DeepRob

Lecture 16
Transformers
University of Minnesota

Slides prepared by Athreyi Badithela and Karthik Desingh
Picture source: Transformers One (2024) movie

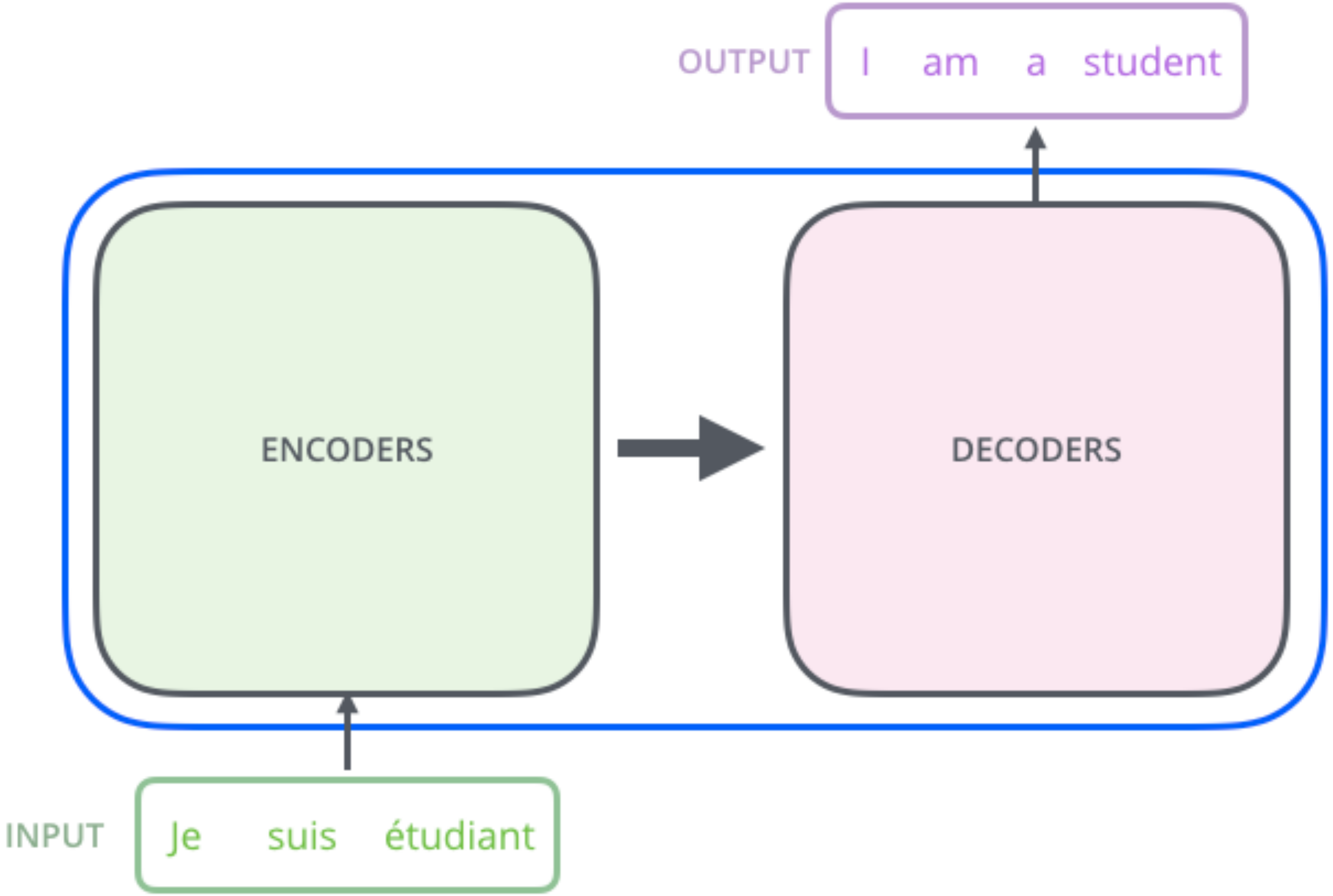
Classification – So Far

- CNNs
- RCNNs
- Faster RCNN
- MaskRCNN

In this lecture:

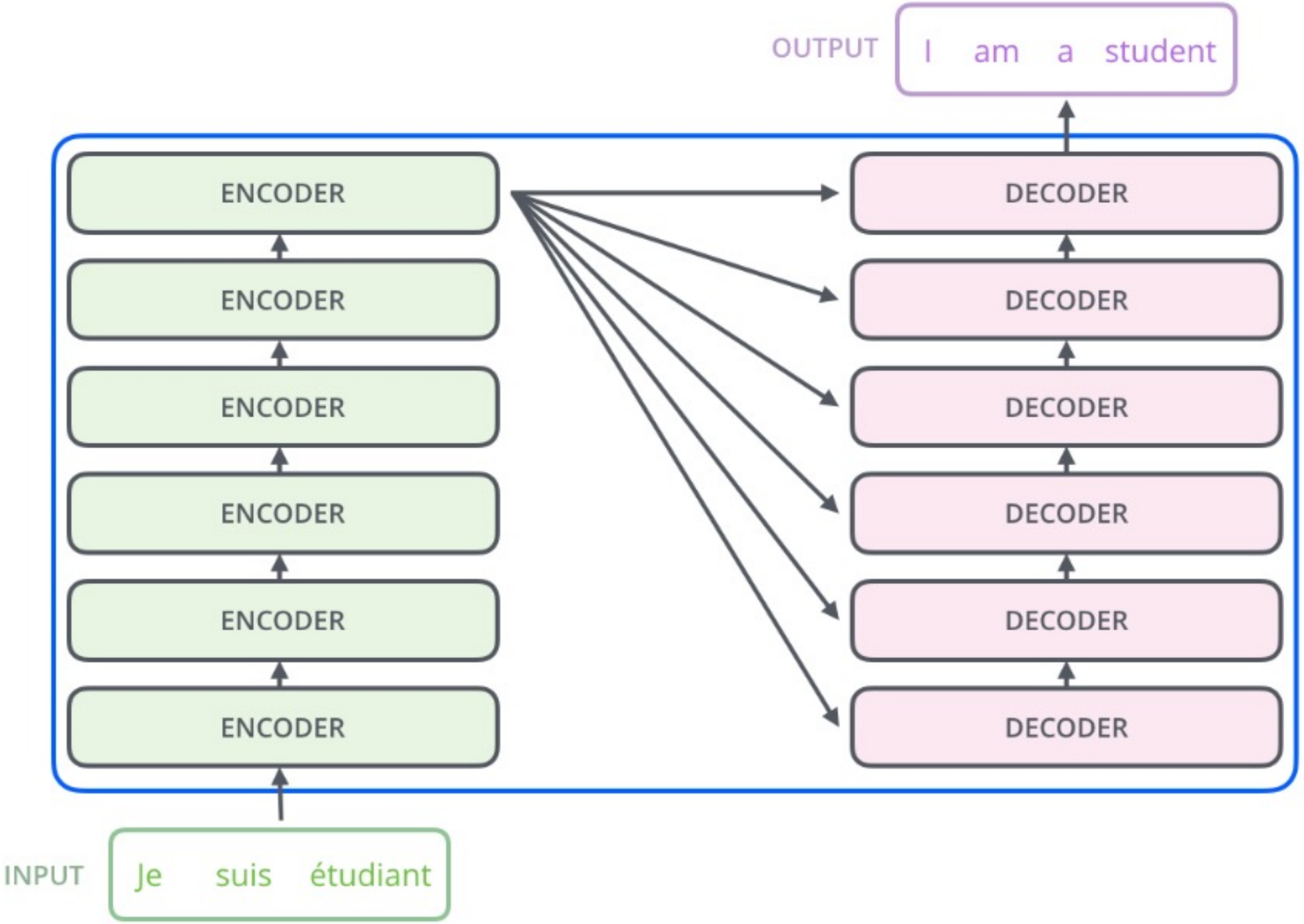
- Transformers in NLP
- Transformers in Vision
- Survey

What are Transformers?



