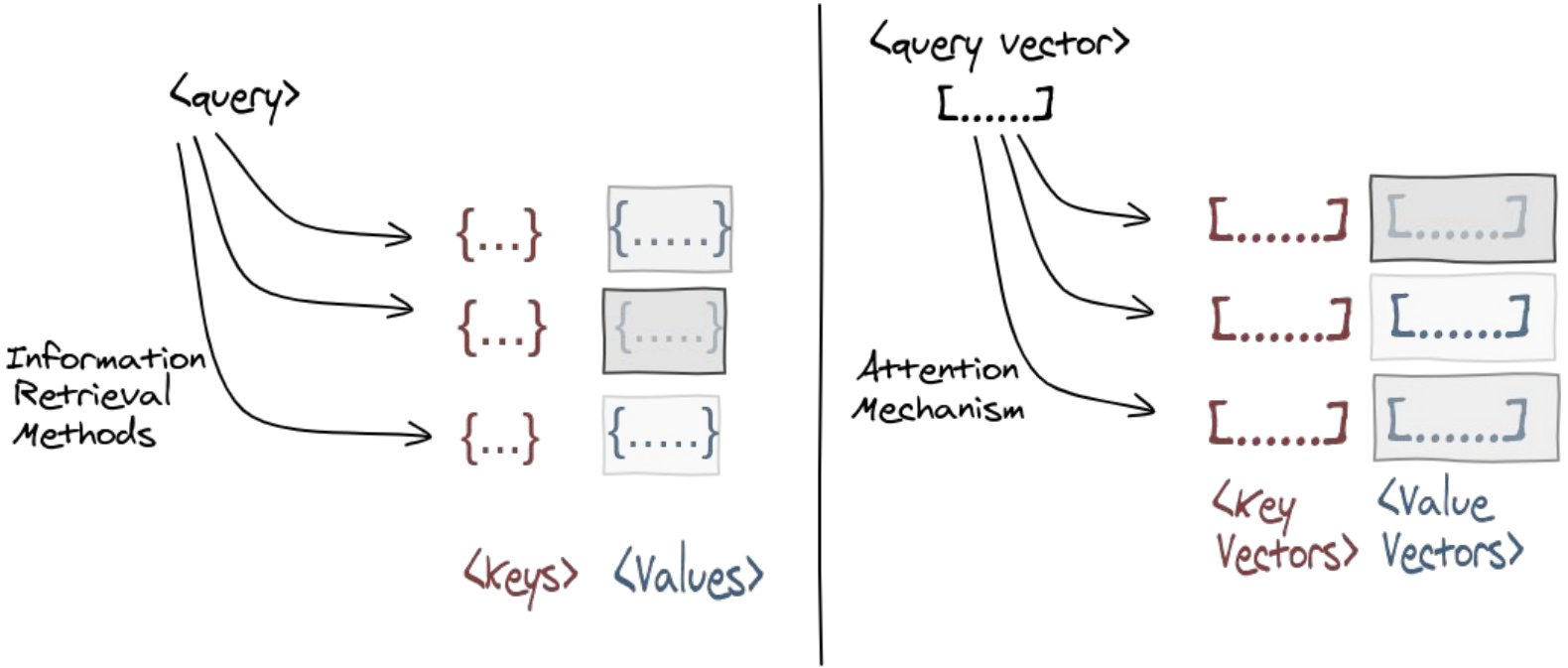
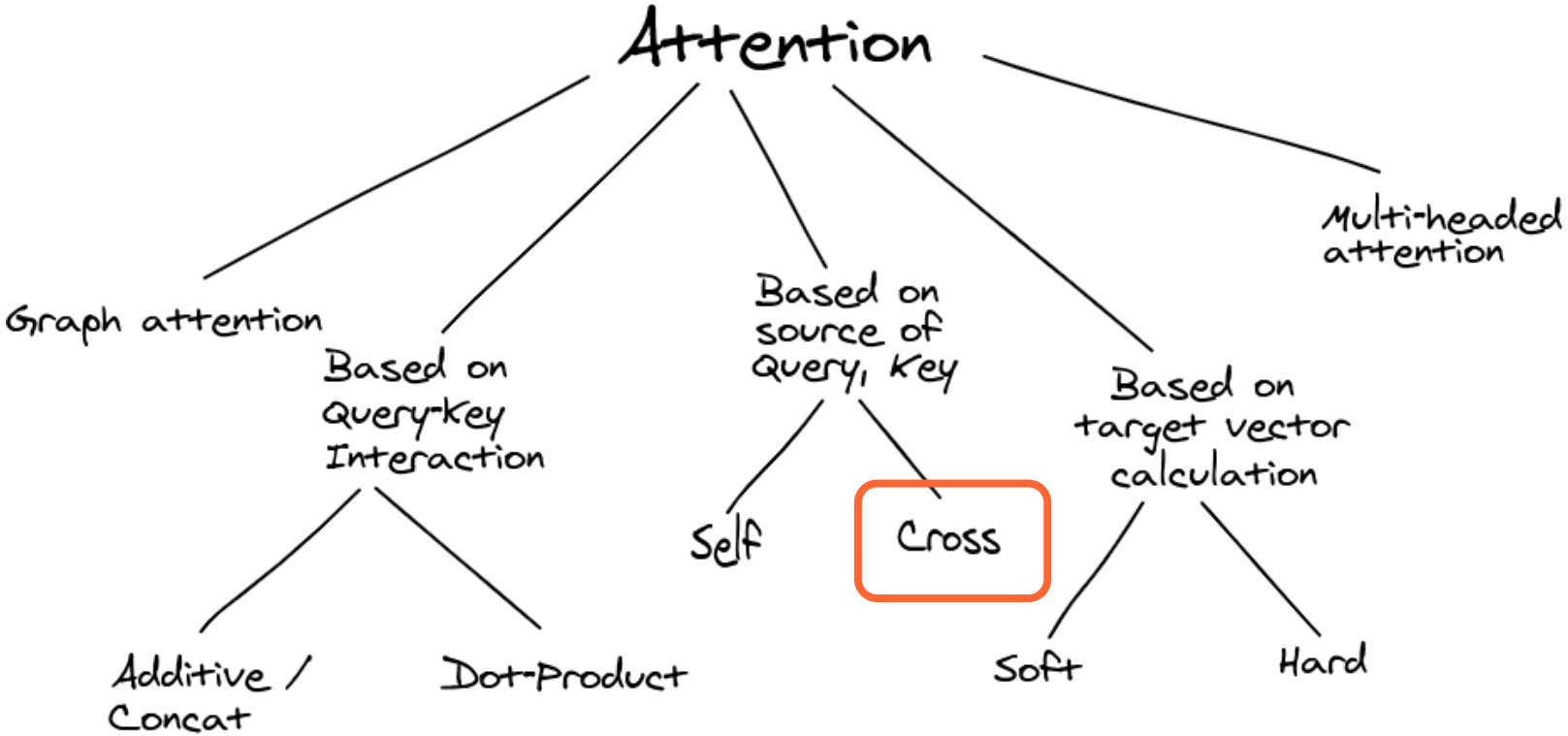


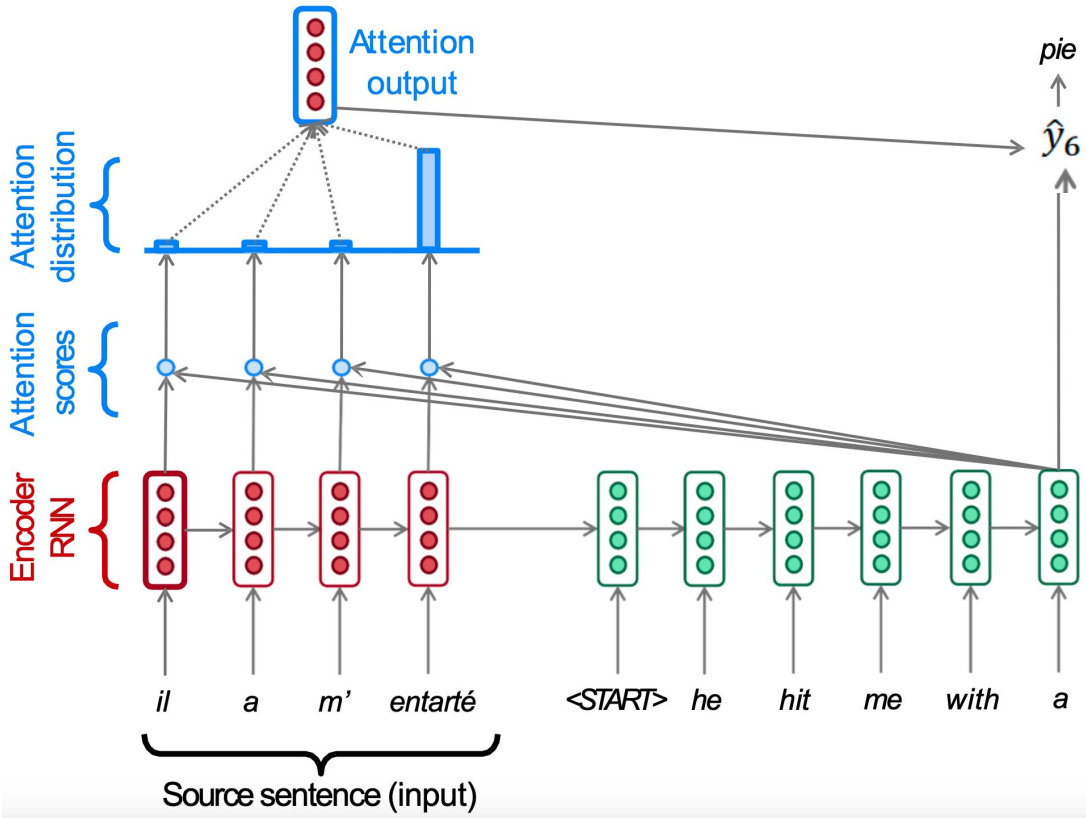
Figure: Nikhil Shah, "Attention? An Other Perspective! [Part 1]" (2020)

Reminder



Nikhil Sha, "Attention ? An other Perspective!". 2020.

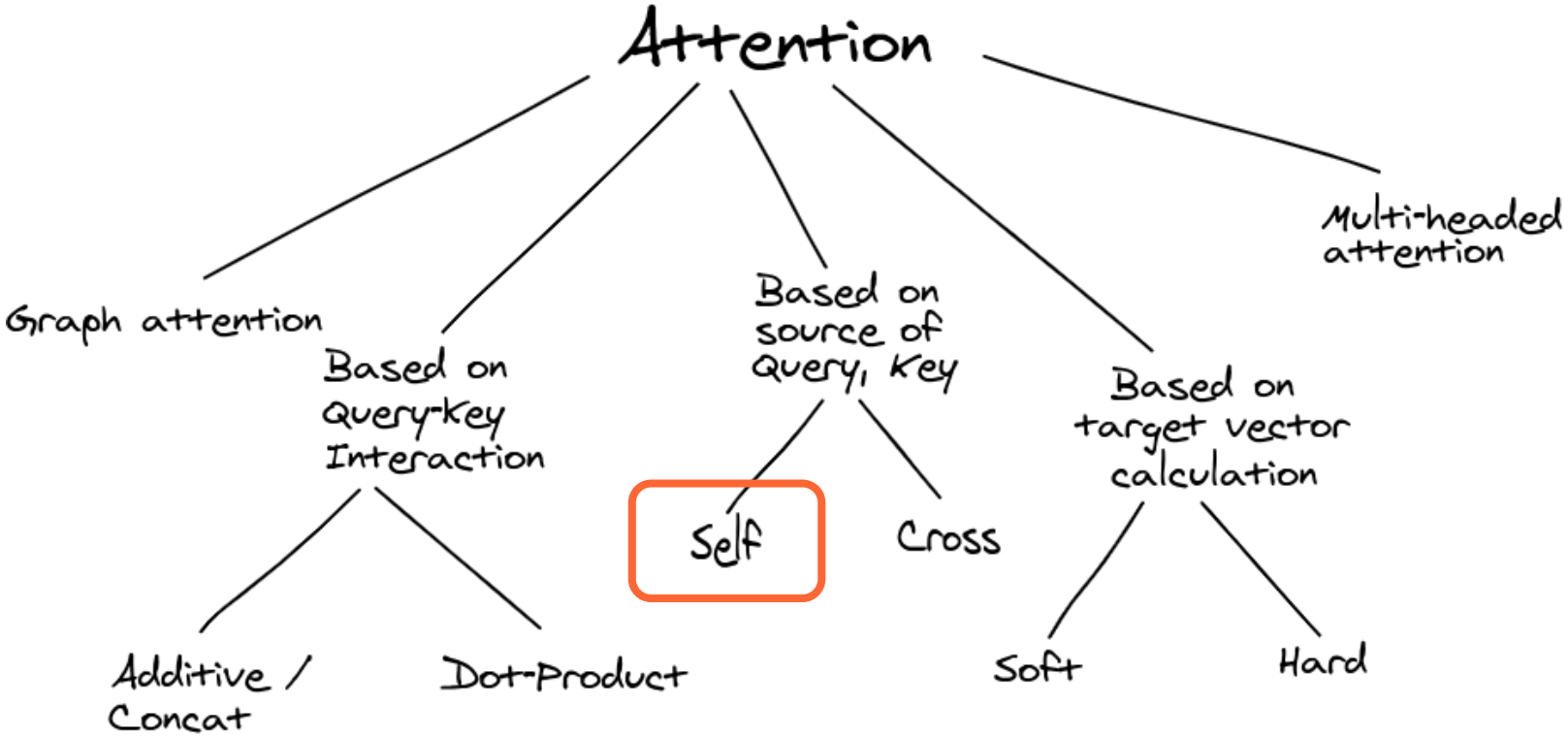
Reminder: Seq2Seq with Cross-Attention



In this case, **cross-attention** refers to the attention between the encoder and decoder states.

Slide concept: Abigail See, Matthew Lamm (Stanford CS224N), 2020

What may the term “self” refer to, as a contrast of “cross”-attention ?



Nikhil Sha, “Attention ? An other Perspective!”. 2020.

Outline

1. Motivation
2. **Self-Attention (SA)**

Self-Attention (or intra-Attention)

Self-attention refers to attending to other elements from the SAME sequence.