Illustrated Transformers (Jay Alammar, 2018)

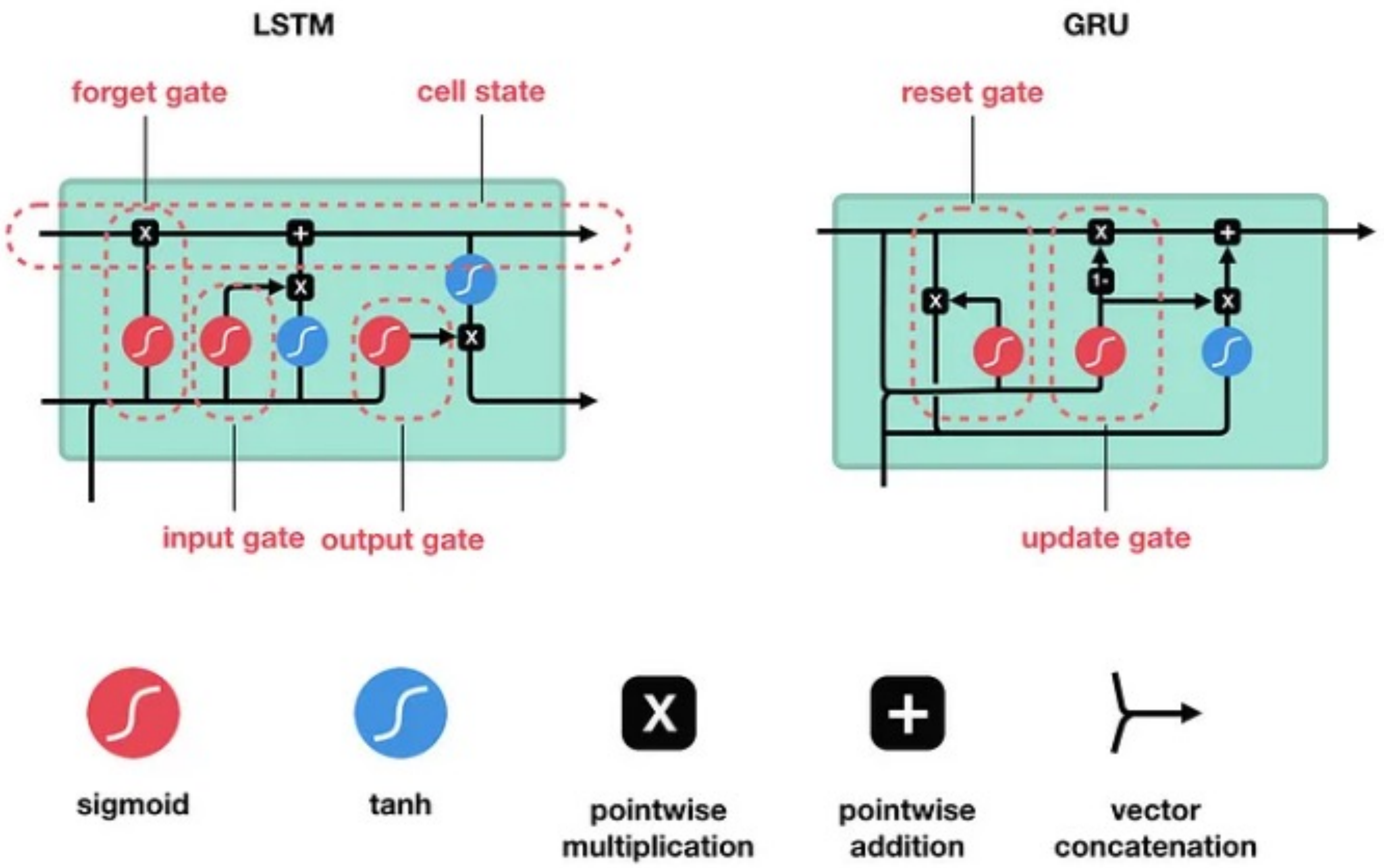
What are Transformers?



[Illustrated Transformers (Jay Alammar, 2018)]

What led to Transformers?

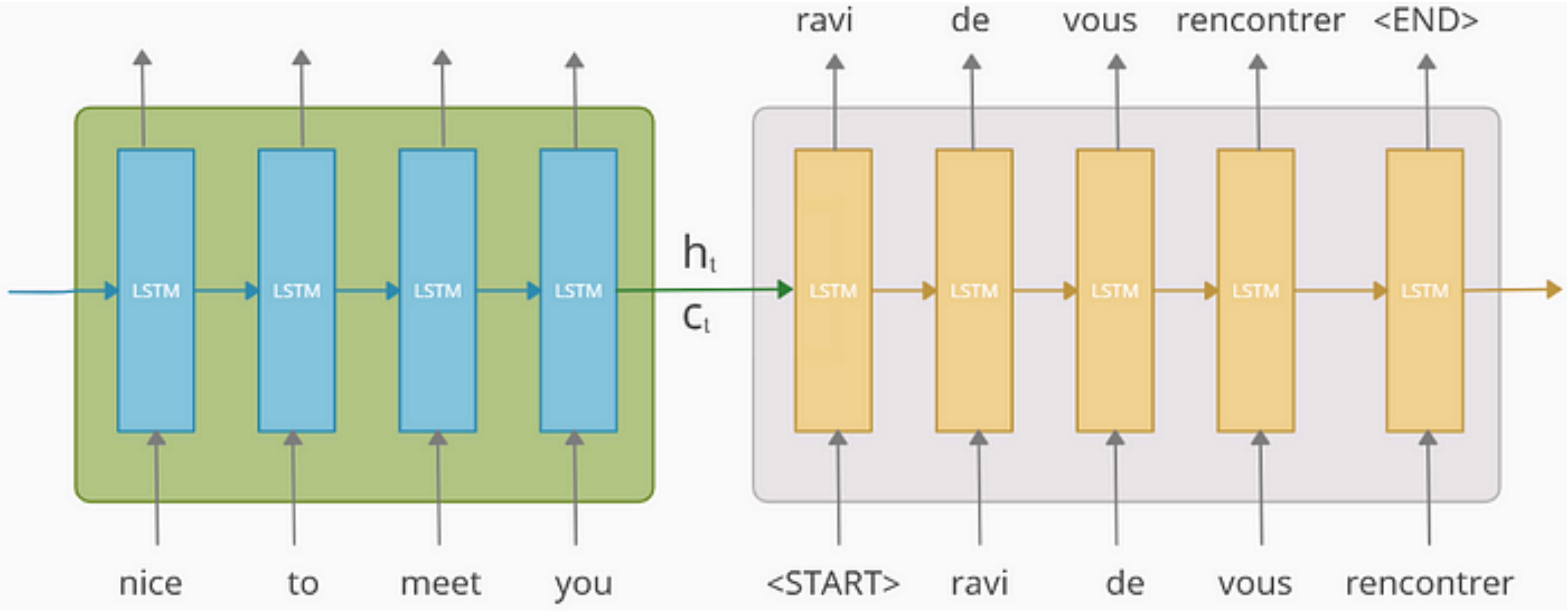
- LSTM
  - Attention mechanism
  - Sequential Processing
  - But hard to parallelize



[Illustrated Guide to LSTMs and GRUs: A step by step explanation]

What led to Transformers:

- Seq2Seq encoder decoder machine translation
    - RNNSearch introduces attention into the encoder-decoder structure

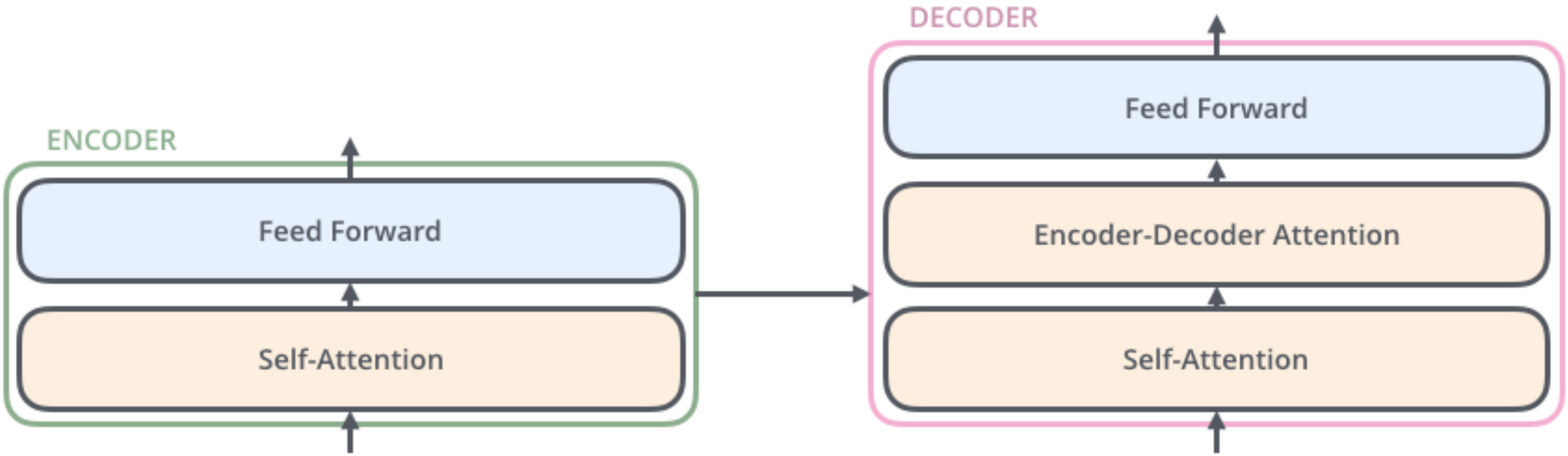


[Encoder-Decoder Seq2Seq Models Clearly Explained!!](#)

What led to Transformers:

- "Attention is All You Need" (2017)
  - Attention without recurrence is sufficient for machine translation, a controversial hypothesis at the time.
  - Apply self-attention to feed-forward networks
  - Parallelizable
  - Encoder-Decoder structure is adaptable

What led to Transformers:



[Illustrated Transformers (Jay Alammar, 2018)]

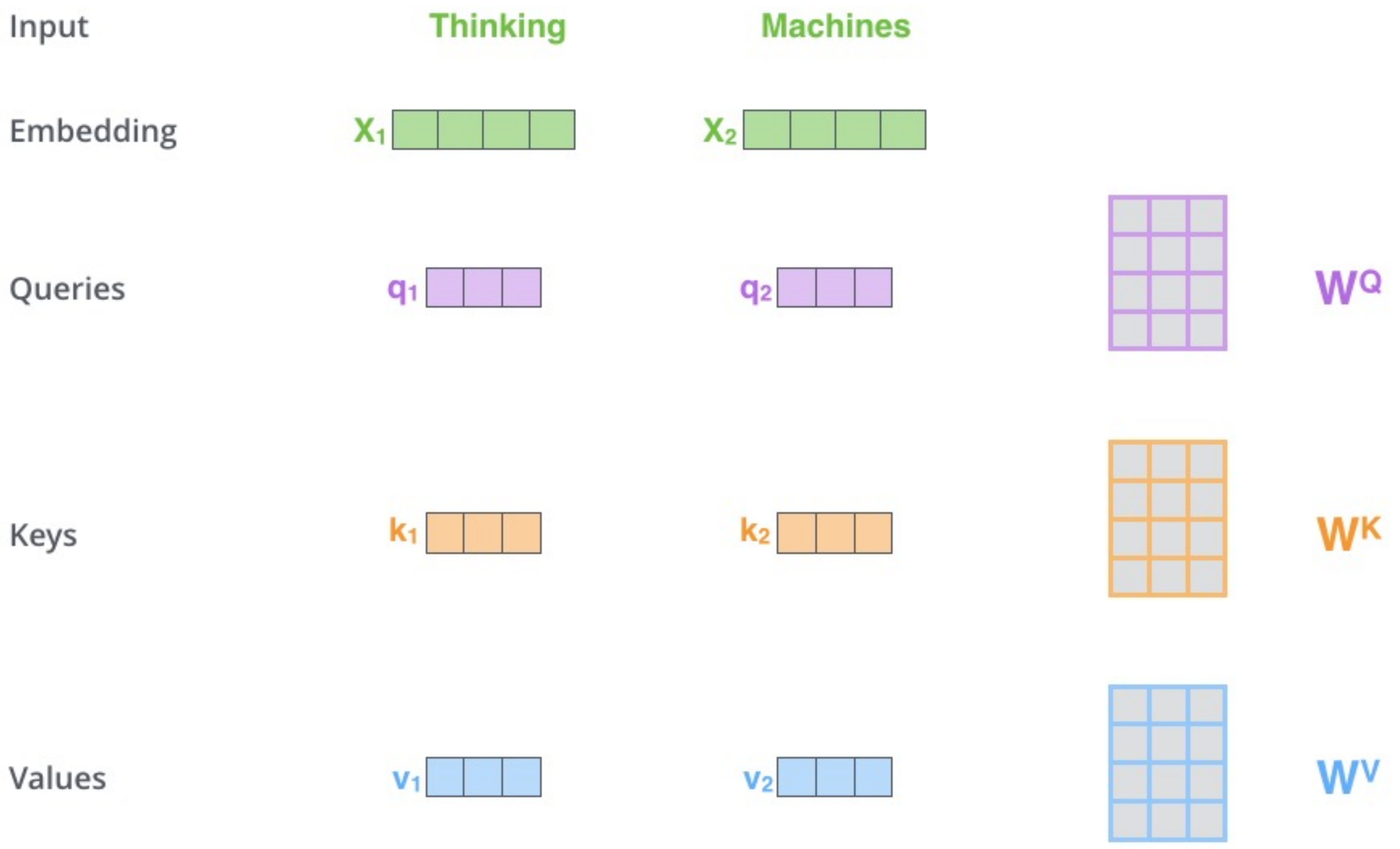
Inputs to Encoder

Queries: What we are looking for

Key: What we “offer”

Values: Value of the word

Obtained from learnable Weight Matrices.

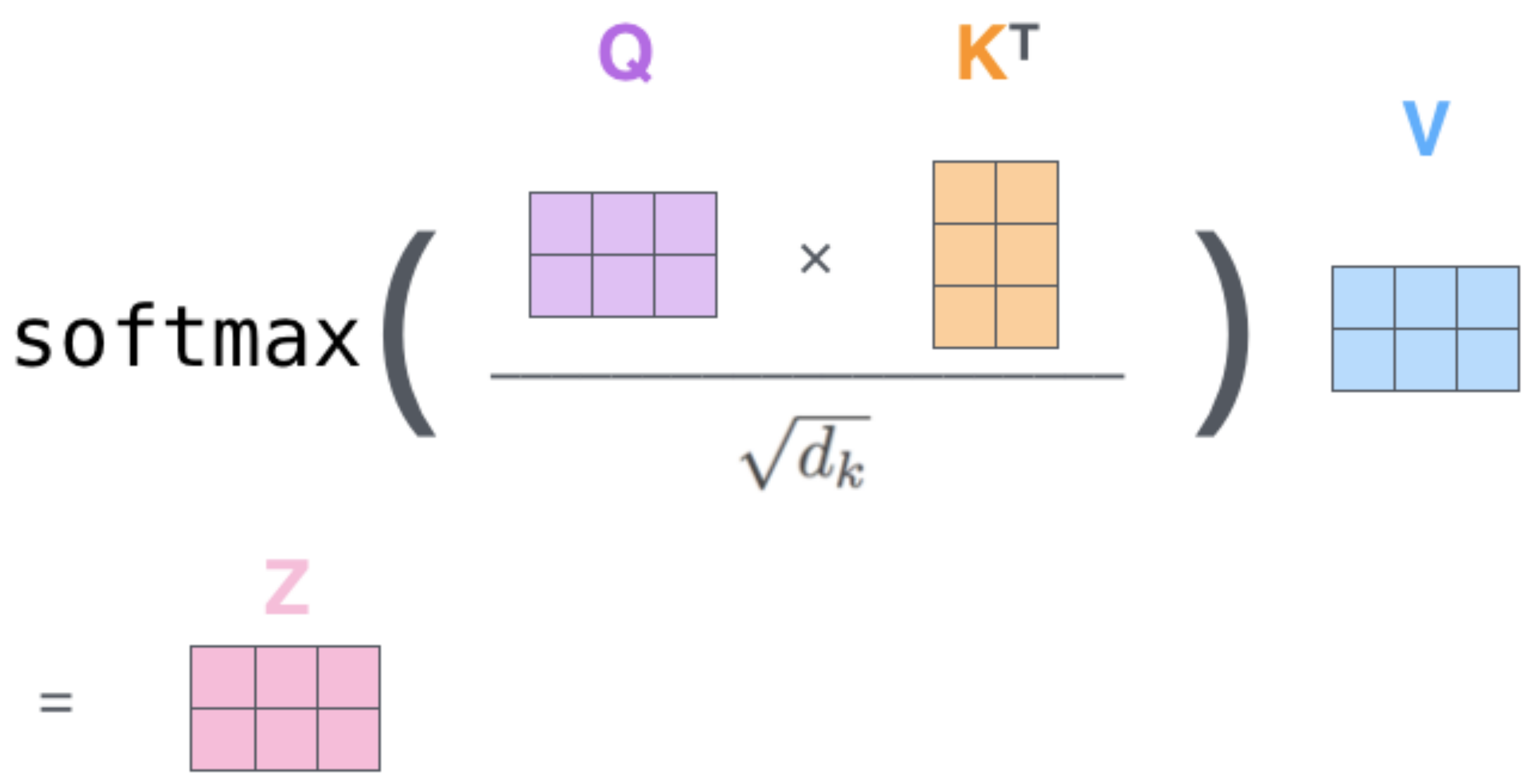


ultiplying x1 by the WQ weight matrix produces q1, the "query" vector associated with that word. We end up creating a "query", a "key", and a "value" projection of each word in the input sentence.

Illustrated Transformers (Jay Alammar, 2018)