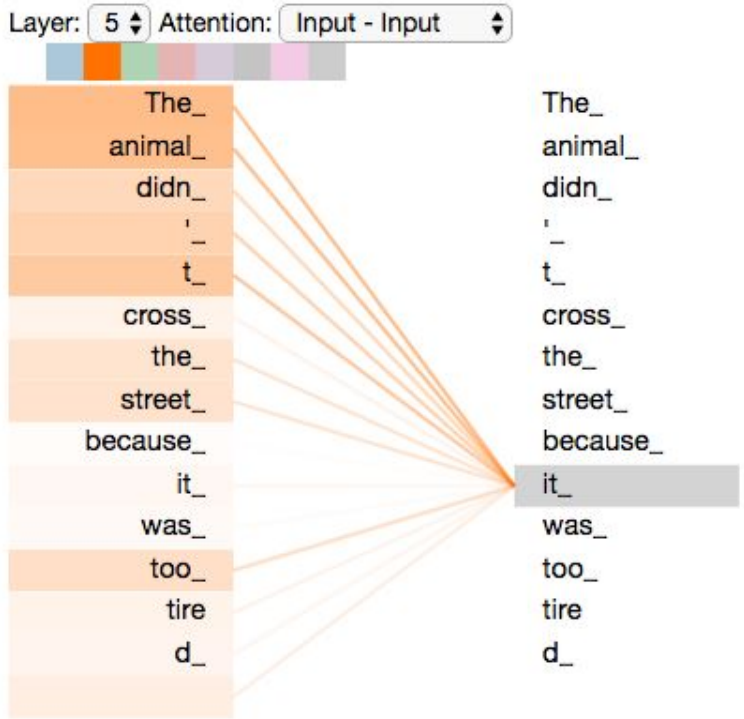
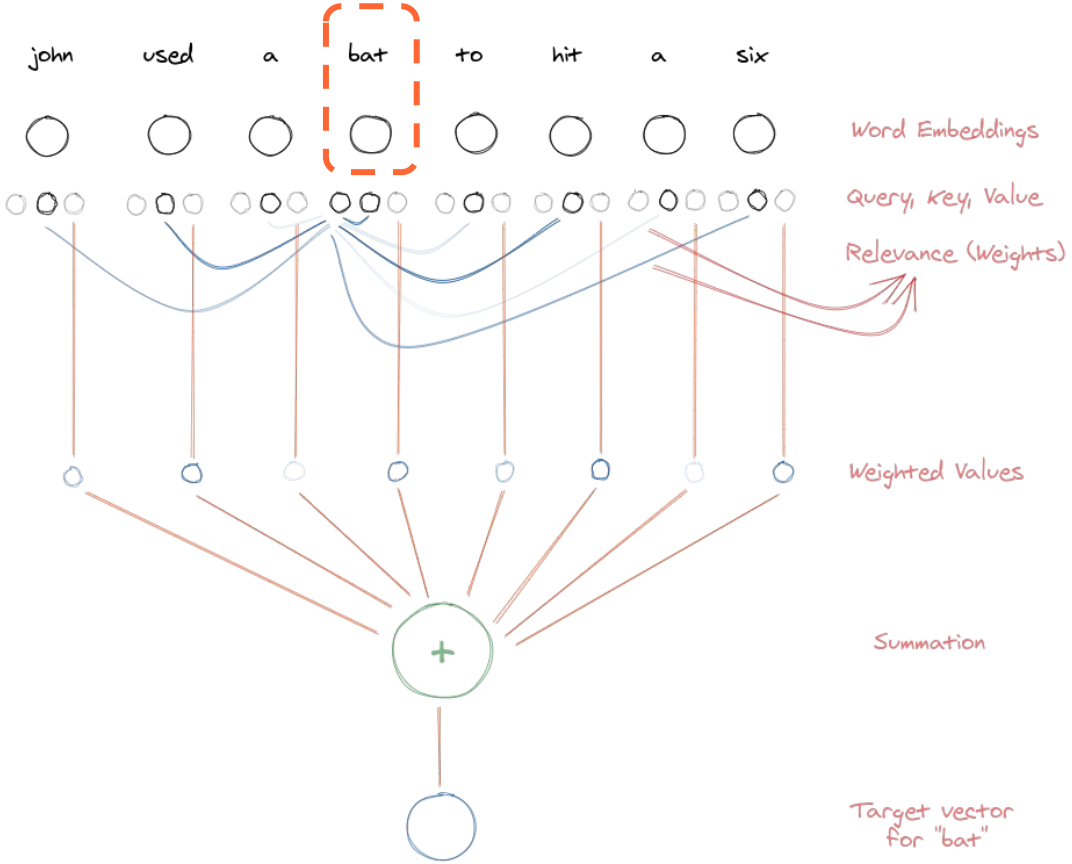


Figure:
Jay Alammar,
"The Illustrated Transformer"

Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. A structured self-attentive sentence embedding. ICLR 2017.

Self-Attention (or intra-Attention)



Query (Q)
$g(x) = W^Q x$

Key (K)
$f(x) = W^K x$

Value (V)
$h(x) = W^V x$

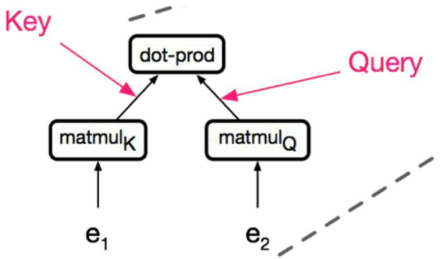
W^Q, W^K and W^V are **projection layers** shared across all words.

Nikhil Sha, "Attention ? An other Perspective!". 2020.

Self-Attention (or intra-Attention)

Which steps are necessary to compute the contextual representation of a word embedding e_2 in a sequences of four words embeddings (e_1, e_2, e_3, e_4) ?

A (scaled) dot-product is computed between each pair of word embeddings (eg. e_1 and e_2)...

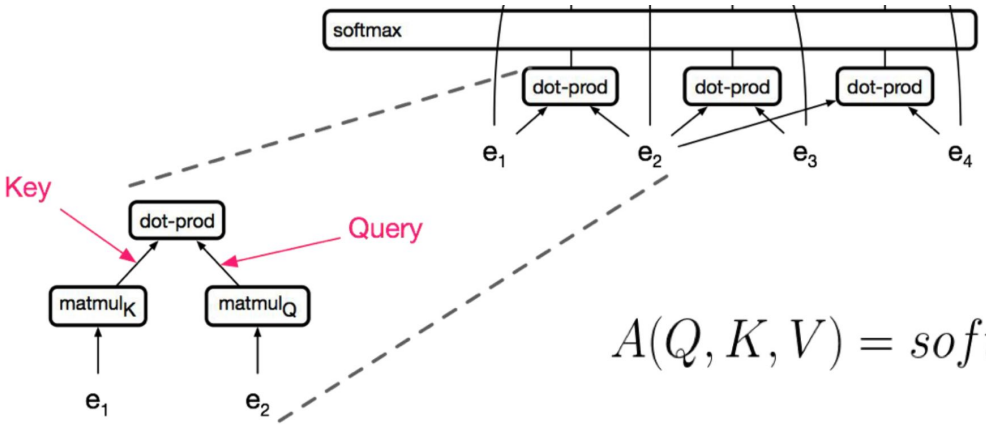


#Transformer Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.. Attention is all you need. NeurIPS 2017.

Self-Attention (or intra-Attention)

Which steps are necessary to compute the contextual representation of a word embedding e_2 in a sequences of four words embeddings (e_1, e_2, e_3, e_4) ?

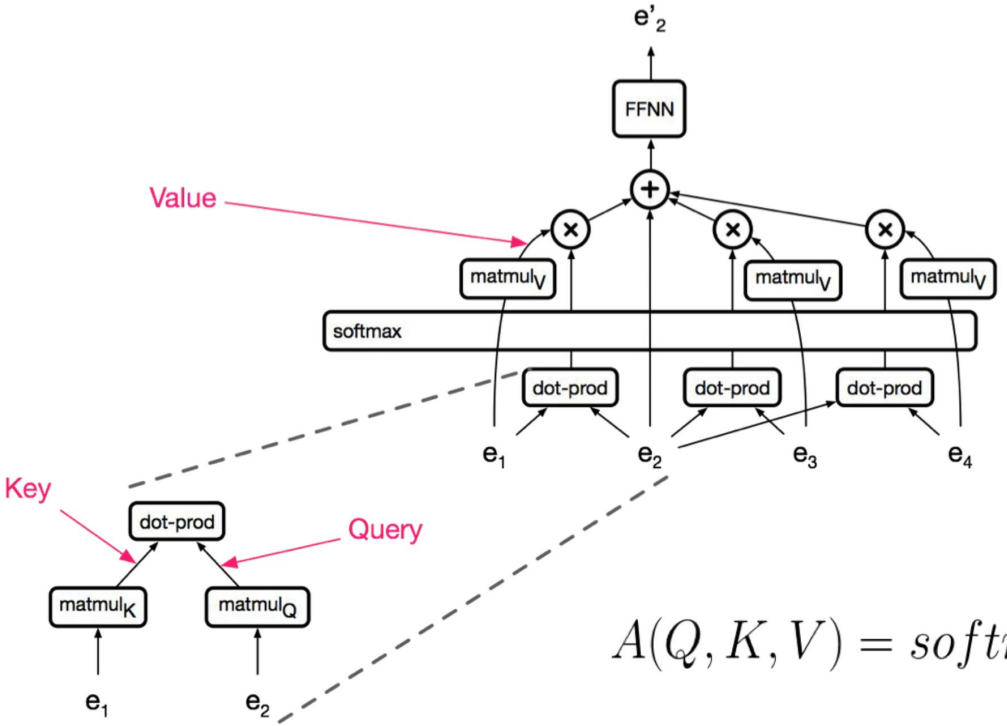
… a softmax layer normalizes the attention scores to obtain the attention distribution...



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

Self-Attention (or intra-Attention)

Which steps are necessary to compute the contextual representation of a word embedding e_2 in a sequences of four words embeddings (e_1, e_2, e_3, e_4) ?



...the same word embeddings are combined to obtain the contextual representation e_2'.

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

Self-Attention (or intra-Attention)

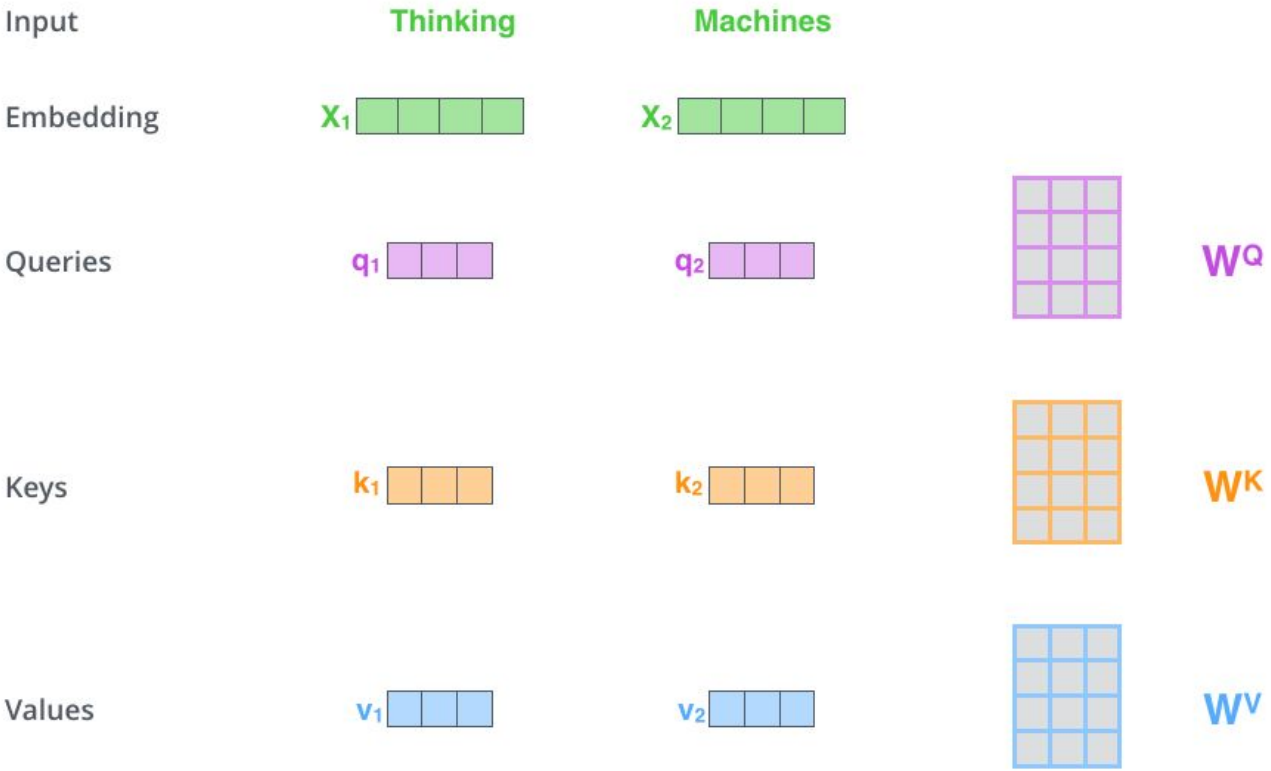


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Self-Attention (or intra-Attention)

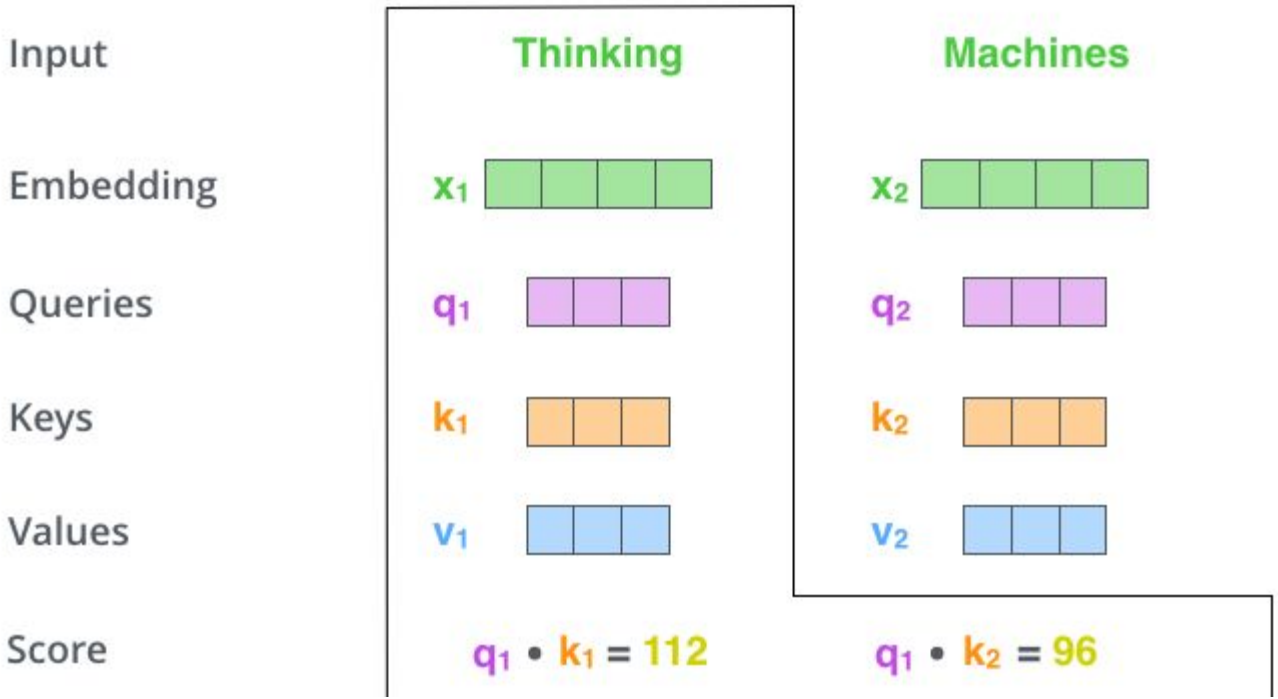


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Self-Attention (or intra-Attention)

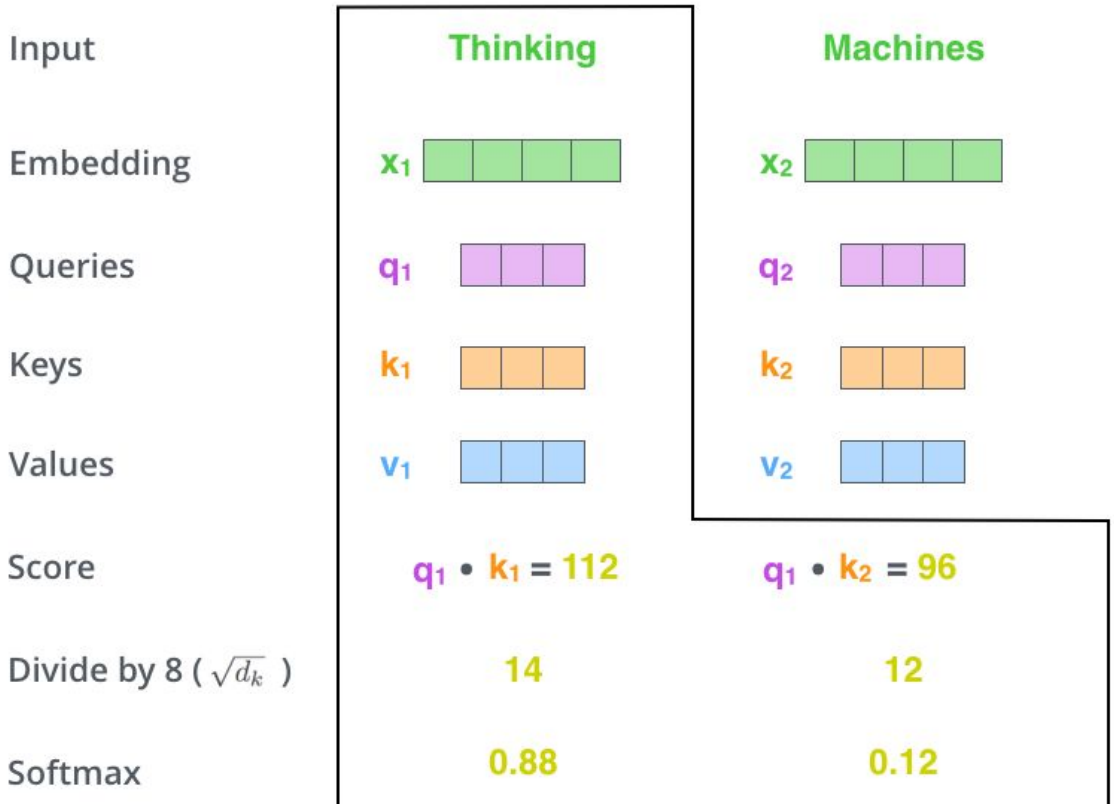


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Self-Attention (or intra-Attention)