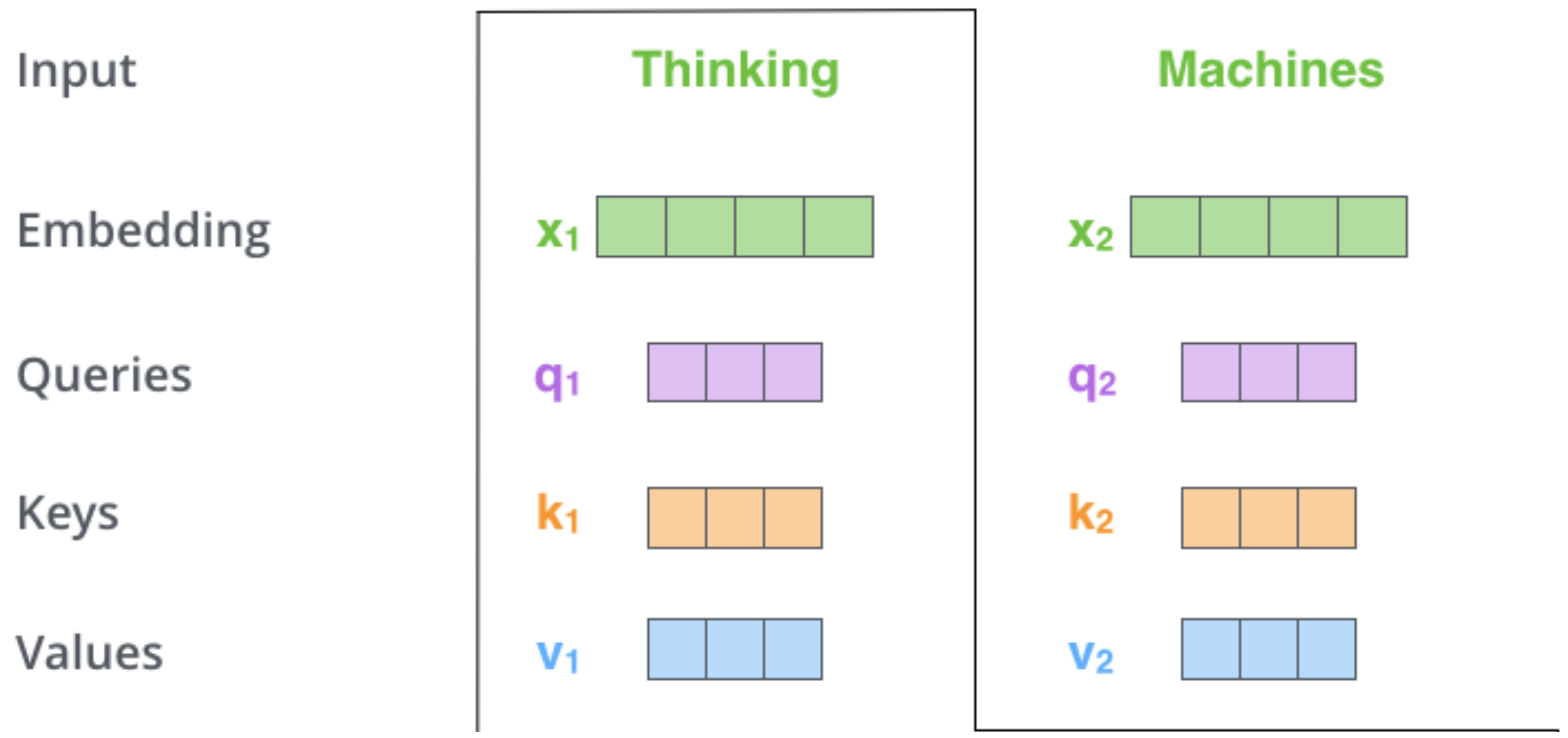
Scaled Dot Product Attention

- For the word thinking:
  - Compute query x key
  - Divide by $\sqrt{d_k}$
  - Take softmax



The self-attention calculation in matrix form

[Illustrated Transformers (Jay Alammar, 2018)]



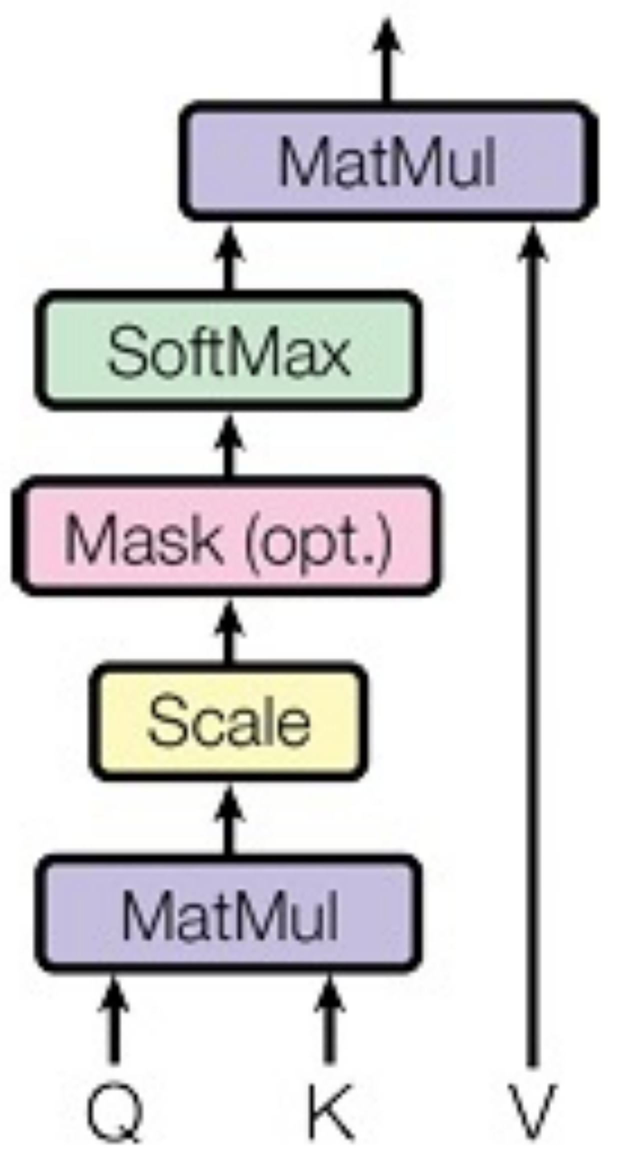
[Illustrated Transformers (Jay Alammar, 2018)]

Multi-Head Attention

1) This is our input sentence*

2) We embed each word*

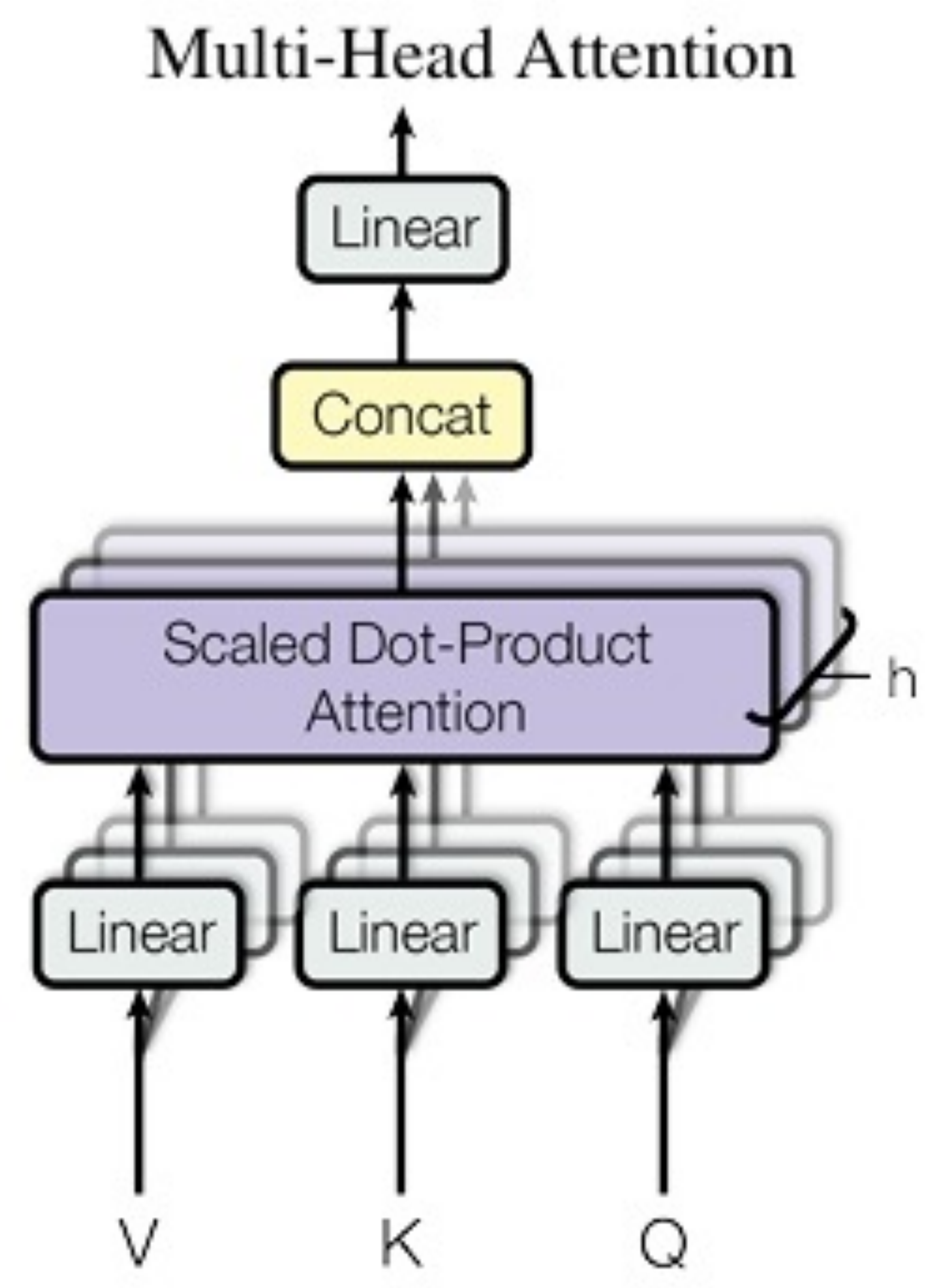
Scaled Dot-Product Attention

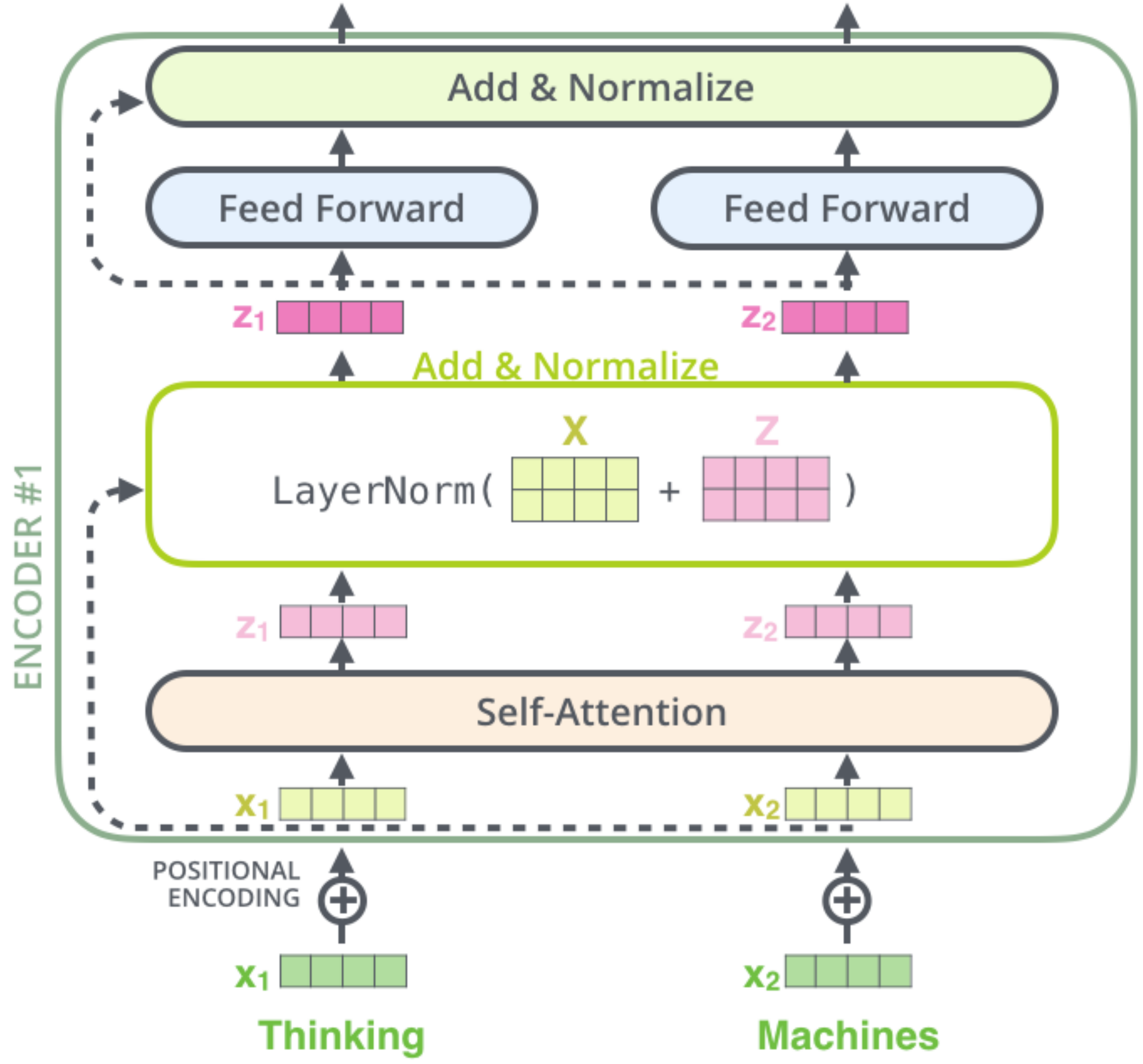


[Illustrated Transformers (Jay Alammar, 2018)]

Questions So Far?

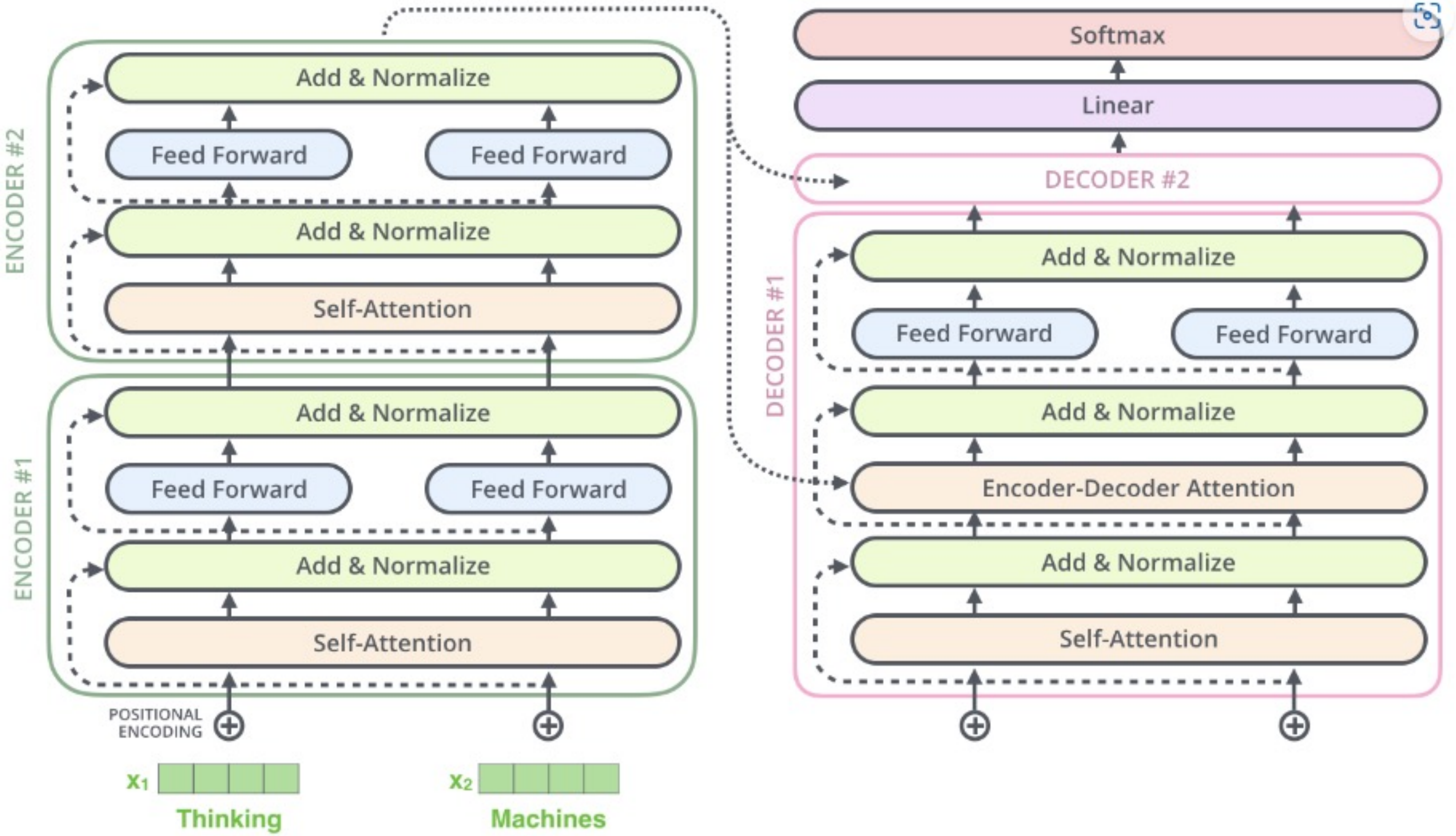
Putting it together





Illustrated Transformers (Jay Alammar, 2018)

Putting it together



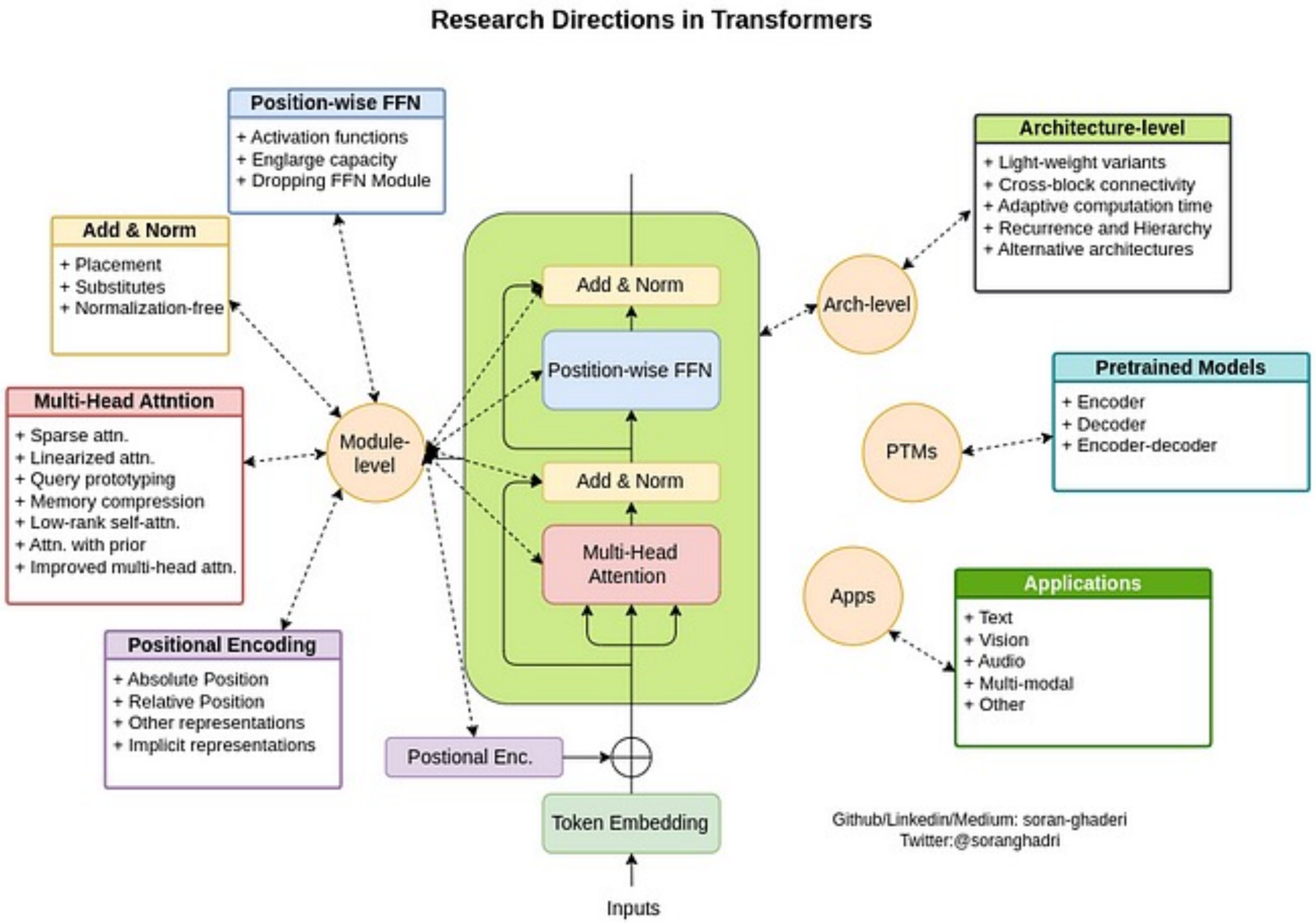
[Illustrated Transformers (Jay Alammar, 2018)]