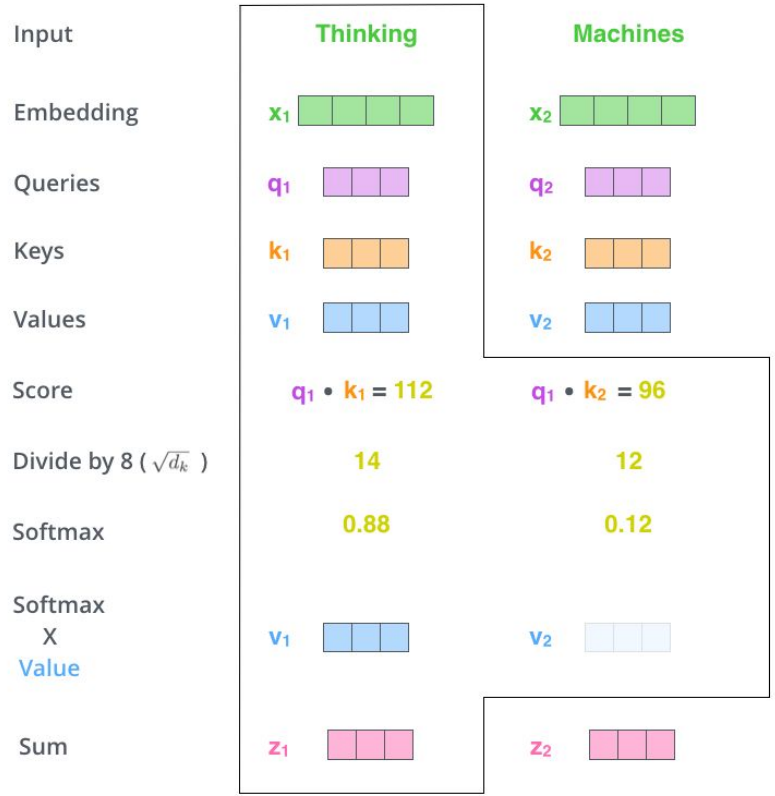


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Self-Attention (or intra-Attention)

1) This is our input sentence

2) We embed each word

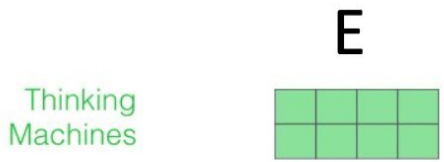


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Self-Attention (or intra-Attention)

1) This is our input sentence

2) We embed each word

3) We multiply X with weight matrices

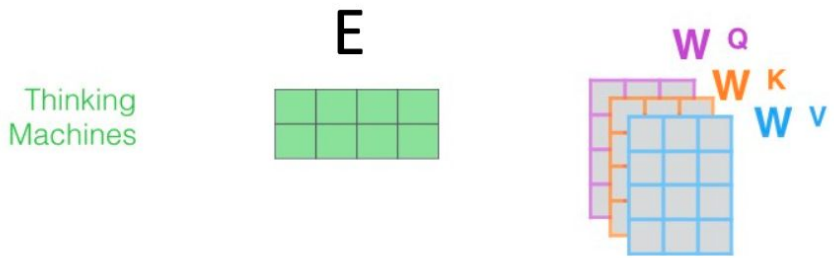


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Self-Attention (or intra-Attention)

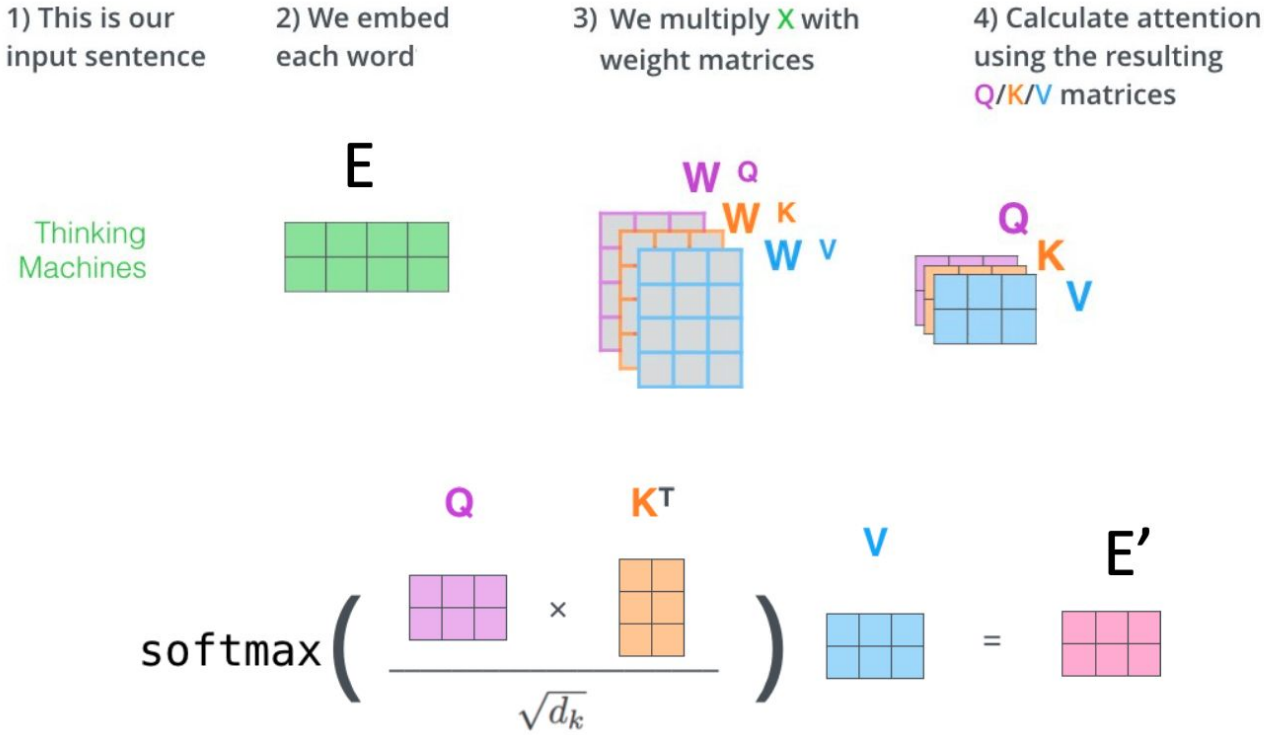


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Self-Attention (or intra-Attention)

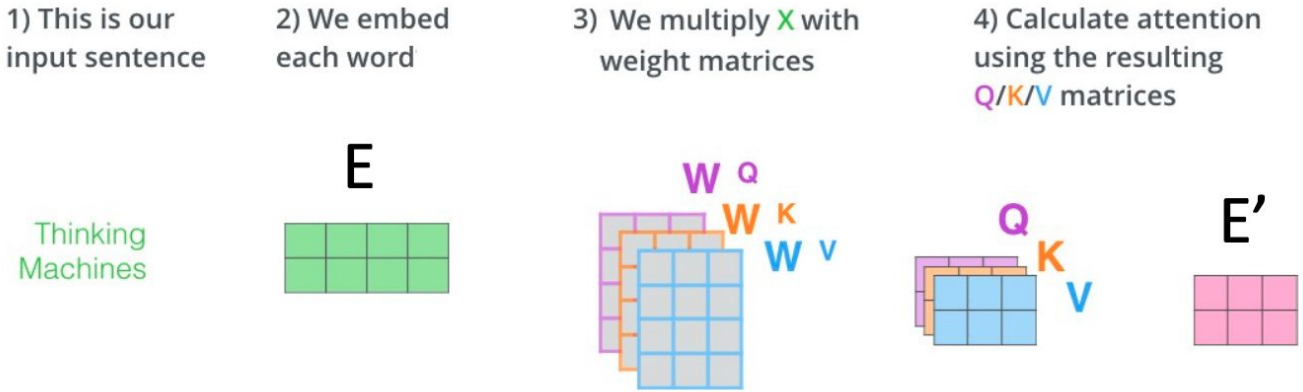


Figure: Jay Alammar, “The illustrated Transformer” (2018)

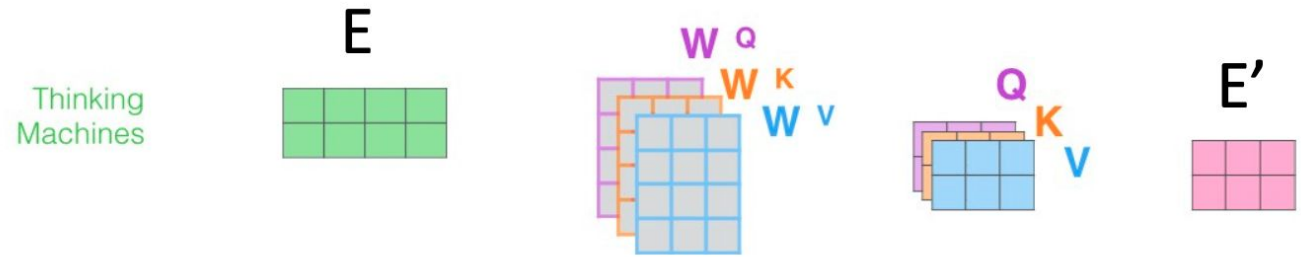
Self-Attention (or intra-Attention)

1) This is our input sentence

2) We embed each word

3) We multiply X with weight matrices

4) Calculate attention using the resulting Q/K/V matrices



Scaled dot-product attention

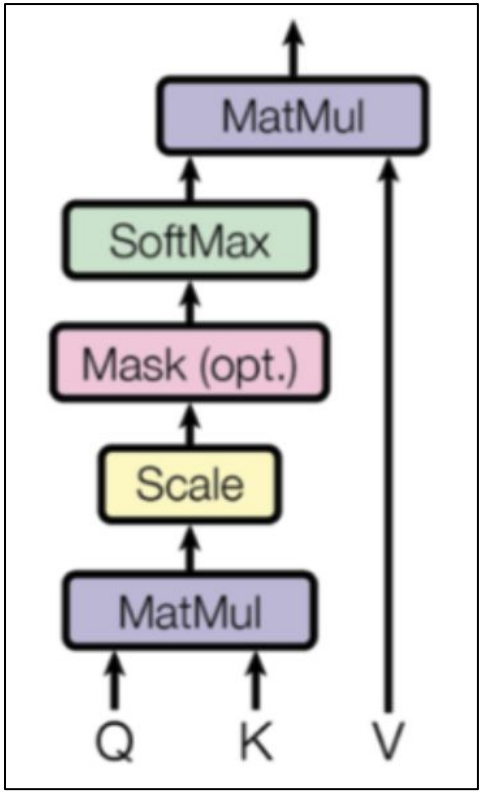
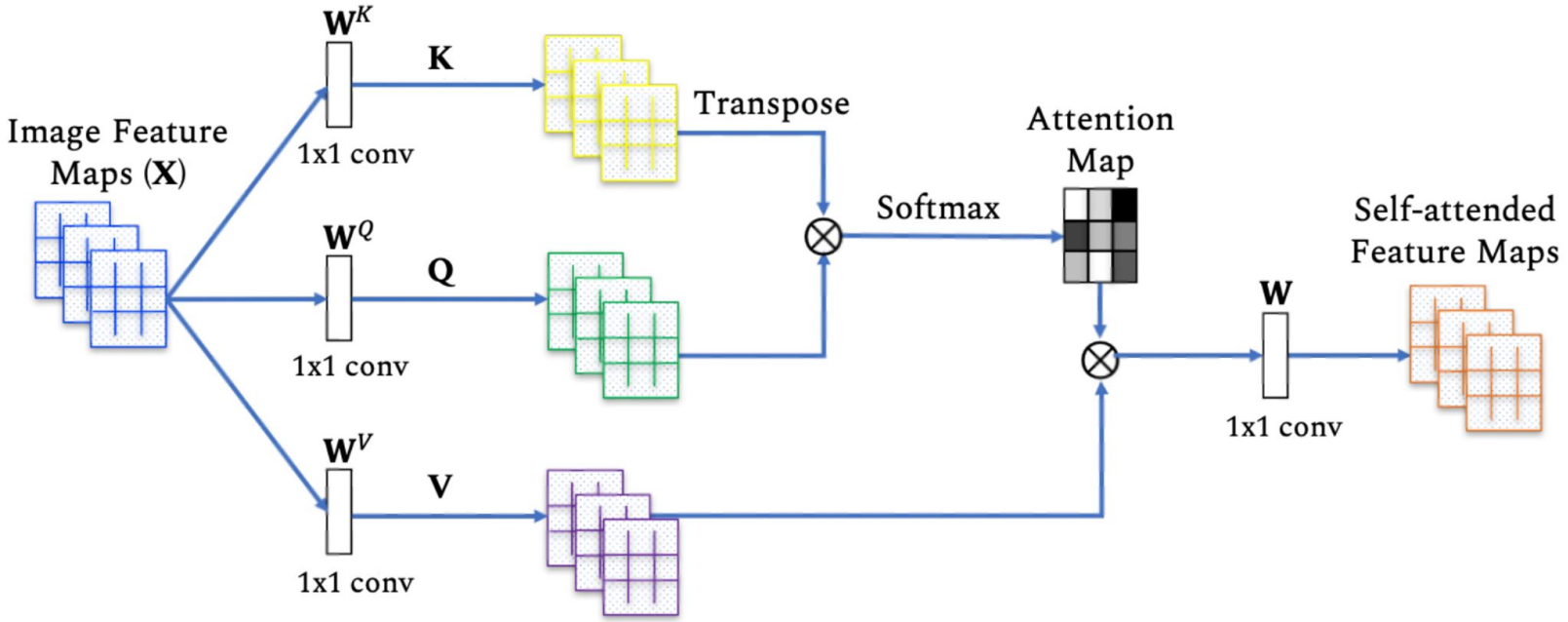


Figure: Jay Alammar, “The illustrated Transformer” (2018)

Self-Attention over 2D Image Feature Maps



#SAGAN Zhang, Han, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. "Self-attention generative adversarial networks." ICML 2019. [video]

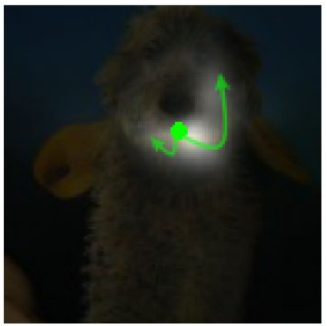
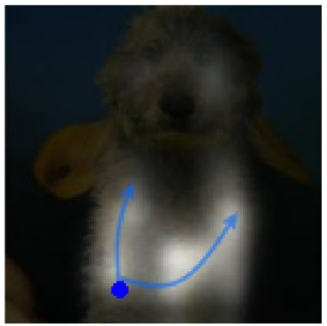
Self-Attention in for image generation

Generator (G): Details can be generated using cues from all feature locations.
Discriminator (D): Can check consistency betweenn features in distant portions of the image.

Query locations

Attention maps for different query locations




#SAGAN Zhang, Han, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. "Self-attention generative adversarial networks." ICML 2019. [video]

Outline

1. Motivation
2. Self-attention (SA)
3. **Multi-head Self-Attention (MHSA)**

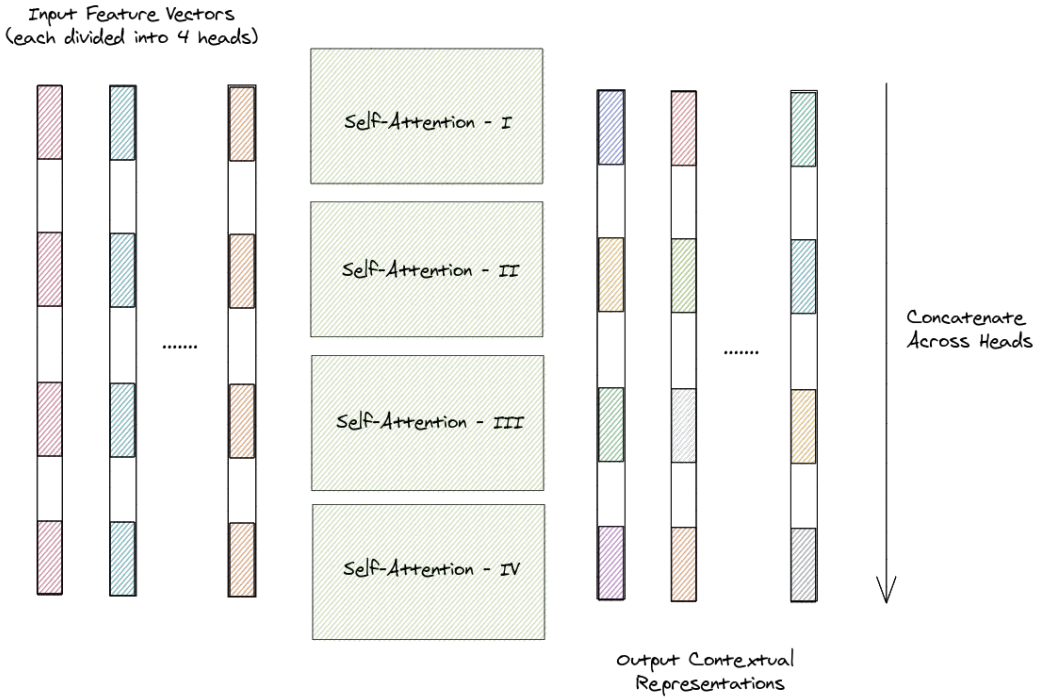
Multi-Head Self-Attention (MHSA)

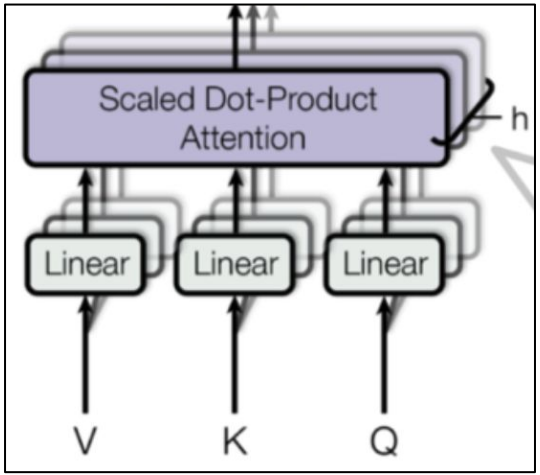
In vanilla self-attention, a single set of projection matrices W^Q, W^K, W^V is used.

Nikhil Sha, "Attention ? An other Perspective!". 2020.

Multi-Head Self-Attention (MHSA)

In multi-head self-attention, multiple sets of projection matrices are used, and can provide different contextual representations for the same input token.





Nikhil Sha, "Attention ? An other Perspective!". 2020.

Multi-Head Self-Attention (MHSA)

Example with 2 heads:

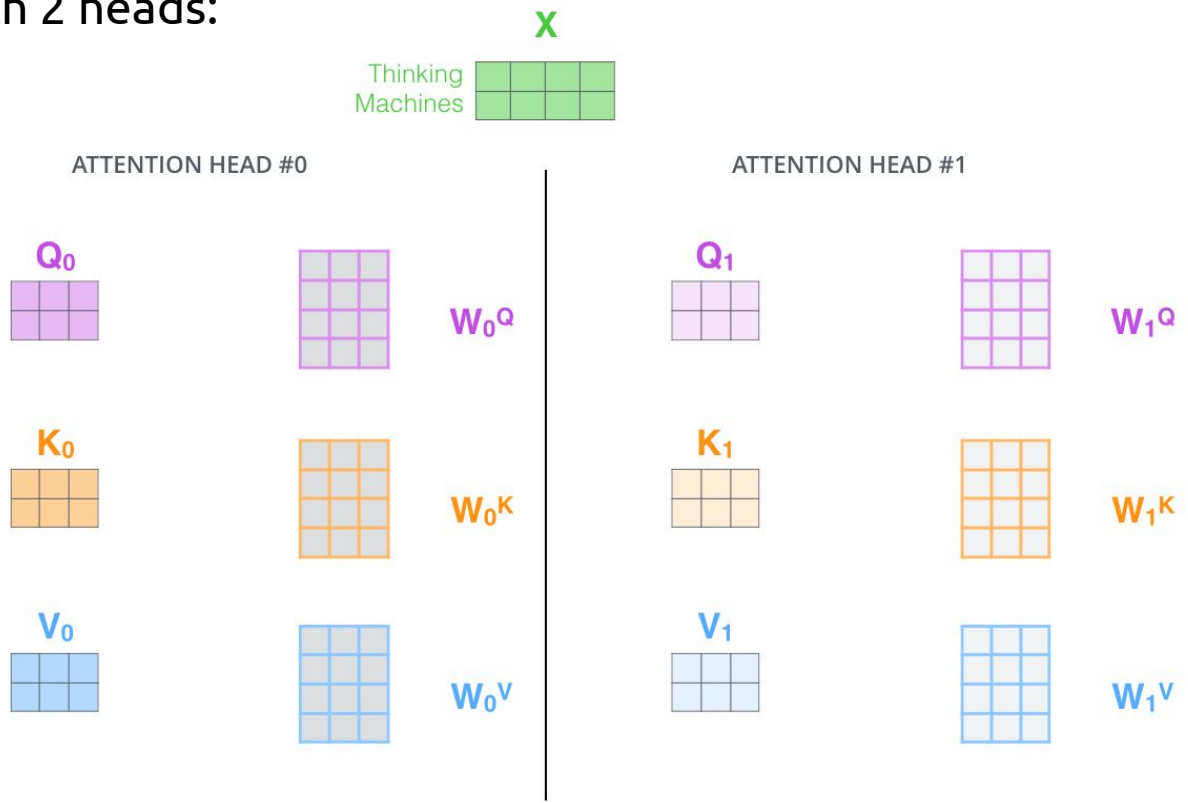


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Multi-Head Self-Attention (MHSA)

Example with 7 heads:

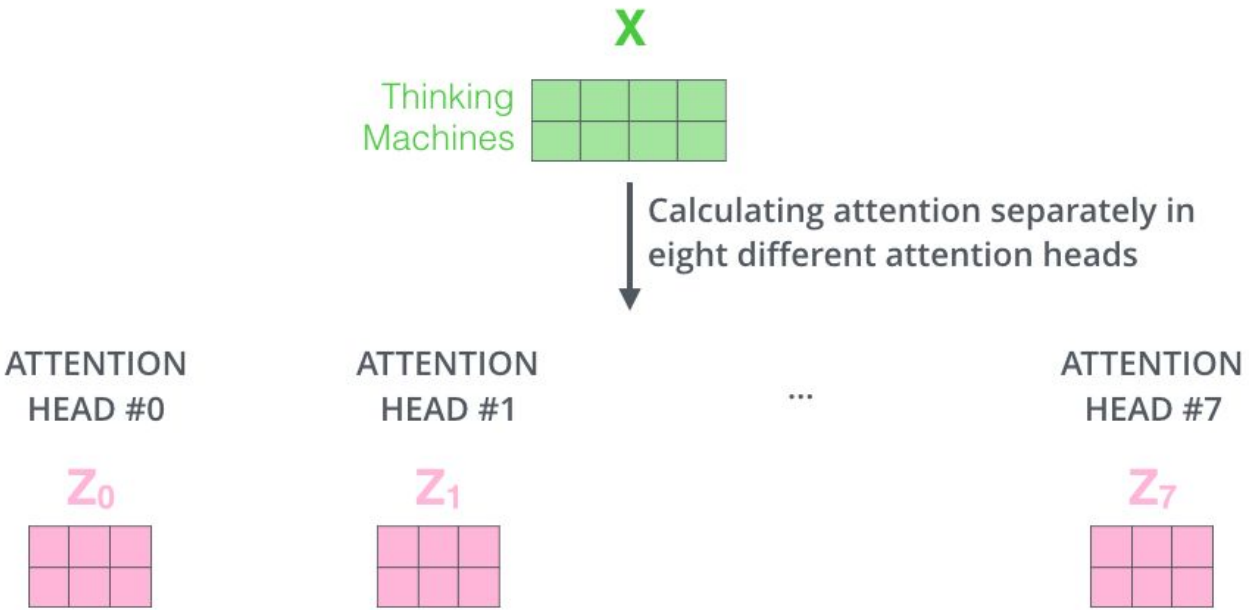
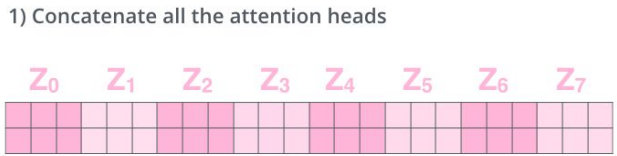


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Multi-Head Self-Attention (MHSA)

The multi-head self-attended z_i are concatenated…



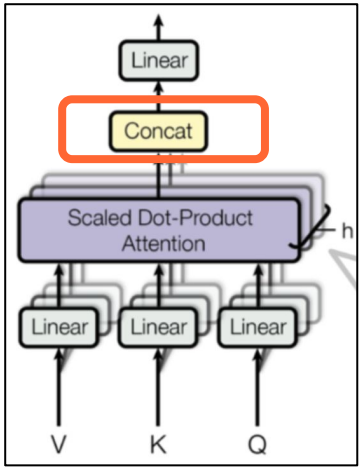
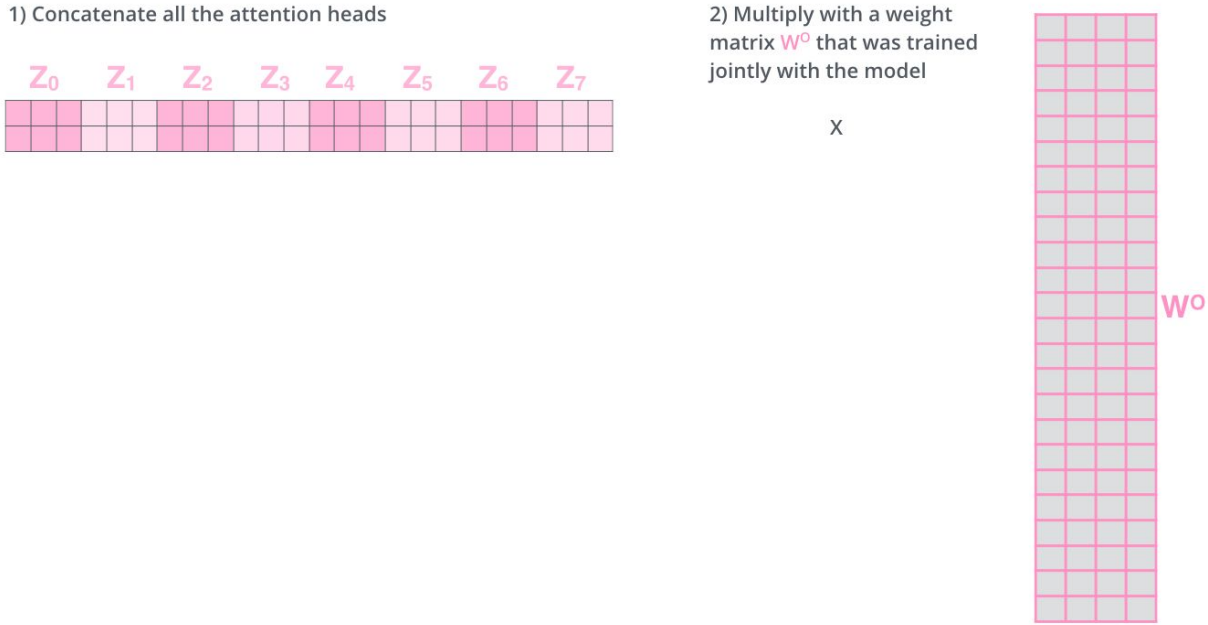


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Multi-Head Self-Attention (MHSA)

…and projected with a linear layer…



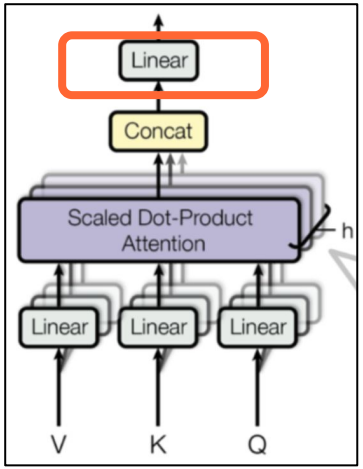
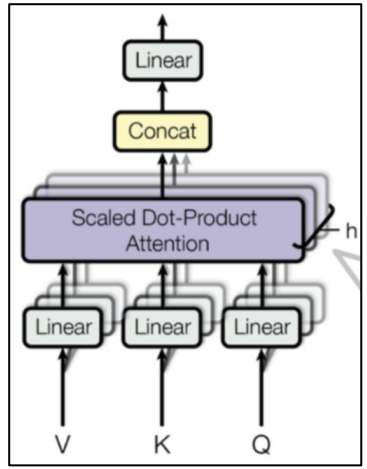


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Multi-Head Self-Attention (MHSA)

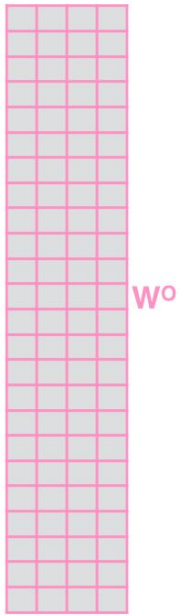
…to build a new single representation *z*.