Some finer details

- Why Residual layers?
  - Transformers are deep
  - Allow for smooth gradients
  - Retains positional information
- Why Positional Encoding?
  - MHA is permutation invariant
  - Position is important in certain tasks (e.g. language)
- Why Masking?
  - Prevent peeking future tokens
  - Improved learning + parallelization
- Why Add and Normalize?
  - Standardize for consistent mean and variance

Questions So Far?

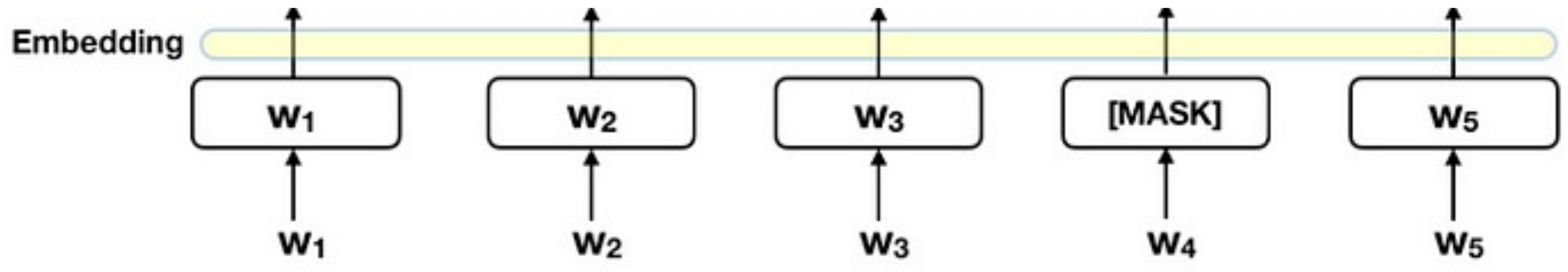
Further advancing Transformers



The Map of Transformers

BERT (2018)

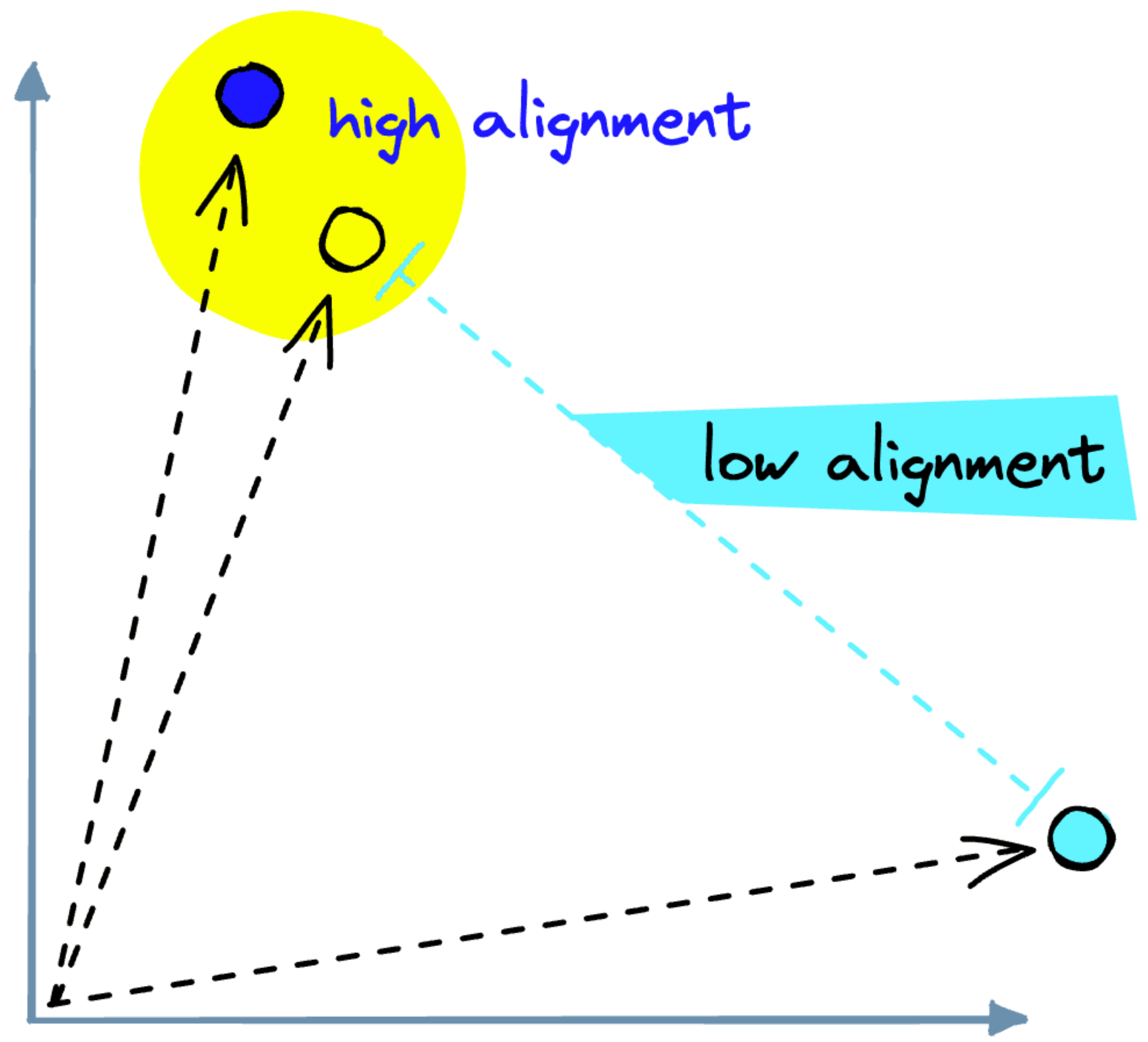
- Use of Encoder-only structure



[BERT Explained: State of the art language model for NLP]

Applying to Images

- Transformers use attention to measure the relationship b/w two vector embeddings.