1) Concatenate all the attention heads

2) Multiply with a weight matrix W^O that was trained jointly with the model

X



3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN

=

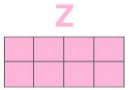
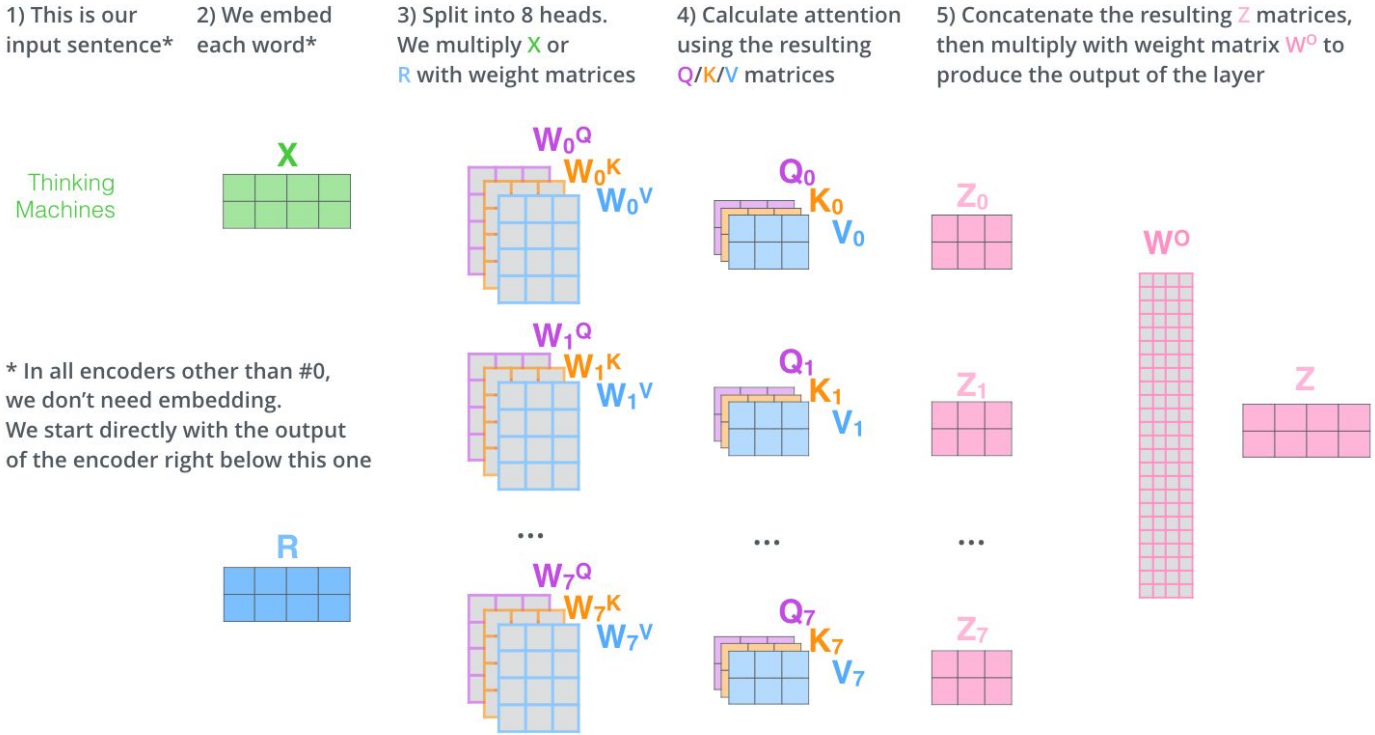


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Multi-Head Self-Attention (MHSA)

The overall MHSA scheme:



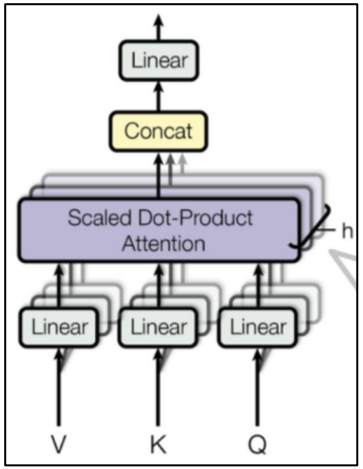
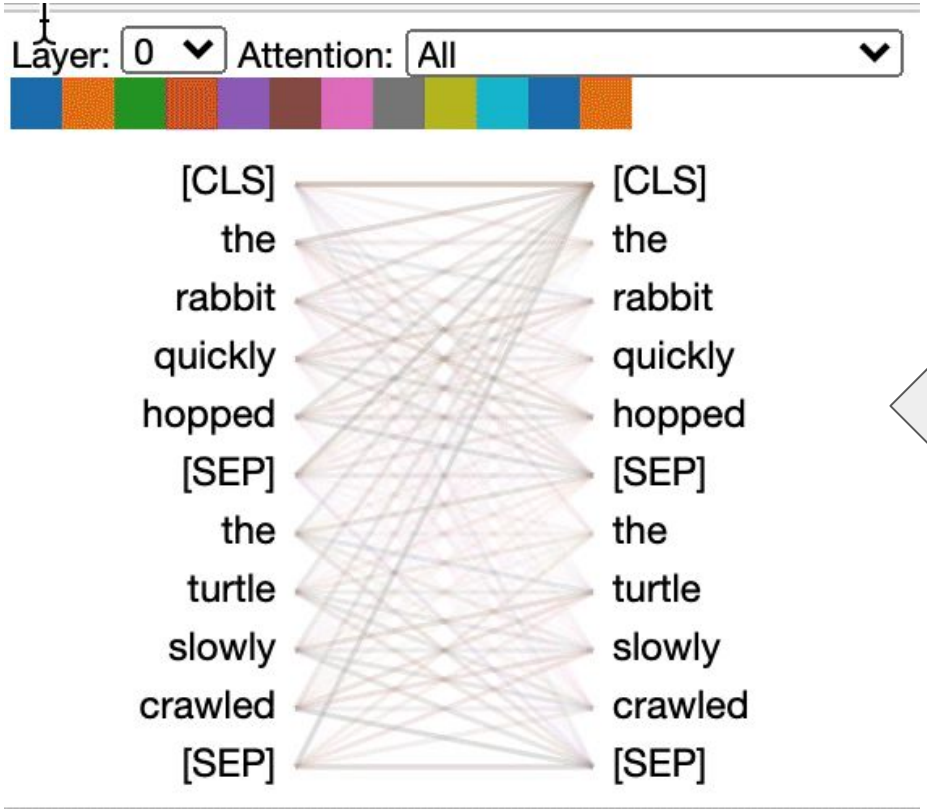


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Multi-head Self-Attention: Visualization

Each colour corresponds to a head.



Blue: First head only.

Multi-color: Multiple heads.

#BertViz Vig, Jesse. "A multiscale visualization of attention in the transformer model." ACL 2019. [code] [tweet]

Self-Attention and Convolutional Layers

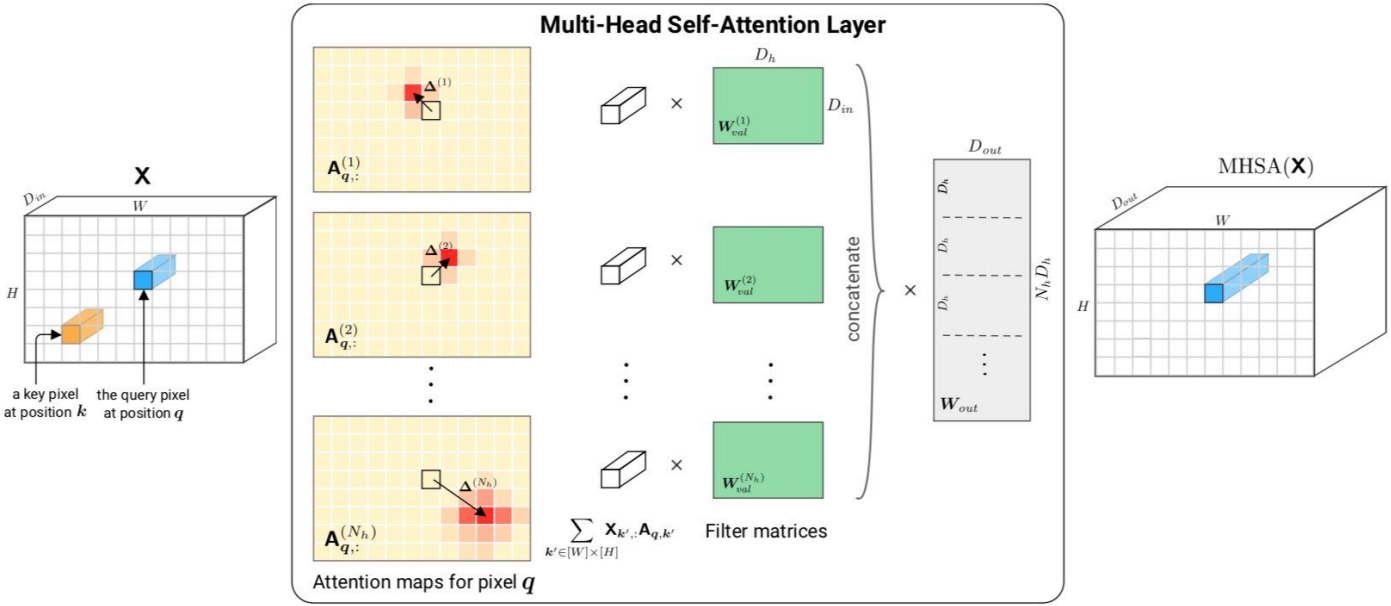


Figure 1: Illustration of a Multi-Head Self-Attention layer applied to a tensor image $\mathbf{X}$. Each head h attends pixel values around shift $\boldsymbol{\Delta}^{(h)}$ and learn a filter matrix $\boldsymbol{W}_{val}^{(h)}$. We show attention maps computed for a query pixel at position $\boldsymbol{q}$.

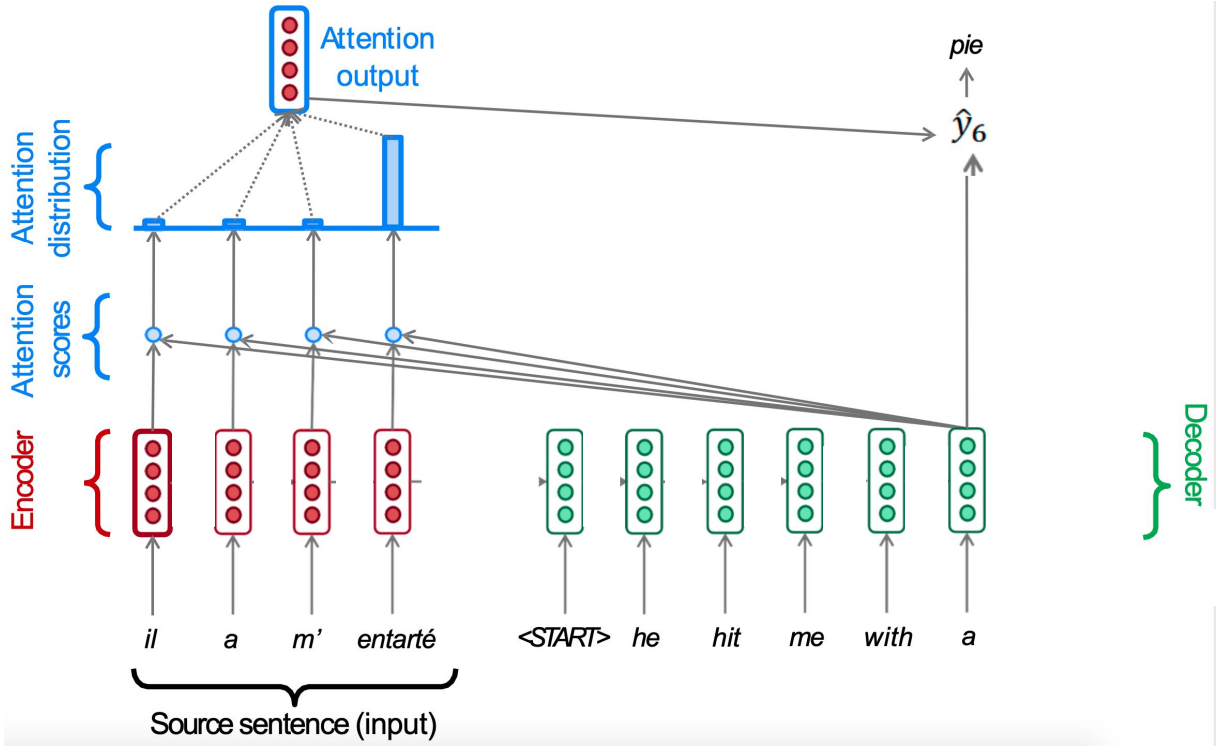
Cordonnier, J. B., Loukas, A., & Jaggi, M. On the relationship between self-attention and convolutional layers. ICLR 2020. [tweet] [code]

Outline

1. Motivation
2. Self-attention
3. Multi-head Attention
4. **Positional Encoding**

Positional Encoding

Given that the attention mechanism allows accessing all input (and output) tokens, we no longer need a memory through recurrent layers.



Positional Encoding

Where is the relative relation in the sequence encoded ?

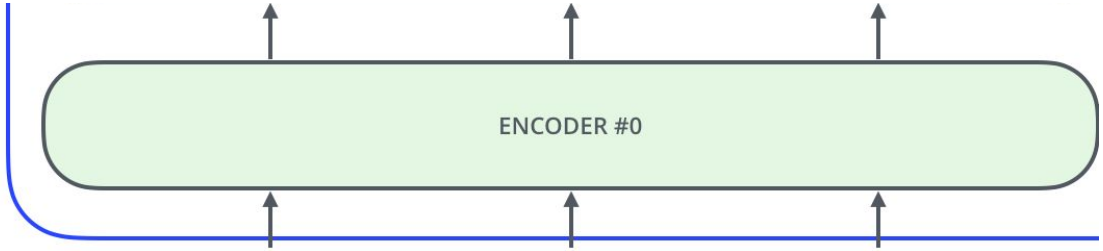


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Positional Encoding

Where is the relative relation in the sequence encoded ?

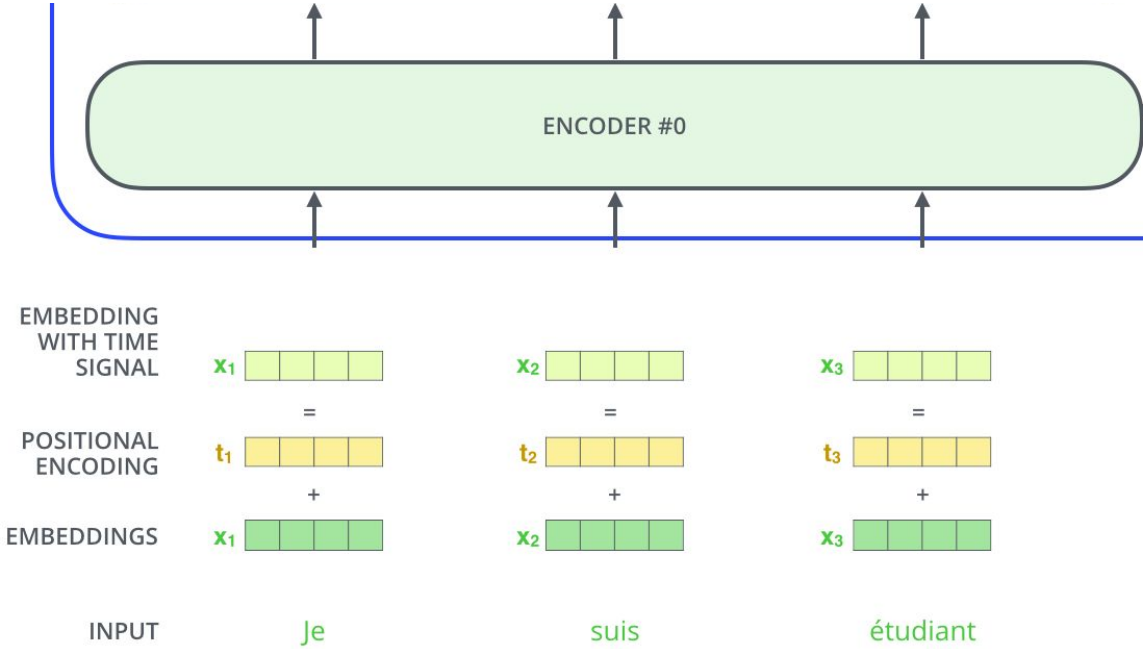
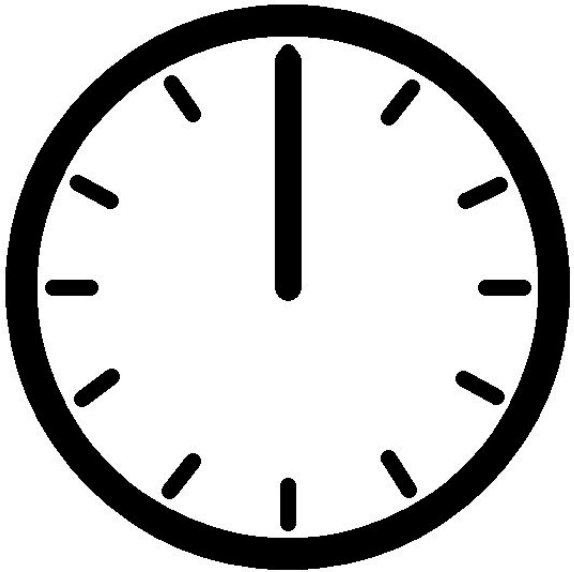
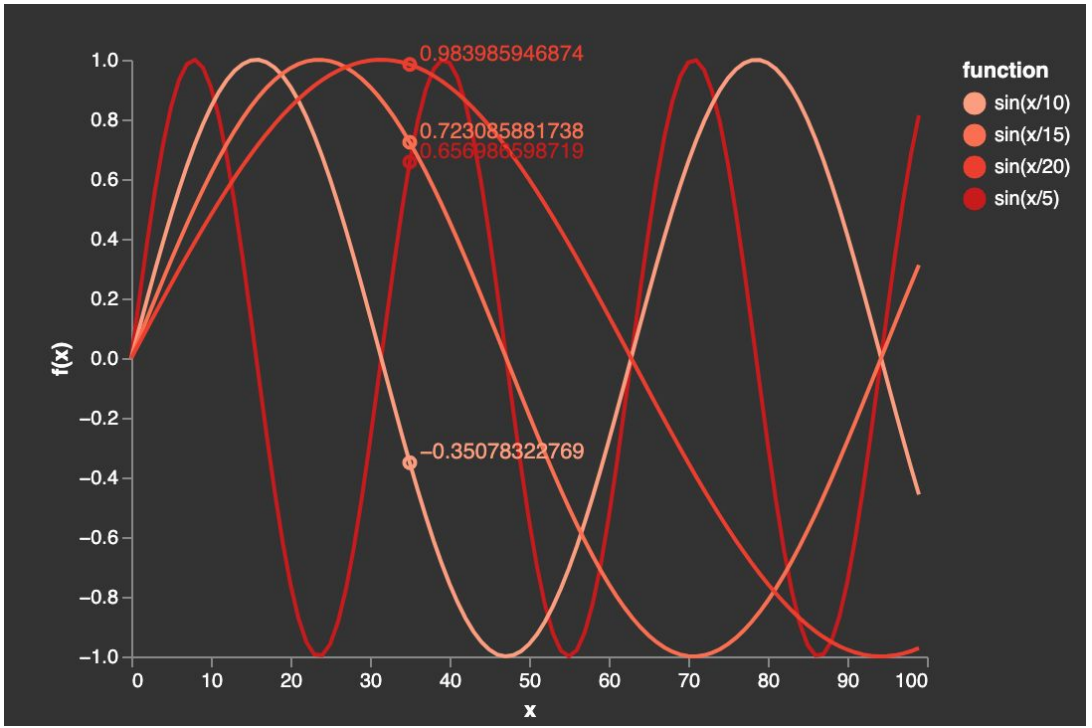


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Positional Encoding

Sinusoidal functions are typically used to provide positional encodings.





Maria Ribalta, Pere-Pau Vàzquez, “Visualization is all you need”. UPC GCED 2020.

Positional Encoding

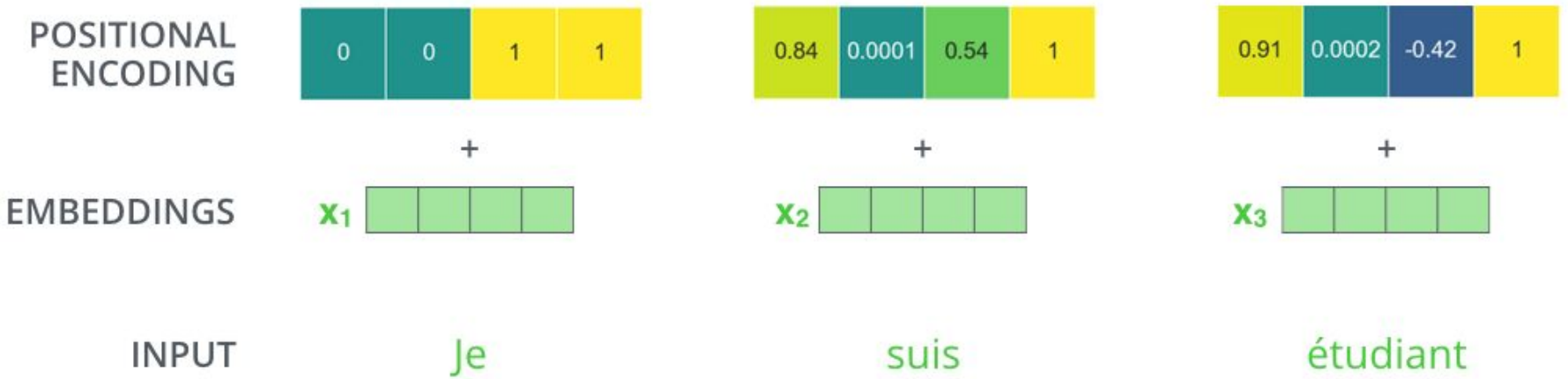


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Outline

1. Motivation
2. Self-attention
3. Multi-head Attention
4. Positional Encoding
5. **Stacks of Encoder & Decoder Layers**

Non-linear & Residual Layers

- The output of the self-attention is fed into a non-linear Feed-Forward layer.
- Both SA & FFD layers include residual connections.
- Data is also normalized with layer-normalization.

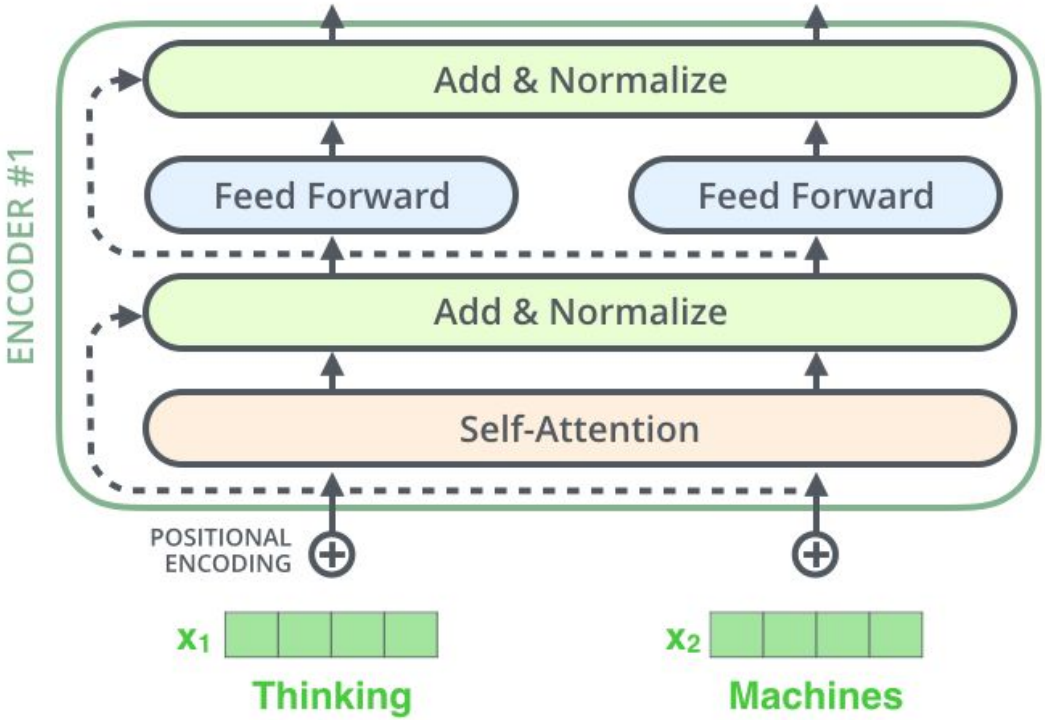


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Stacking Layers

Multiple stacks of SA + FF layers can be stacked in both encoder and decoder.
Example: 2 Stacked Encoders and Decoders:

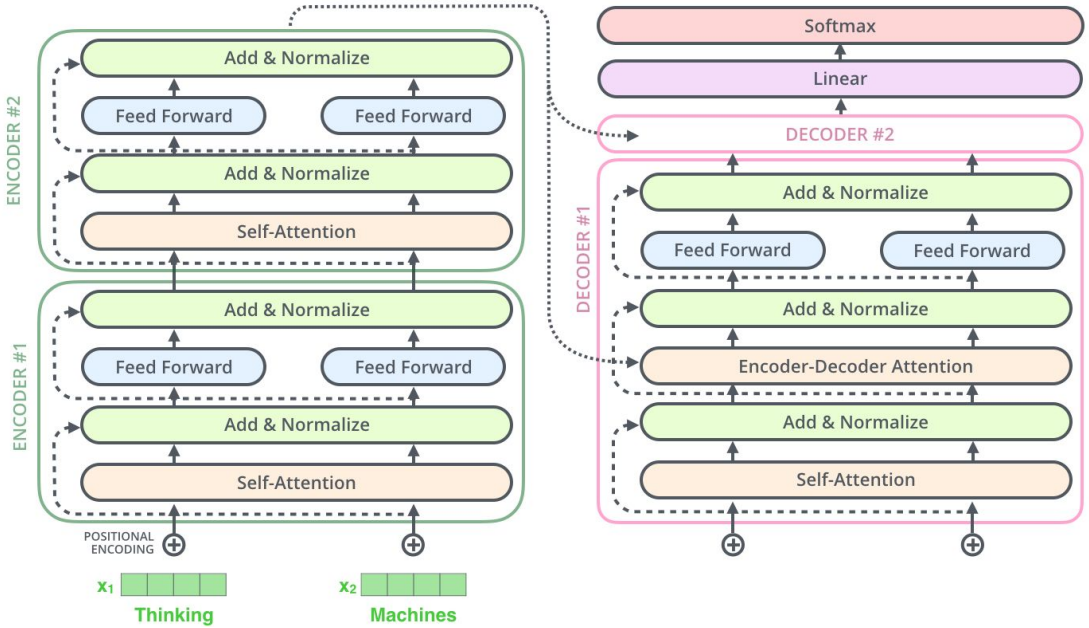


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Stacks of Encoder & Decoder Layers

Multiple stacks of SA + FF layers can be stacked in both encoder and decoder.
Example: 2 Stacked Encoders and Decoders:

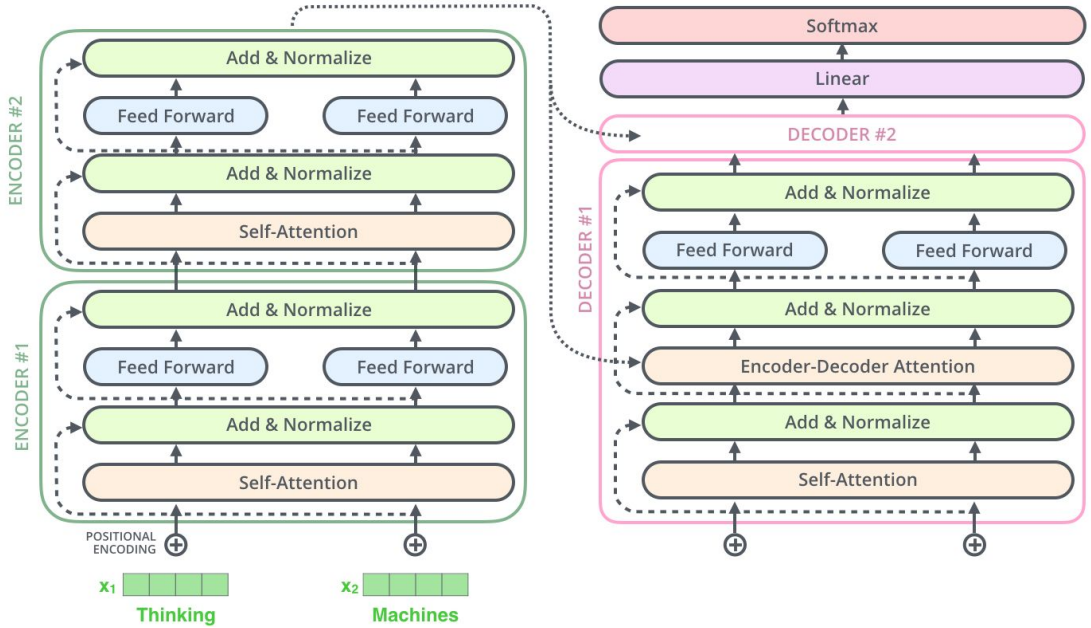


Figure: Jay Alammar, "The illustrated Transformer" (2018)

Stacks of Encoder & Decoder Layers

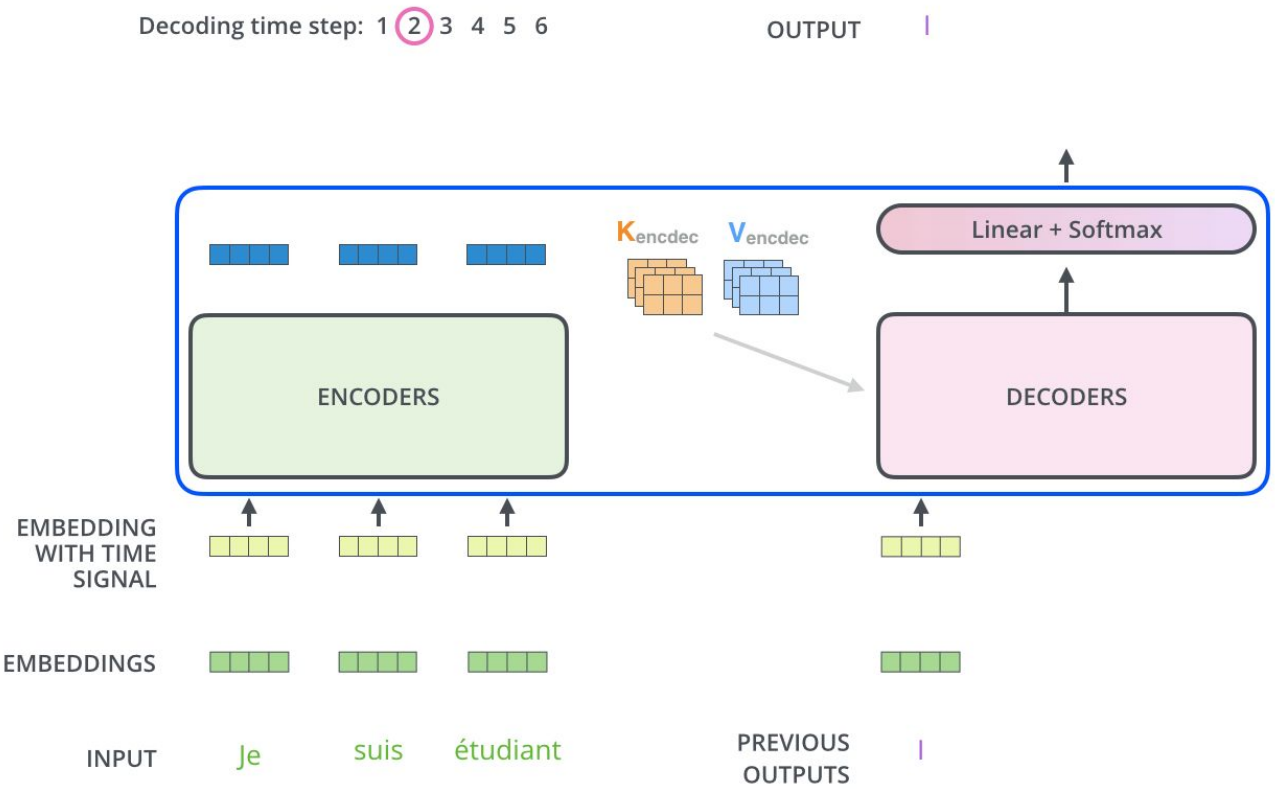


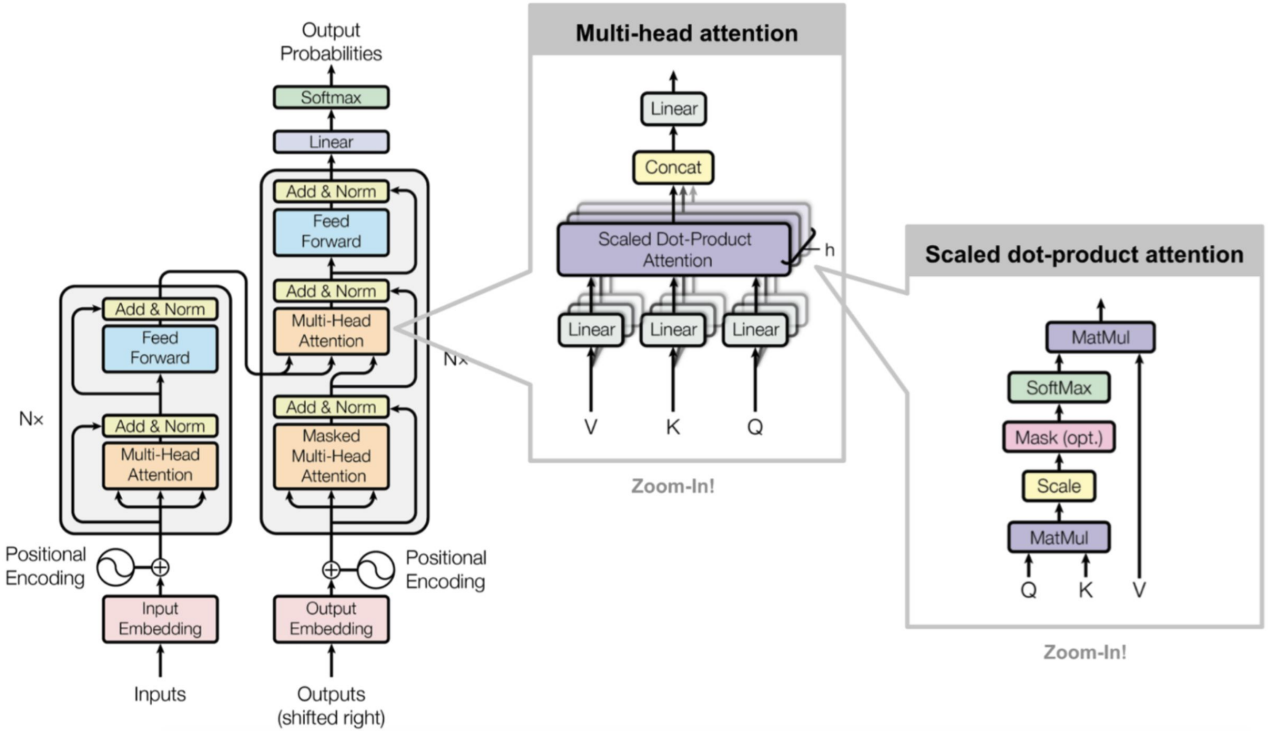
Figure: Jay Alammar, "The illustrated Transformer" (2018)