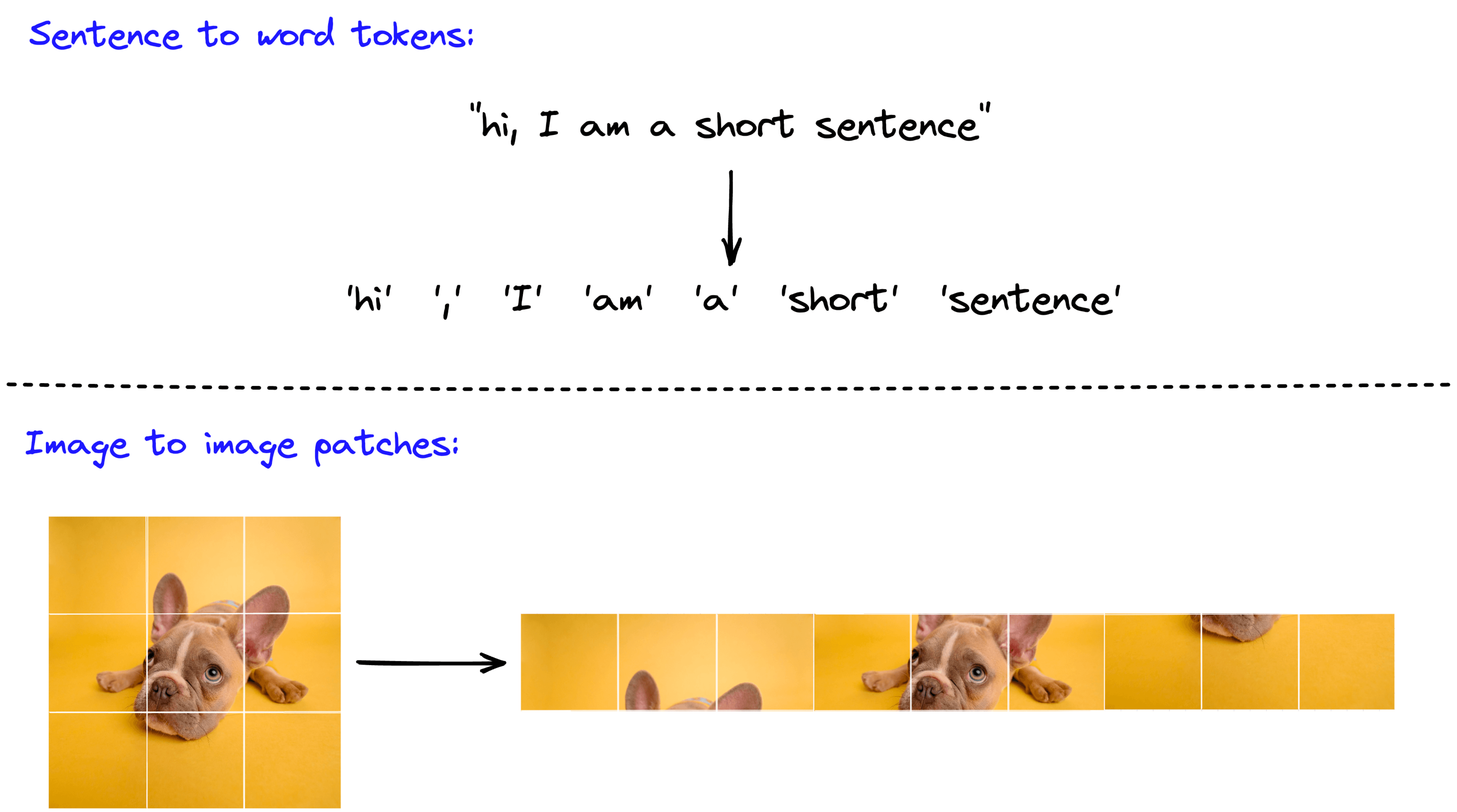
Vision Transformer (ViT) Explained

Applying to Images

- As we have seen, in NLP those pairs are tokens.
- In Vision the smallest unit for analysis would be a pixel.
- Self-attention is a quadratic operation.
- Pixel-wise self-attention is computationally expensive.

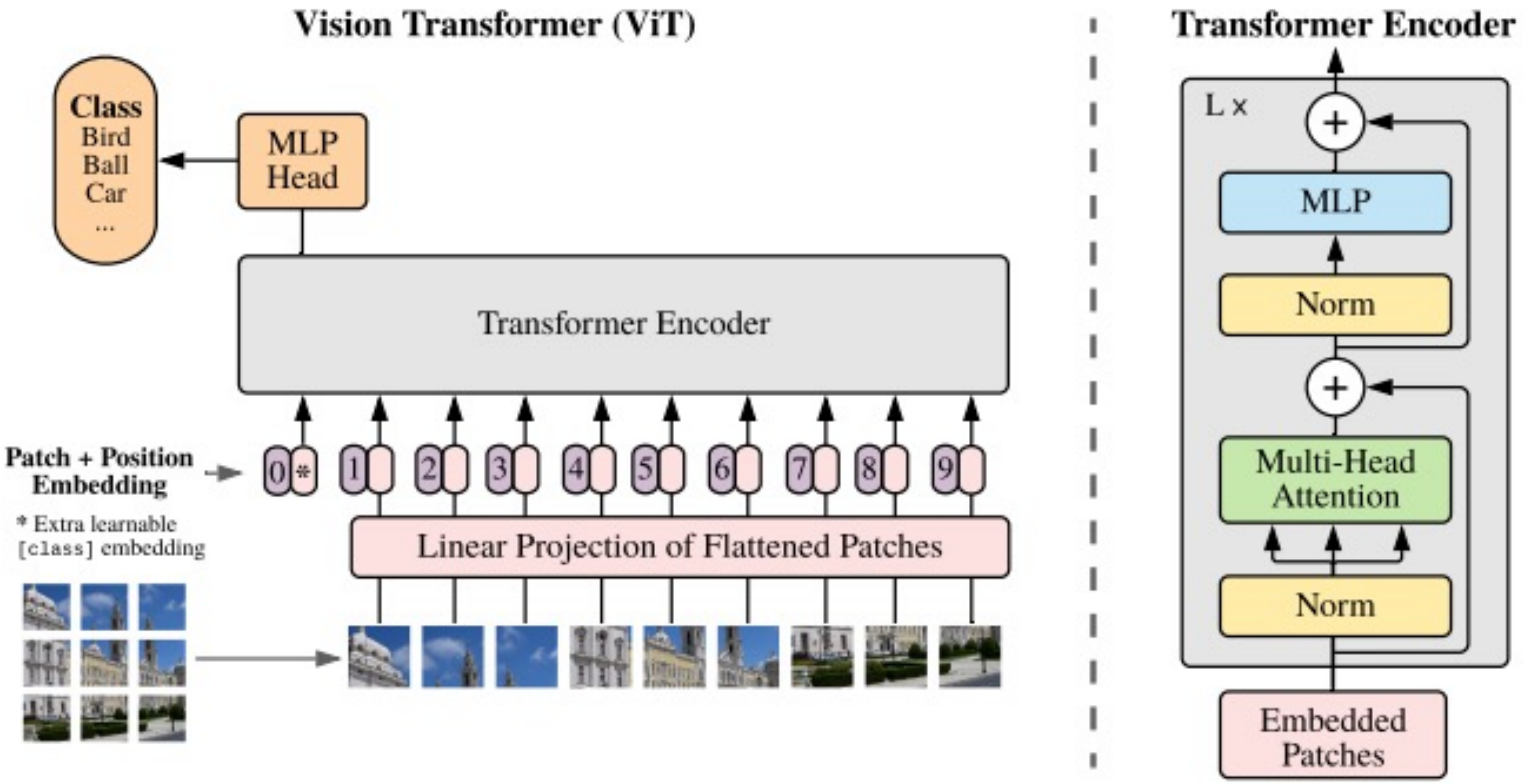
Applying to Images

- Instead of pixel, split the image into patches
- Create vector embeddings of image patches



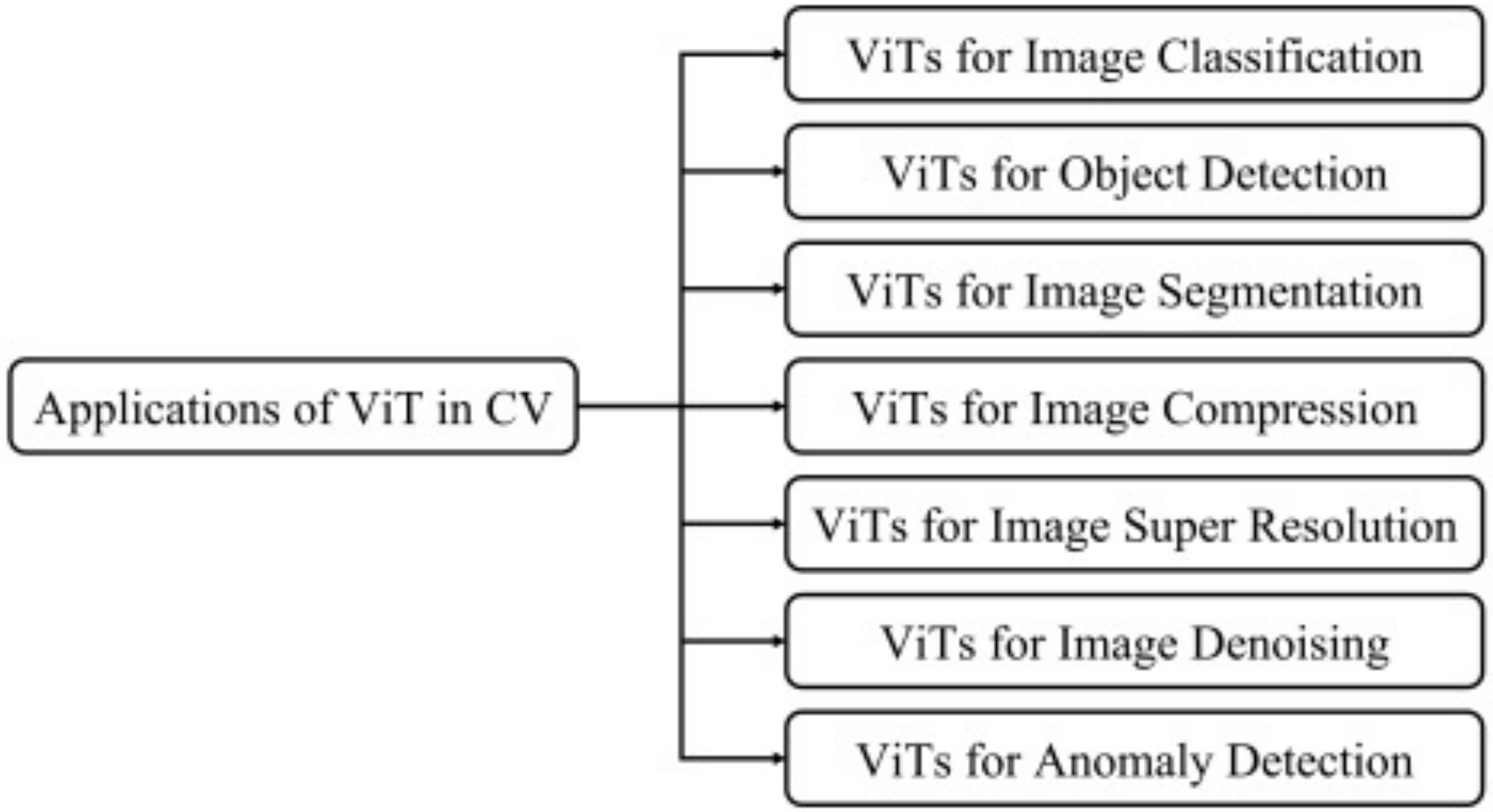
Vision Transformer (ViT) Explained

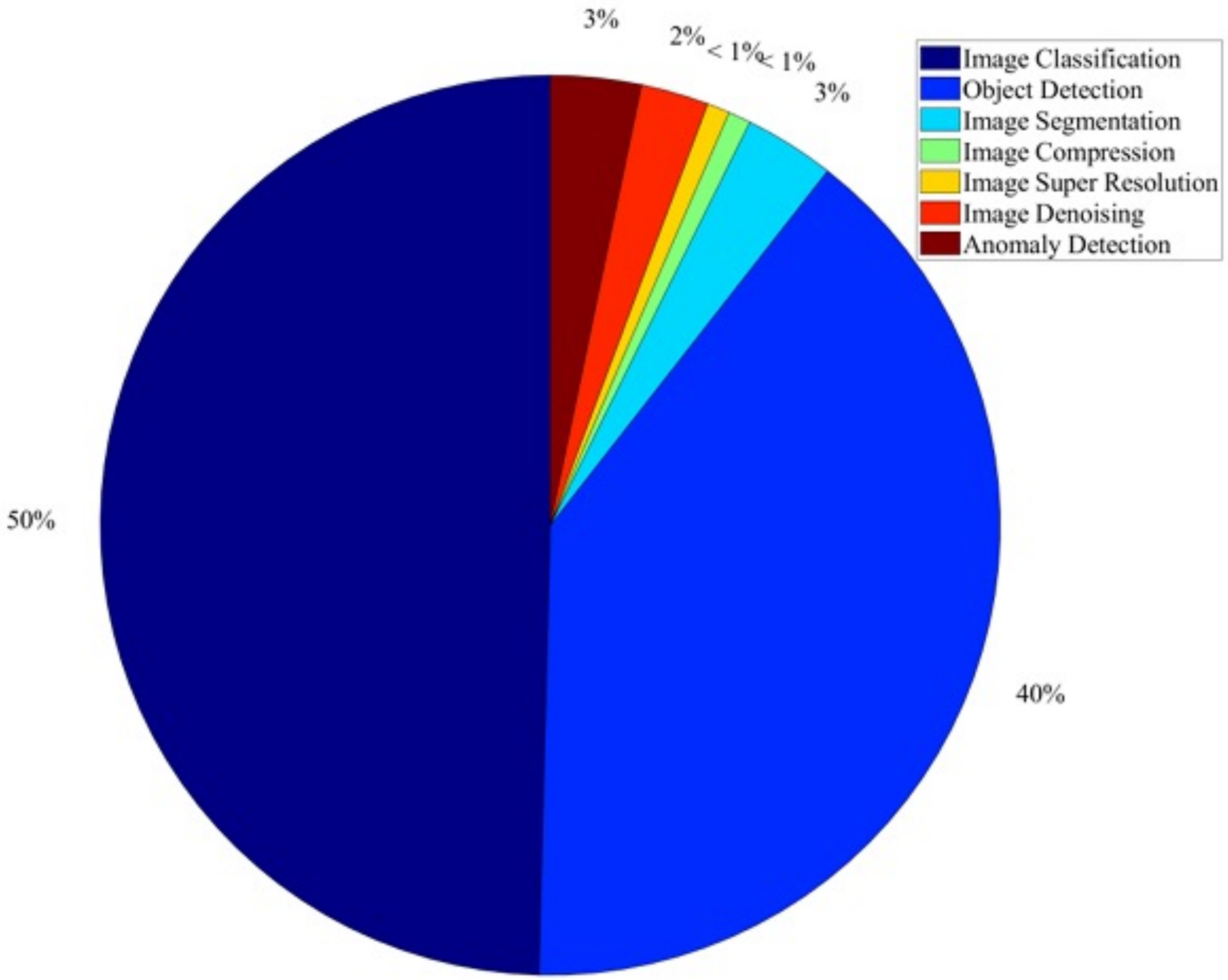
An Image is worth 16 x 16 words: Transformers for Image Recognition at Scale (2021)



ViT model from "An Image is worth 16 x 16 words: Transformers for Image Recognition at Scale"

Transformer for Computer Vision





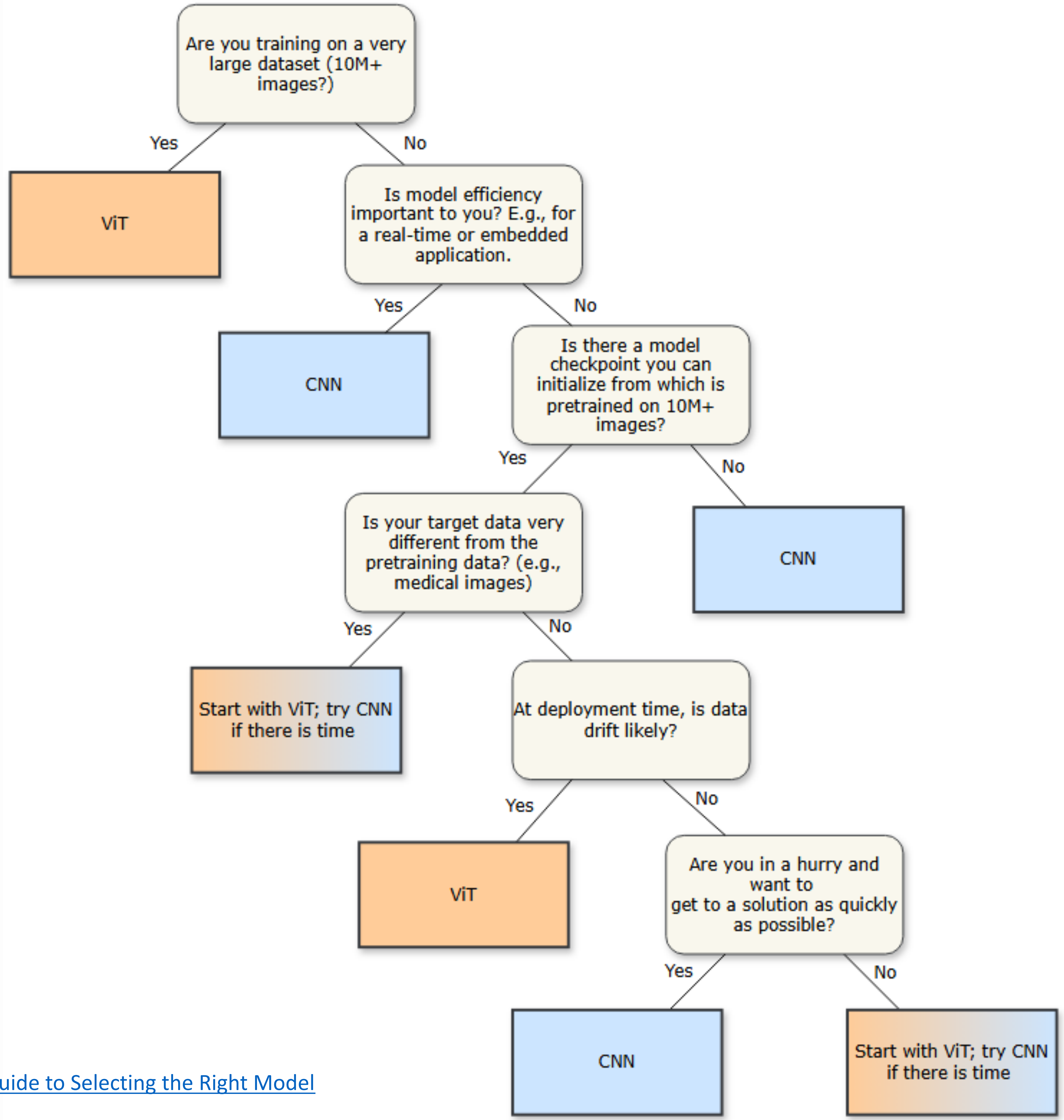
A Comprehensive Survey of Transformers for Computer Vision Applications (2022)

ViT vs CNNs – Image Classification

Field is rapidly growing!

- CNNs are compute efficient
- ViTs are more robust
- Hybrid ViT-CNN architectures.

Tons of variants – literature scales more than one can read.



CNN vs Vision Transformers: A Practitioner's Guide to Selecting the Right Model

Thank you!!

Next Lecture:

Pre-training Representations

DeepRob

Lecture 16
Transformers
University of Minnesota

Slides prepared by Athreyi Badithela and Karthik Desingh
Picture source: Transformers One (2024) movie