Outline

1. Motivation
2. Self-attention
3. Multi-head Attention
4. Positional Encoding
5. Stacks of Encoder & Decoder Layers
6. **The Transformer**

The Transformer

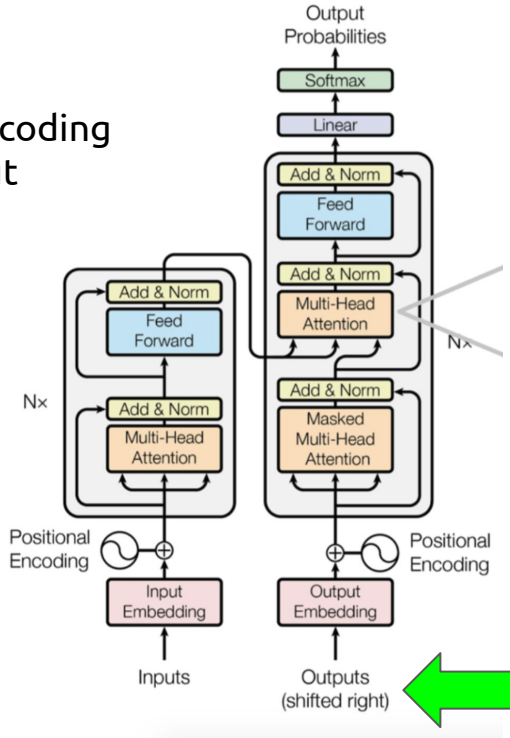
The Transformer removed the recurrency mechanism thanks to self-attention.



#Transformer Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.. Attention is all you need. NeurIPS 2017.

The Transformer

Positional Encoding over the input sequence.



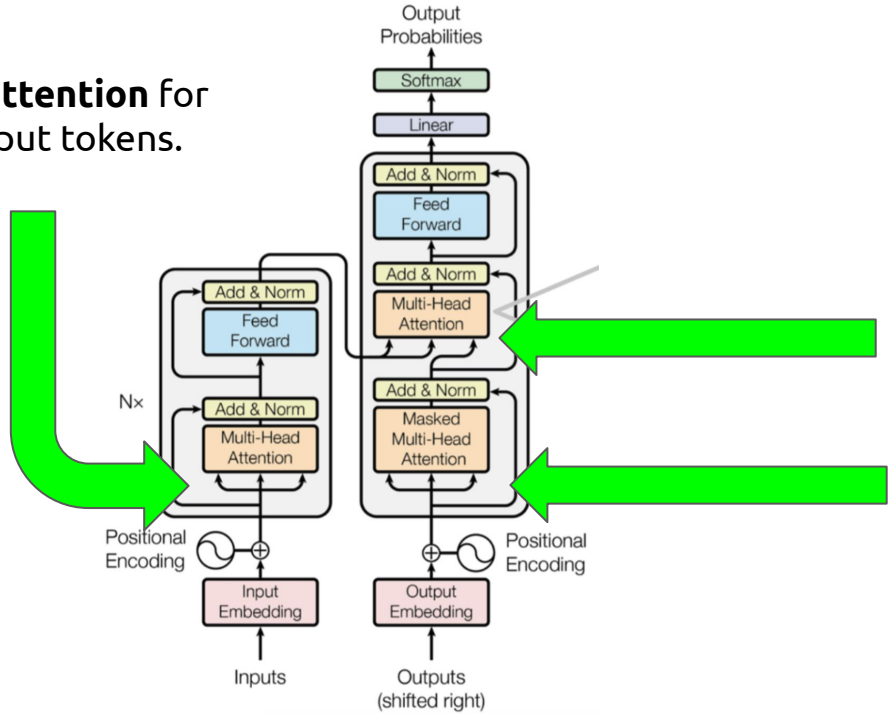
Positional Encoding over the output sequence.

Auto-regressive (at test).

#Transformer Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.. Attention is all you need. NeurIPS 2017.

The Transformer

Self-attention for the input tokens.

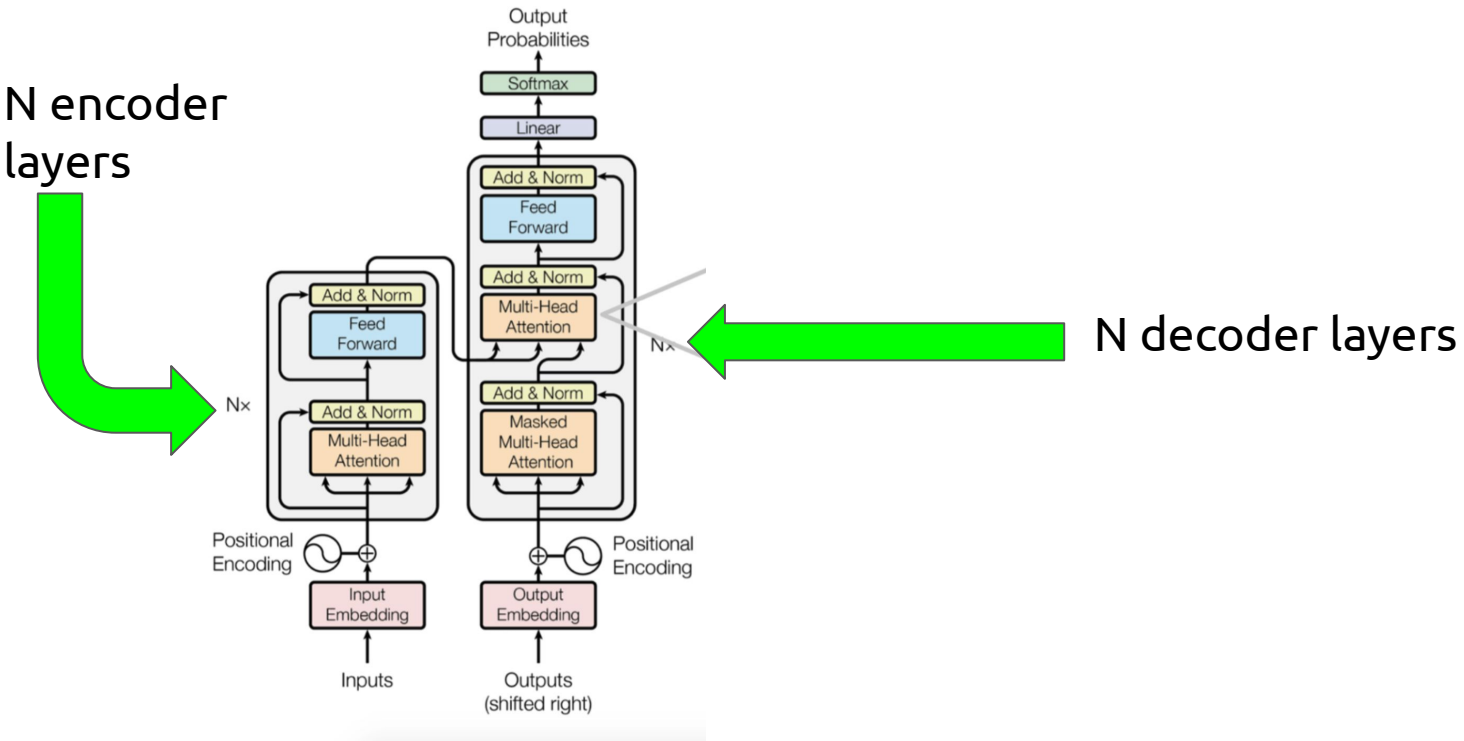


Cross-Attention (or inter-attention) between input and output tokens

Self-attention for the output tokens.

#Transformer Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.. Attention is all you need. NeurIPS 2017.

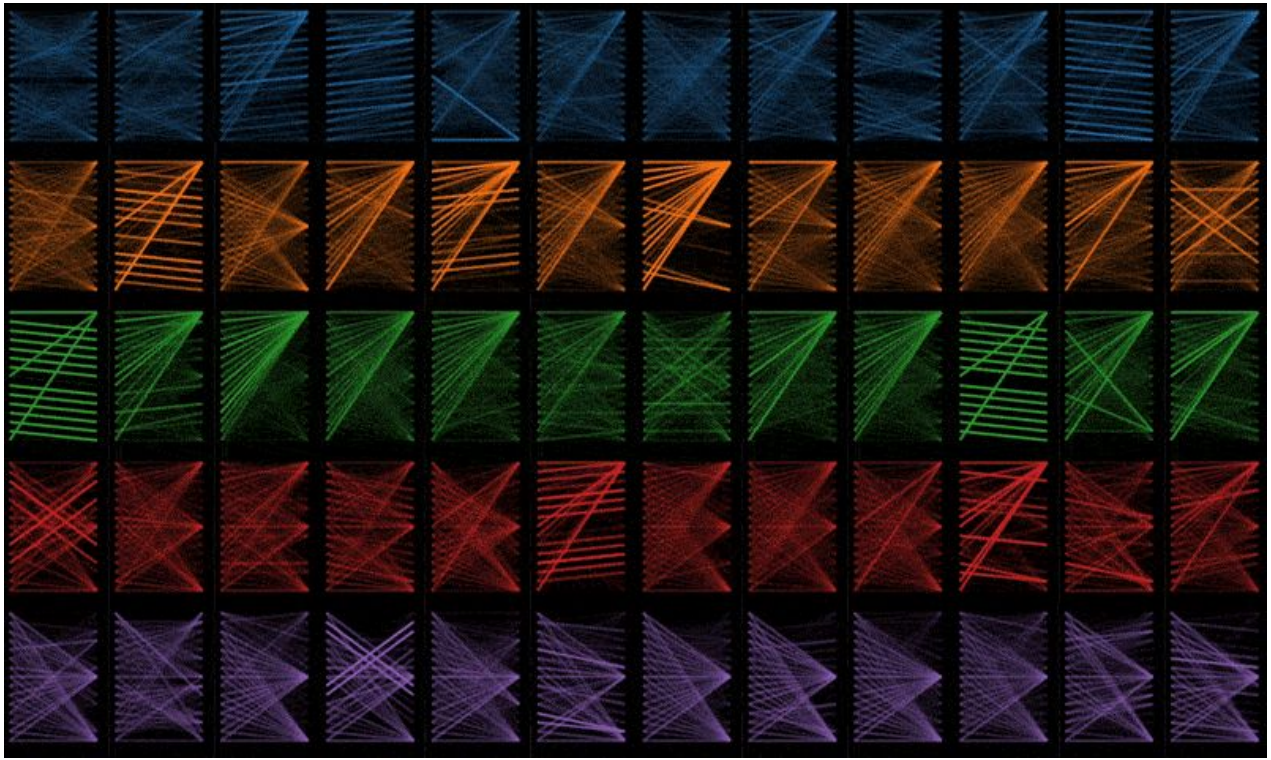
The Transformer: Layers



#Transformer Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.. Attention is all you need. NeurIPS 2017.

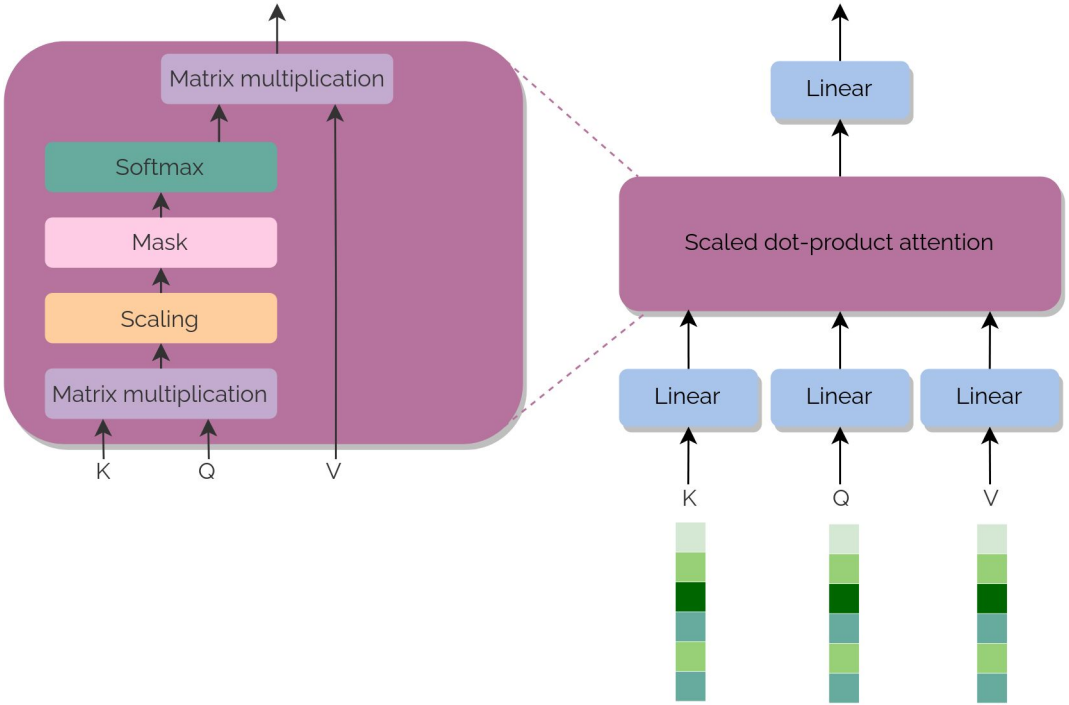
The Transformer: Layers

A birds-eye view of attention across all of the model's layers and heads



#BertViz Vig, Jesse. "A multiscale visualization of attention in the transformer model." ACL 2019. [code] [tweet]

The Transformer: Visualization



Maria Ribalta, Pere-Pau Vàzquez, "Visualization is all you need". UPC GCED 2020.

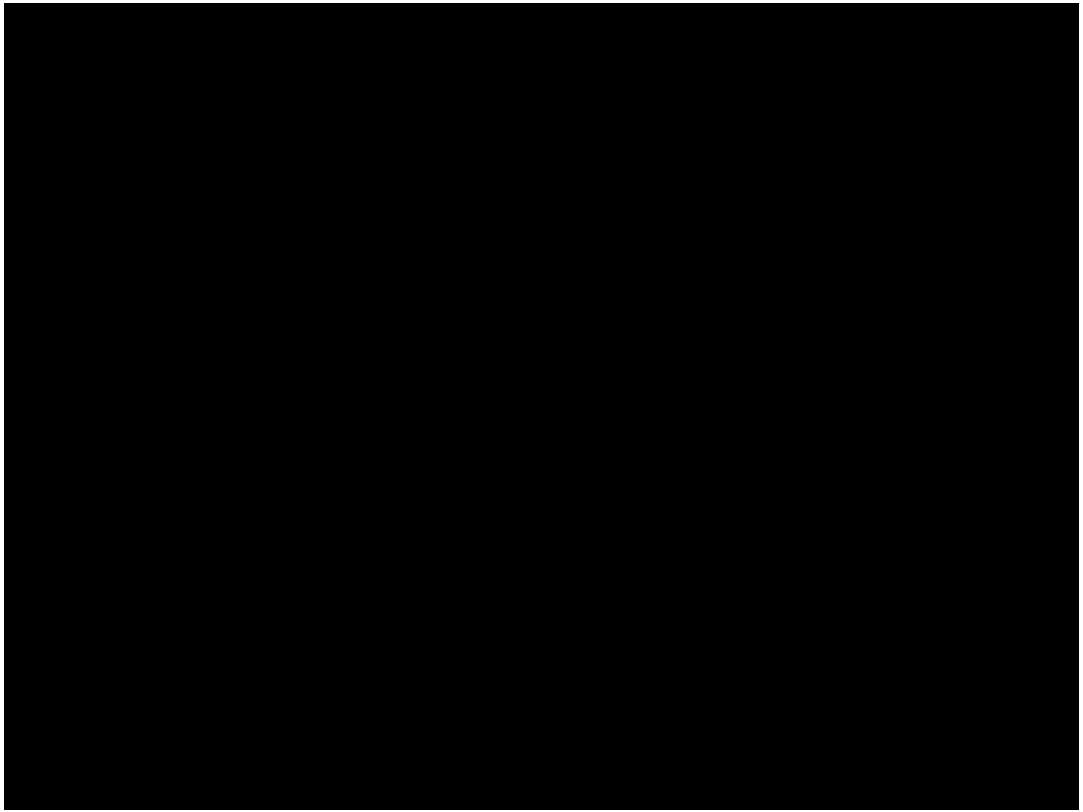
Are Transformers for Language only ? NO !!

#ViT Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." ICLR 2021. [blog] [code]

Outline

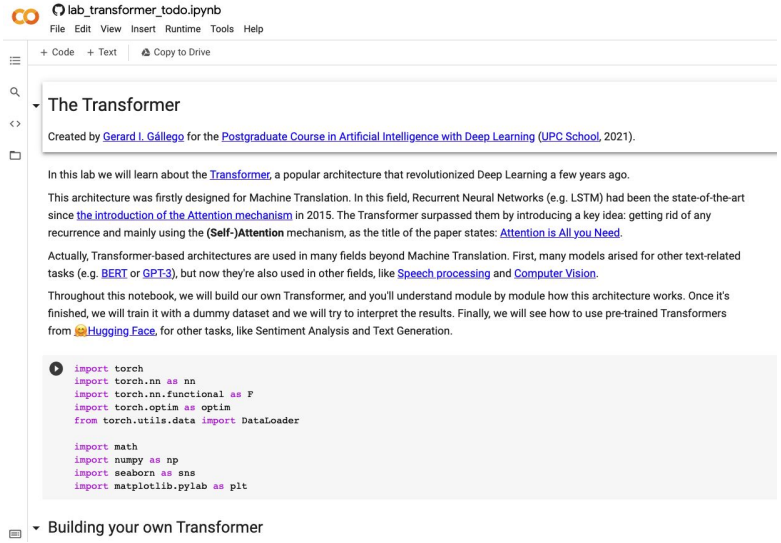
1. Motivation
2. Self-attention
3. Multi-head Attention
4. Positional Encoding
5. Stacks of Encoder & Decoder Layers
6. The Transformer

Teaser of Kostas Derpanis lecture



Kostas Derpanis, York University (2022) - @CSProfKGD

(extra) PyTorch Lab on Google Colab



Gerard Gallego
gerard.ion.gallego@upc.edu

PhD Candidate
Universitat Politecnica de Catalunya
Technical University of Catalonia

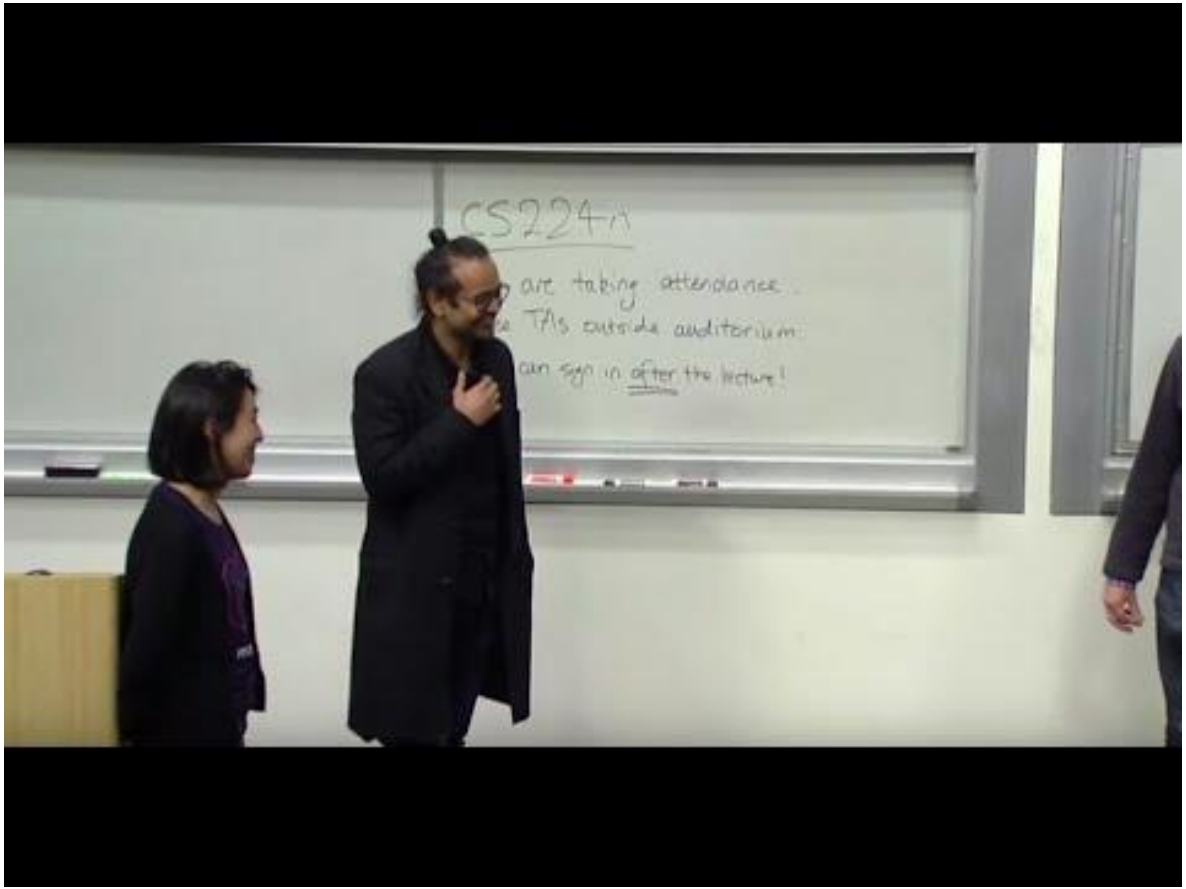
DL resources from UPC Telecos:

- Lectures (with Slides & Videos)
- Labs

Software

- Transformers in HuggingFace.
- GPT-Neo by EleutherAI
  - Similar results to GPT-3, but smaller and open source.
- Andrej Karpathy, minGPT (2020).

Learn more

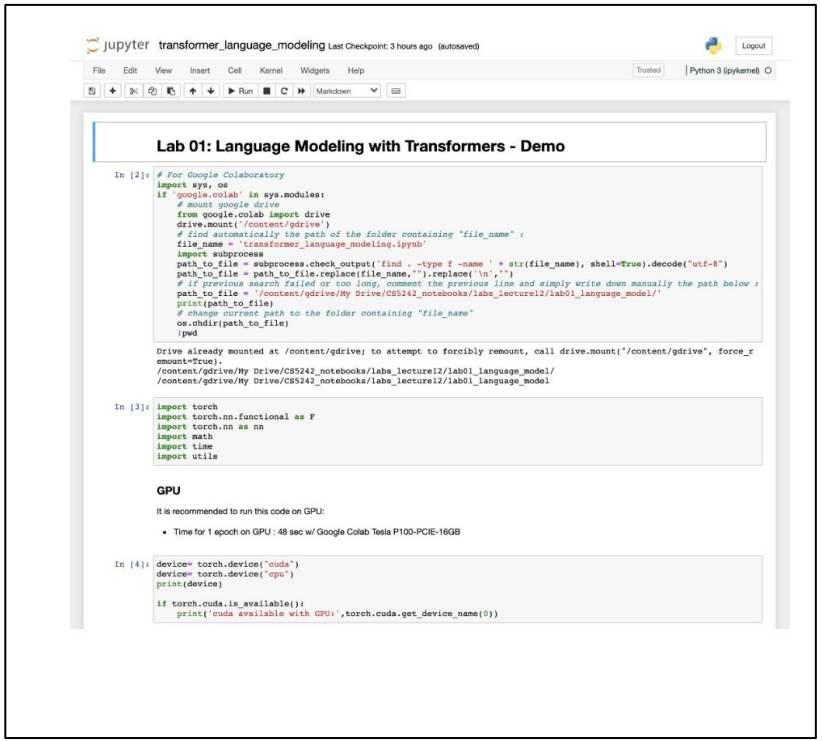


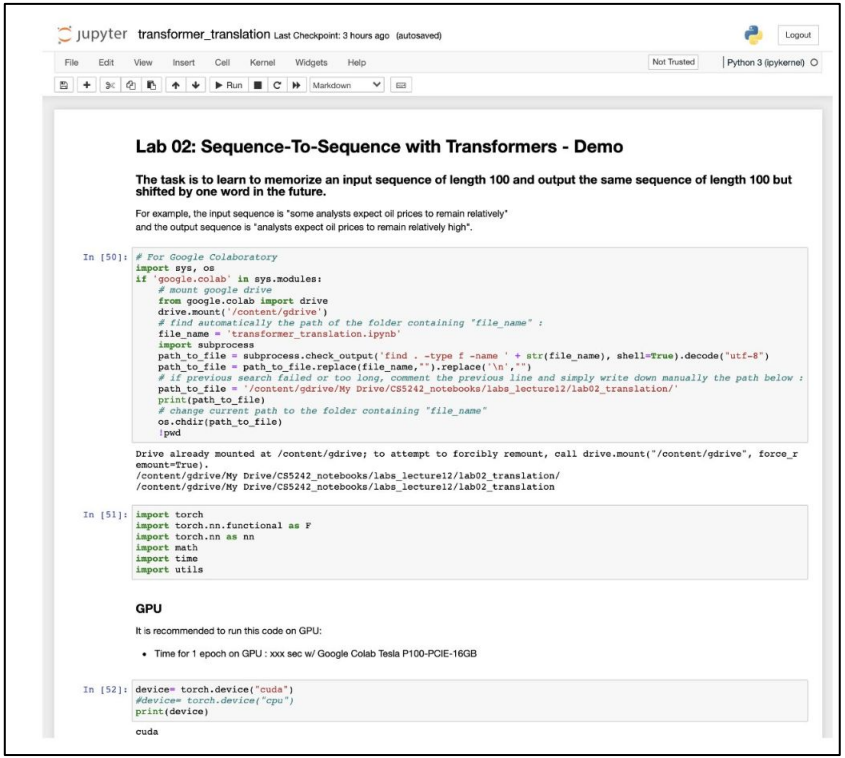
Ashish Vaswani, Stanford CS224N 2019.

Learn more

- Tutorials
  - Lilian Weg, The Transformer Family. Lil'Log 2020
- Twitter threads
  - Christian Wolf (INSA Lyon)
- Scientific publications
  - **#Perceiver** Jaegle, Andrew, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. "Perceiver: General perception with iterative attention." arXiv preprint arXiv:2103.03206 (2021).
  - **#Longformer** Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).
  - Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. Transformers are RNNs: Fast autoregressive transformers with linear attention. ICML 2020.
  - Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, Razvan Pascanu, "Multiplicative Interactions and Where to Find Them". ICLR 2020. [tweet]
  - Self-attention in language
    - Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733.
  - Self-attention in images
    - Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, Ł., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. ICML 2018.
    - Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks." In CVPR 2018.

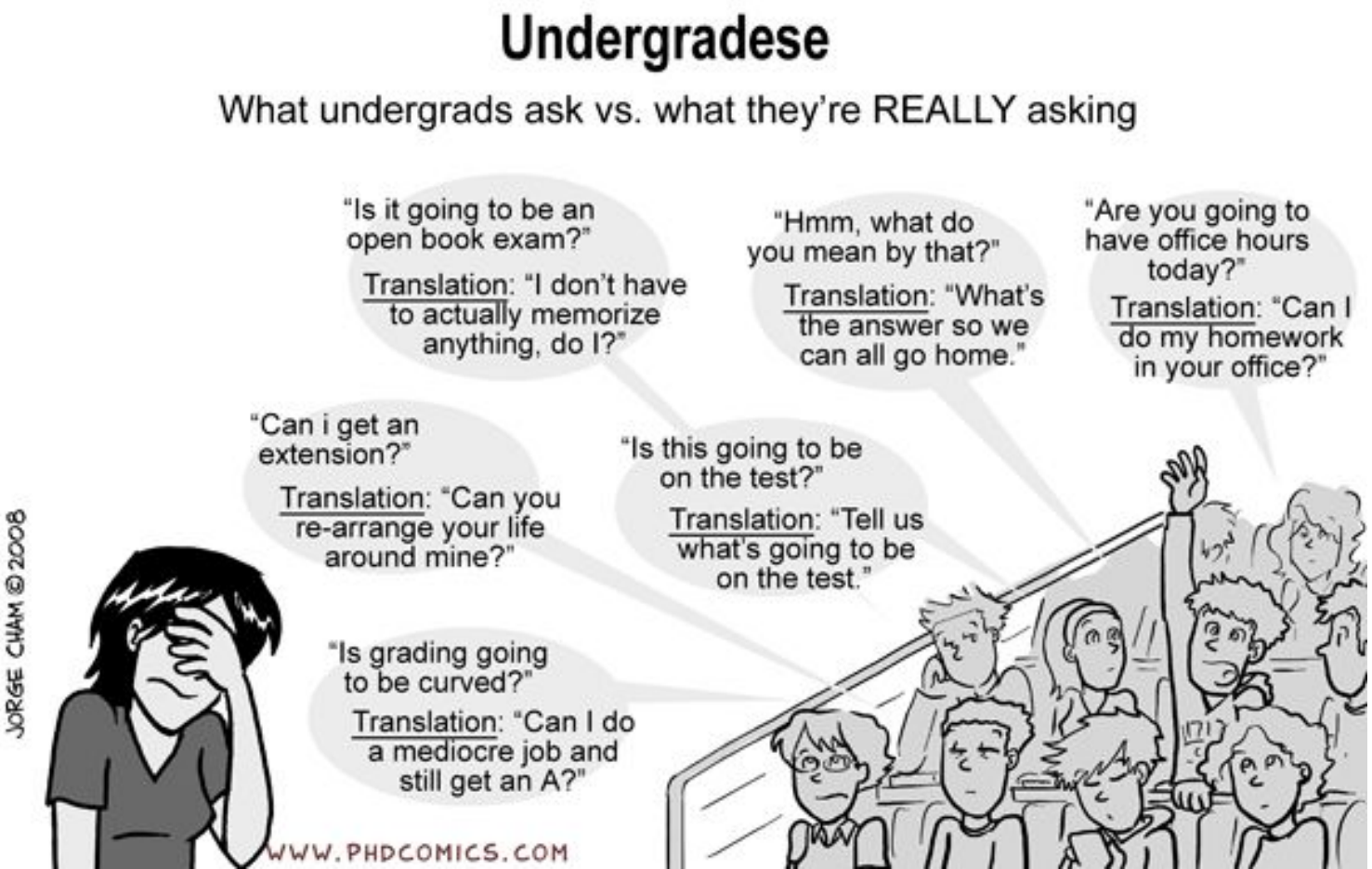
Learn more





Xavier Bresson, NUS 2021. [tweet]

Questions ?


Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM