Introduction to transformers

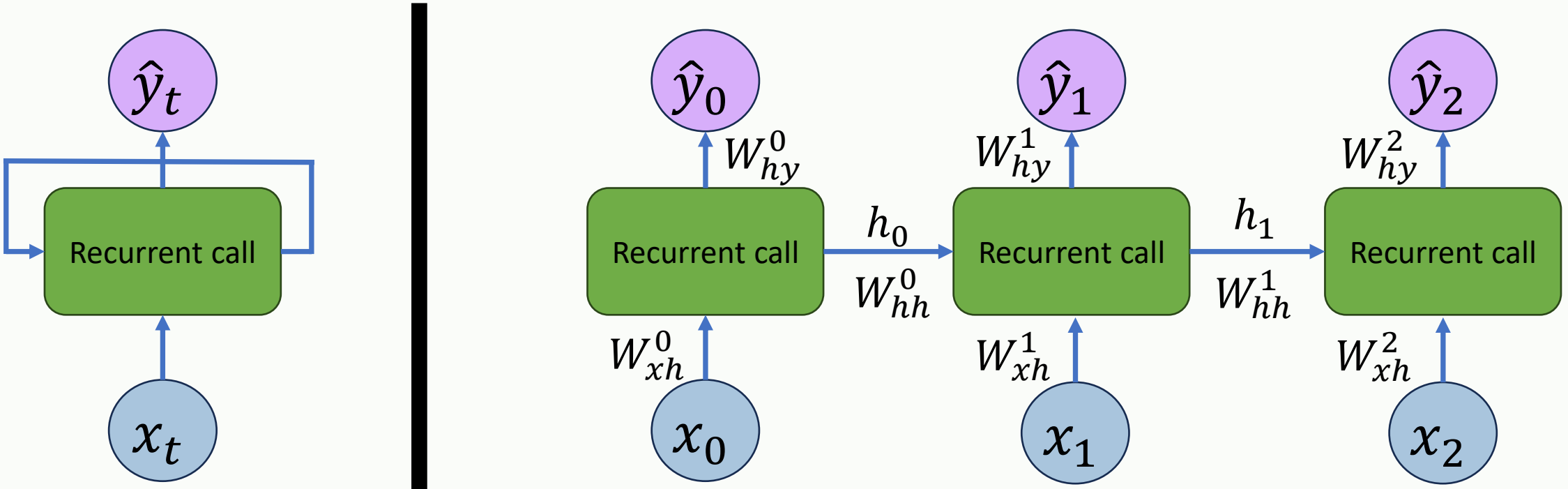
Peyman Tahghighi

Winter 2024

Learning objectives

- Why do we need an attention mechanism?
- How does attention work?
- What are encoders and decoders in transformers?
- Vision transformers.

Recurrent Neural Networks



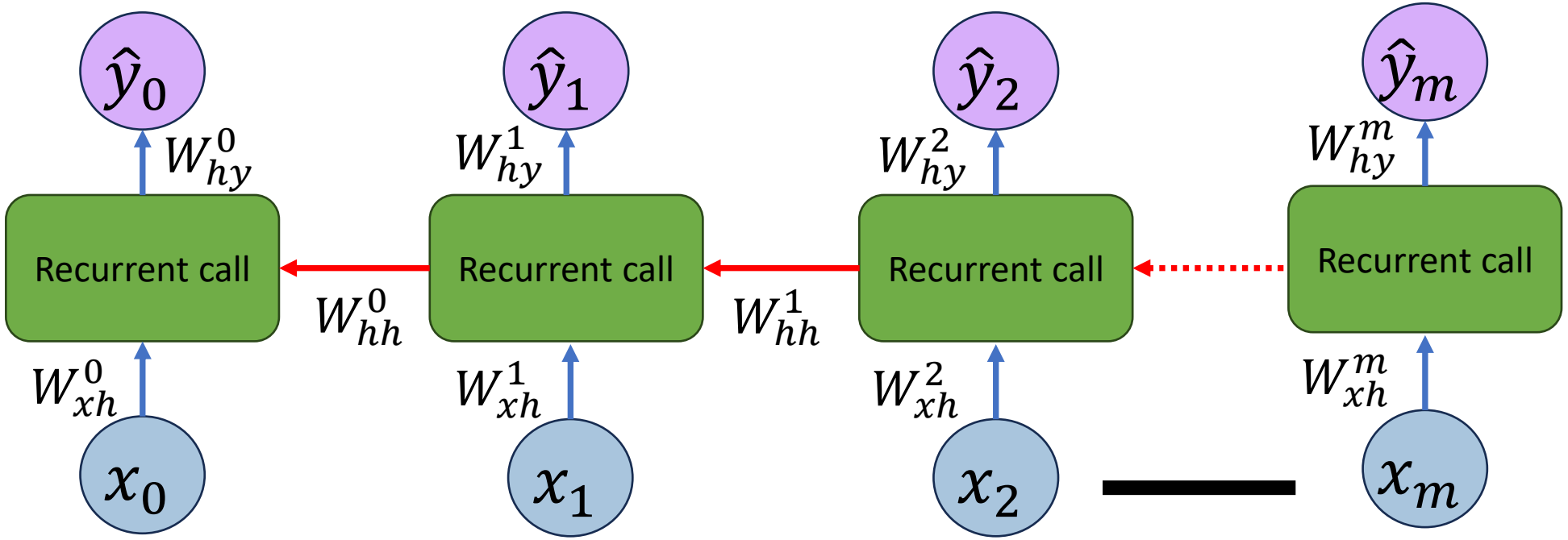
Output vector

$$\hat{y}_t = W_{hy}^T h_t$$

Hidden state update

$$h_t = \tanh(W_{hh}^T h_{t-1} + W_{xh}^T x_t)$$

Recurrent Neural Networks

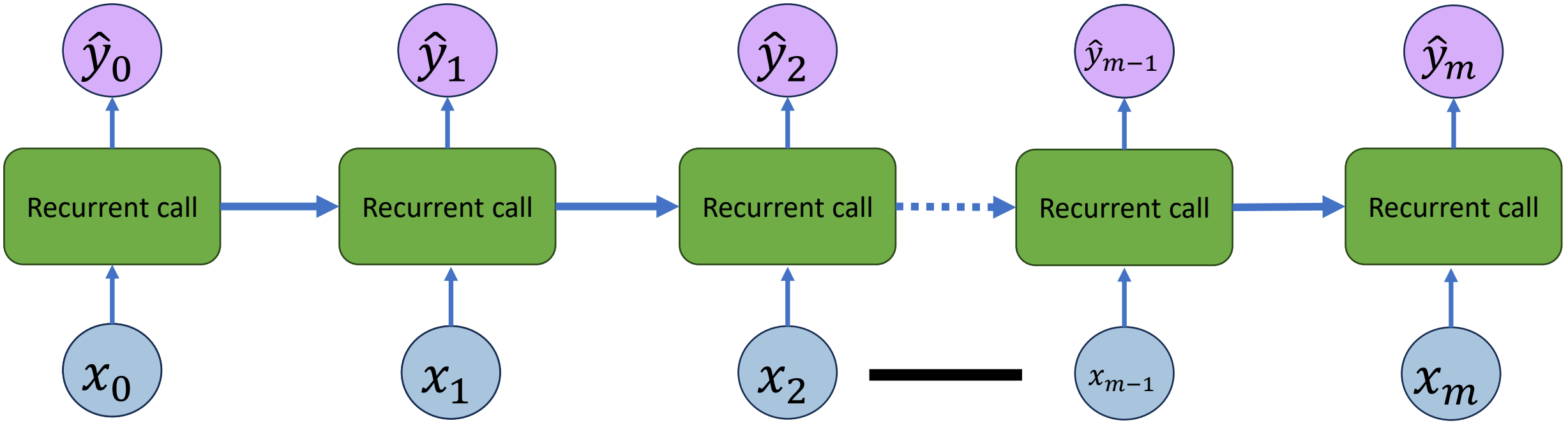


Many values > 1:
Exploding gradients

Many values < 1:
Vanishing gradients

Recurrent Neural Networks

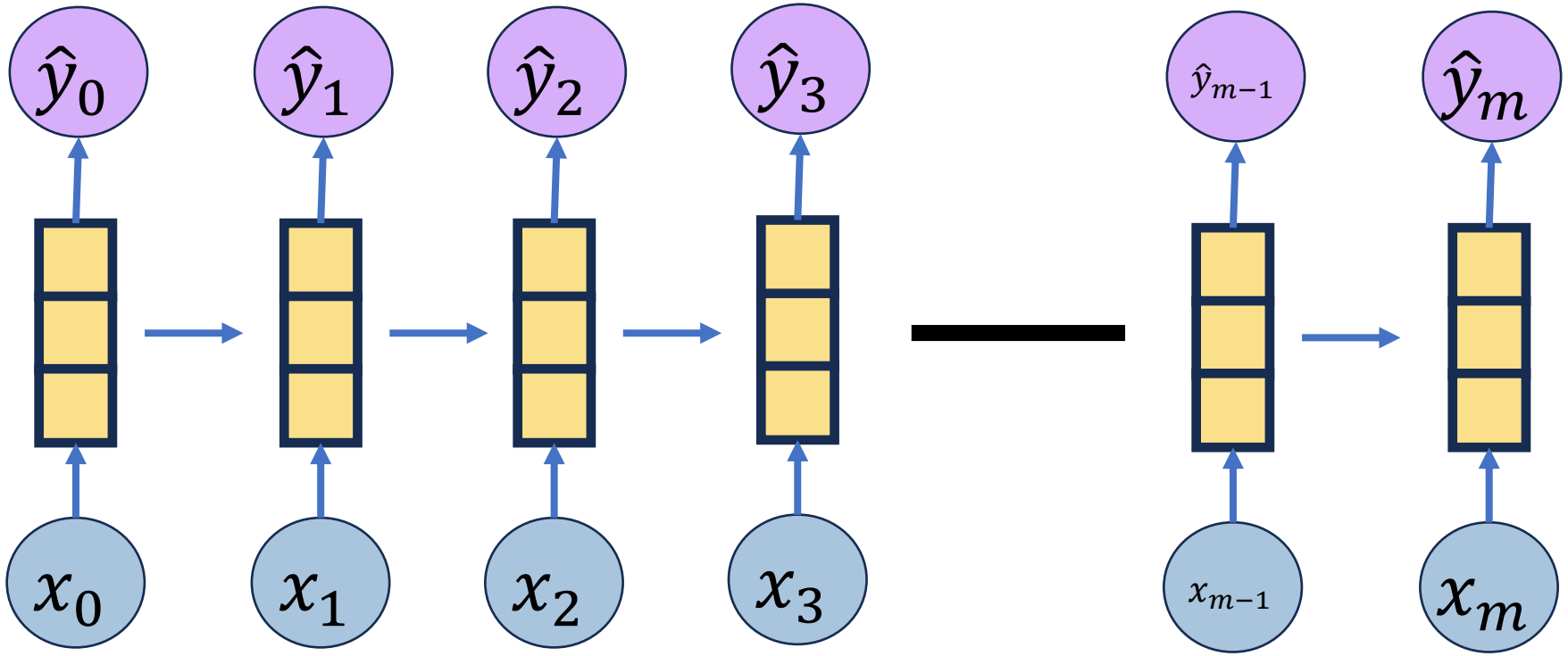
I grew up in France…. And I speak fluent….



Goal of Sequence Modeling

Limitations of RNN:

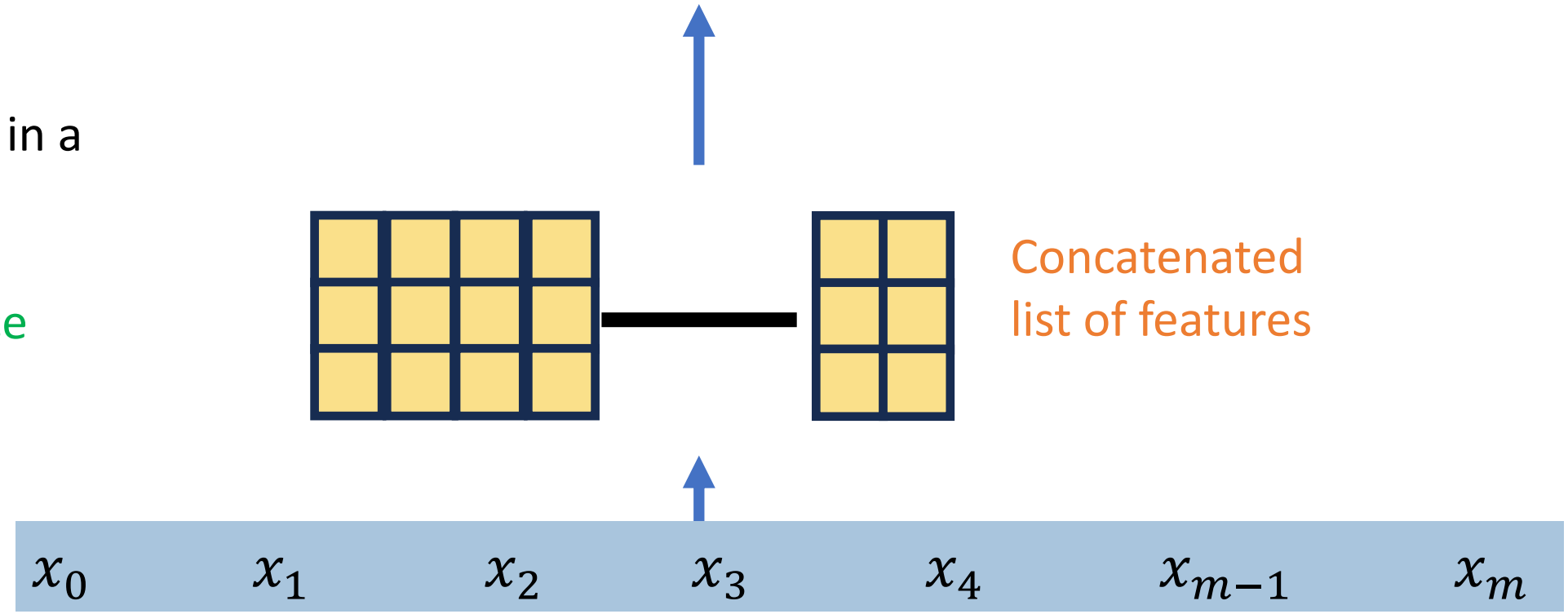
- Weak Long- Term Memory.
- Vanishing/Explodi ng Gradients.
- Not easy to parallelize.



Goal of Sequence Modeling

Feed Everything in a dense network:

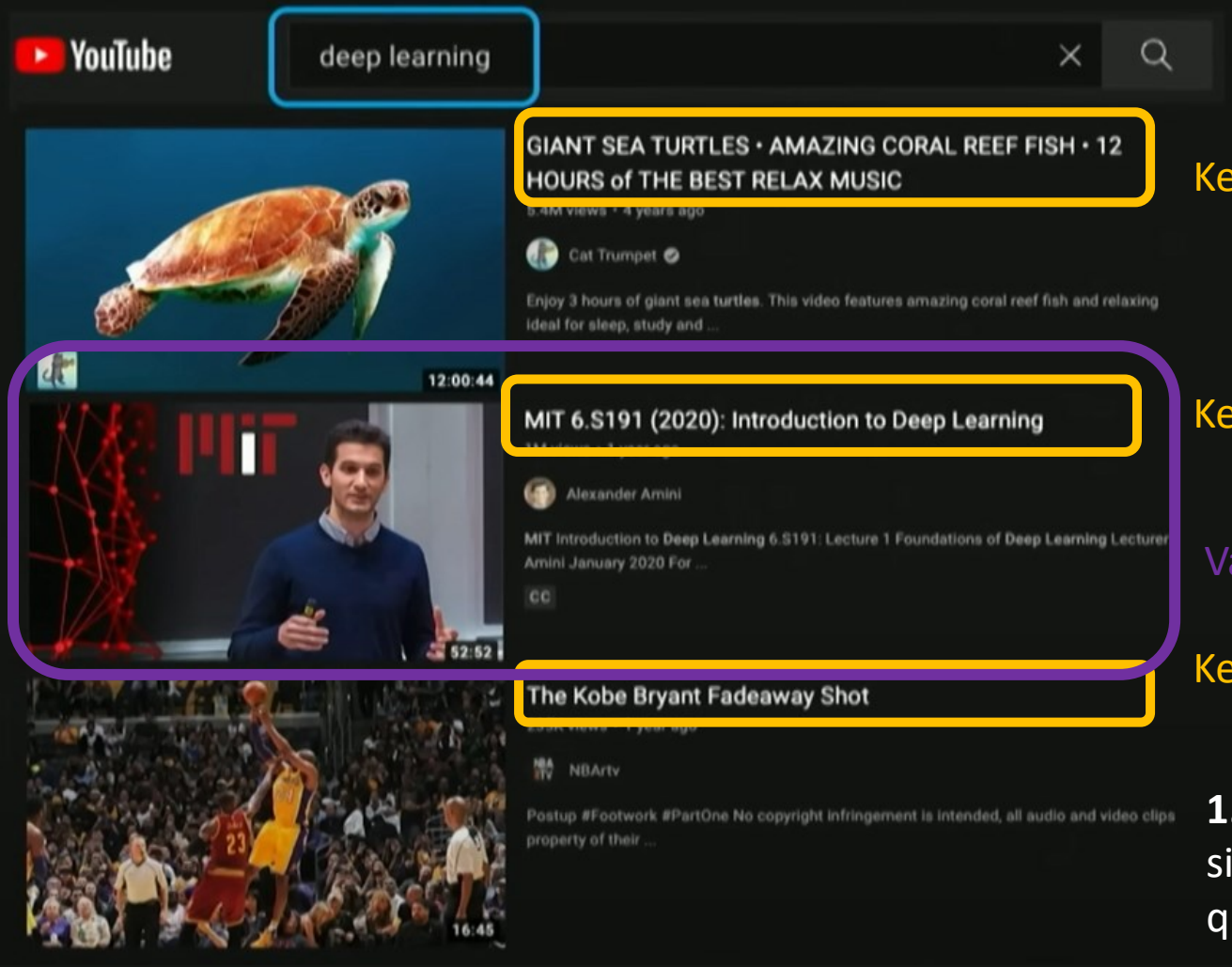
- No Recurrence
- Not scalable
- No order



Attention, the core of transformers

1. Identify which parts to attend to
2. Extract the features with higher attention

What is attention?



Query (Q)

Key (K_1)

Key (K_2)

Value (V)

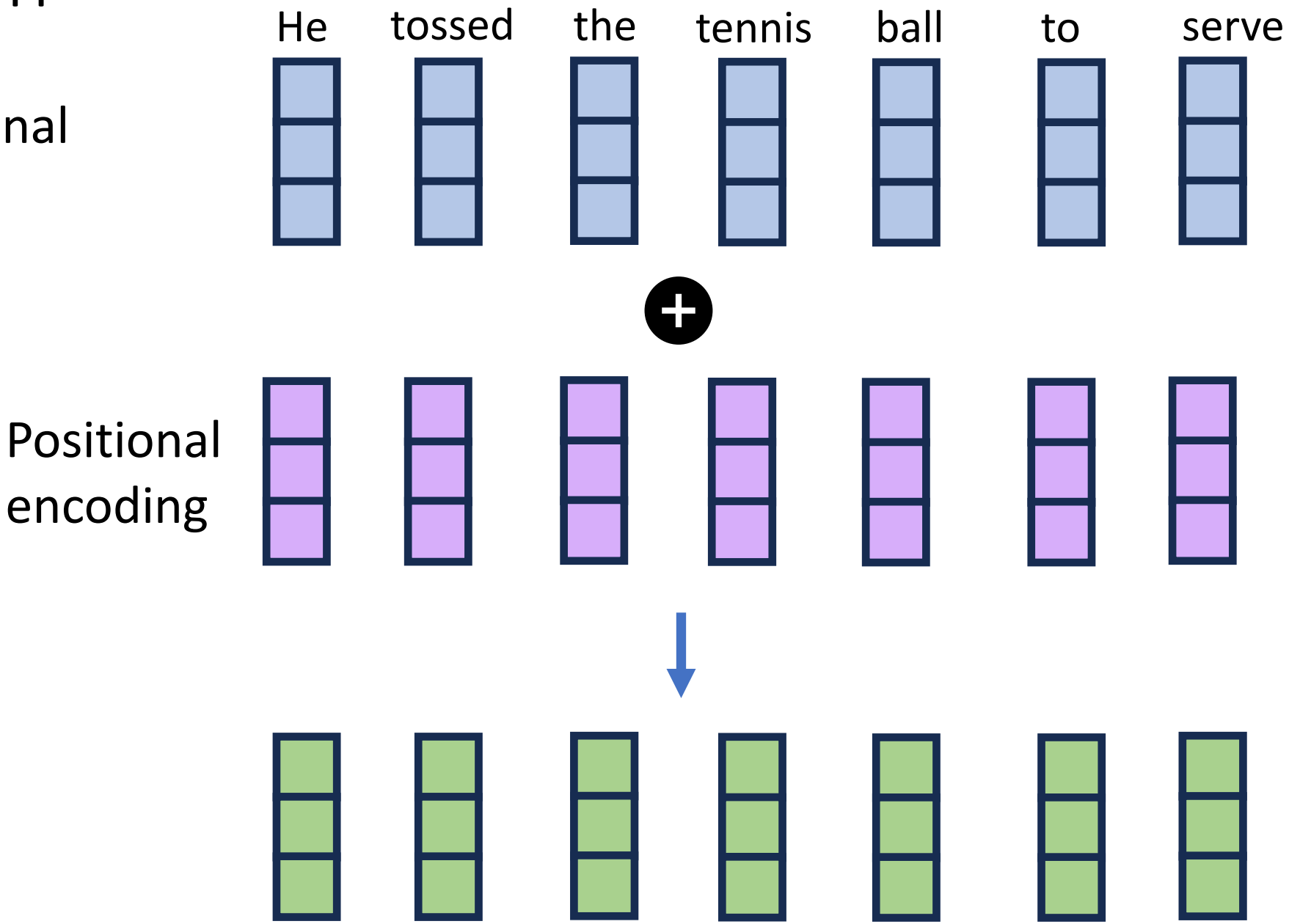
Key (K_3)

1.Attention mask: How similar each key is to the query?

2.Extract values based on attention: Return a combination of values according to the mask.

Self-attention

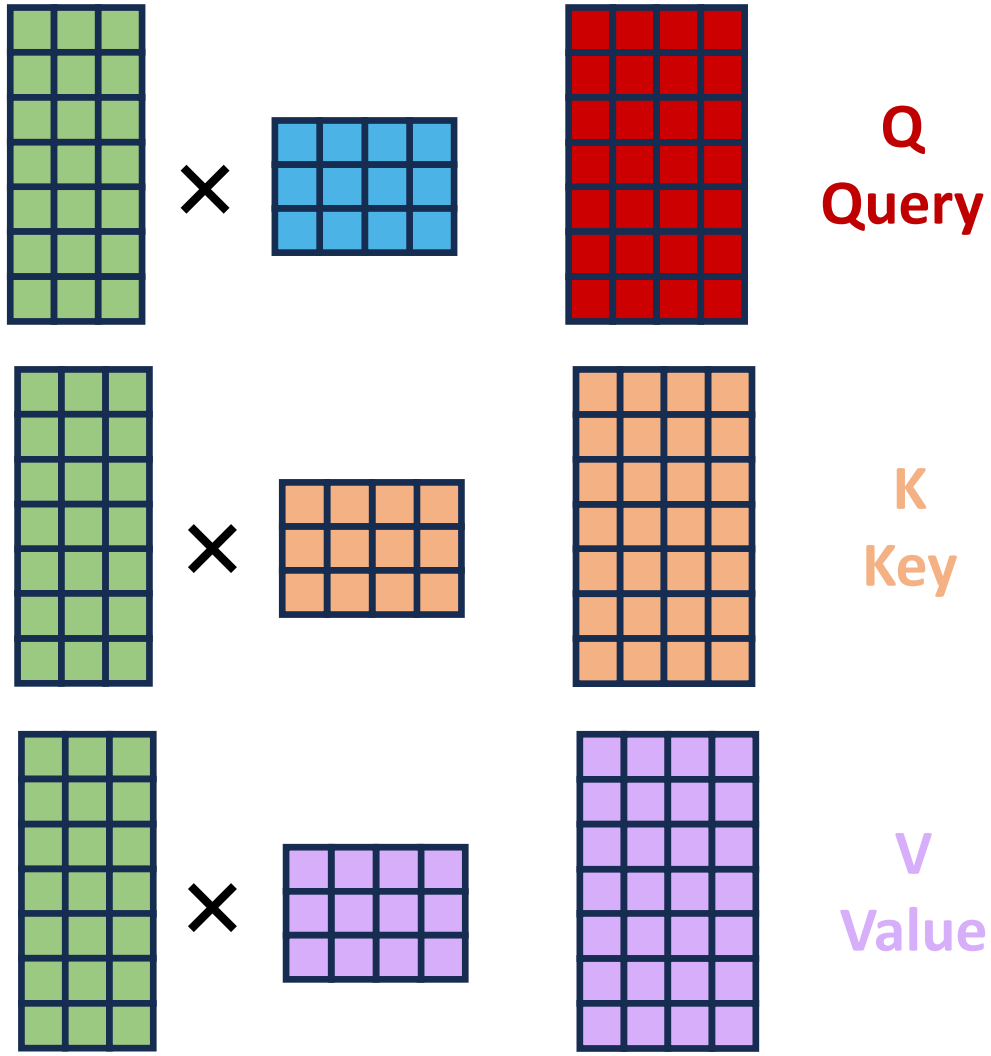
1- Encode positional information



Self-attention

1- Encode positional information

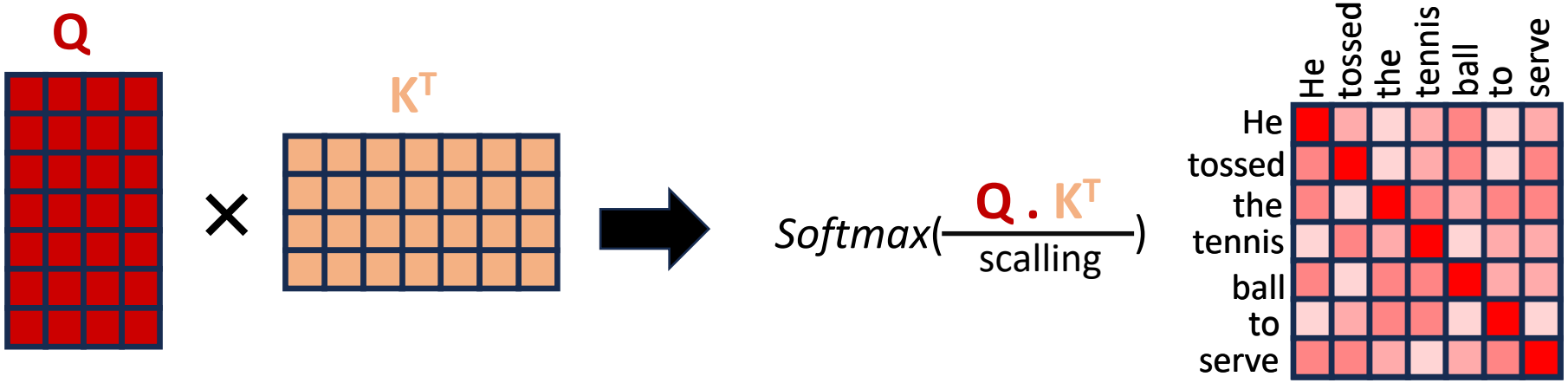
2-Extract **Key**, **Query** and **Value**



Self-attention

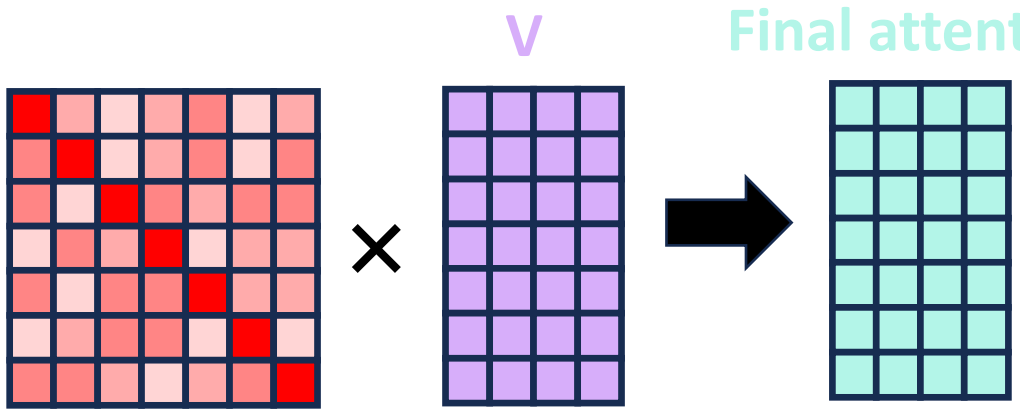
1- Encode positional information

2-Extract **Key**, **Query** and **Value**



3-Compute attention weights

4-Extract features with high attention



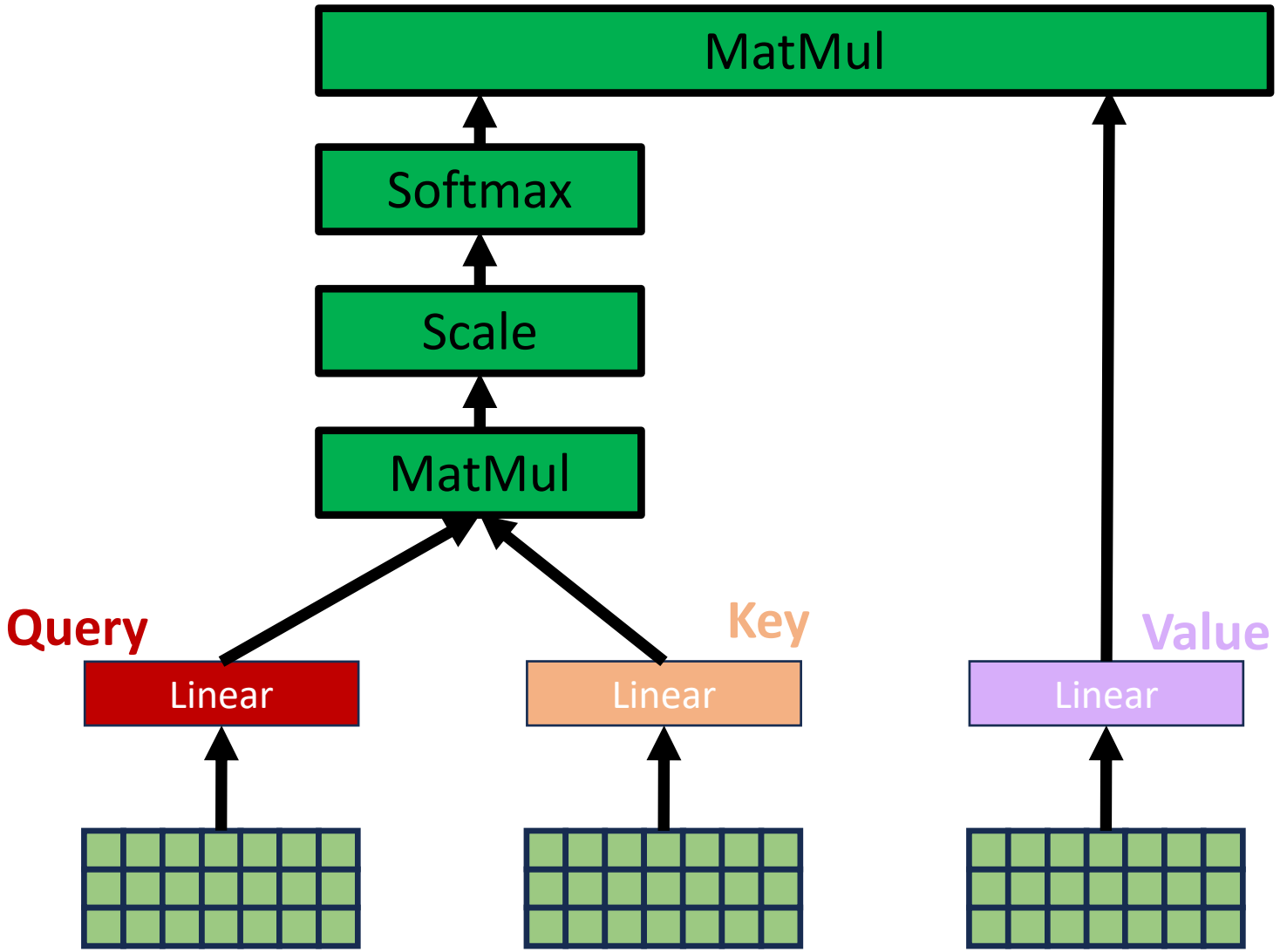
Recap

1- Encode positional information

2-Extract **Key**, **Query** and **Value**

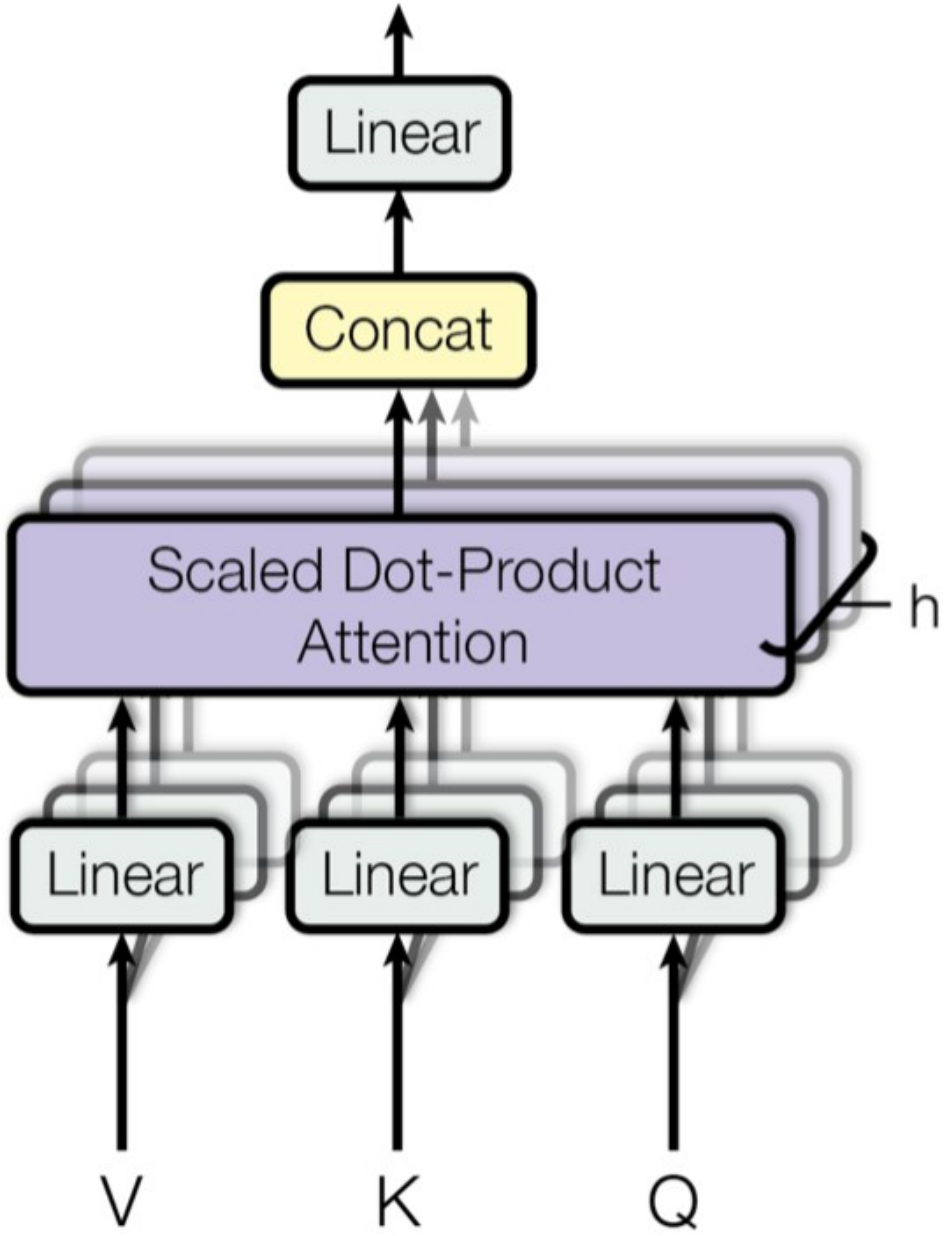
3-Compute attention weights

4-Extract features with high attention



Multiheaded attention

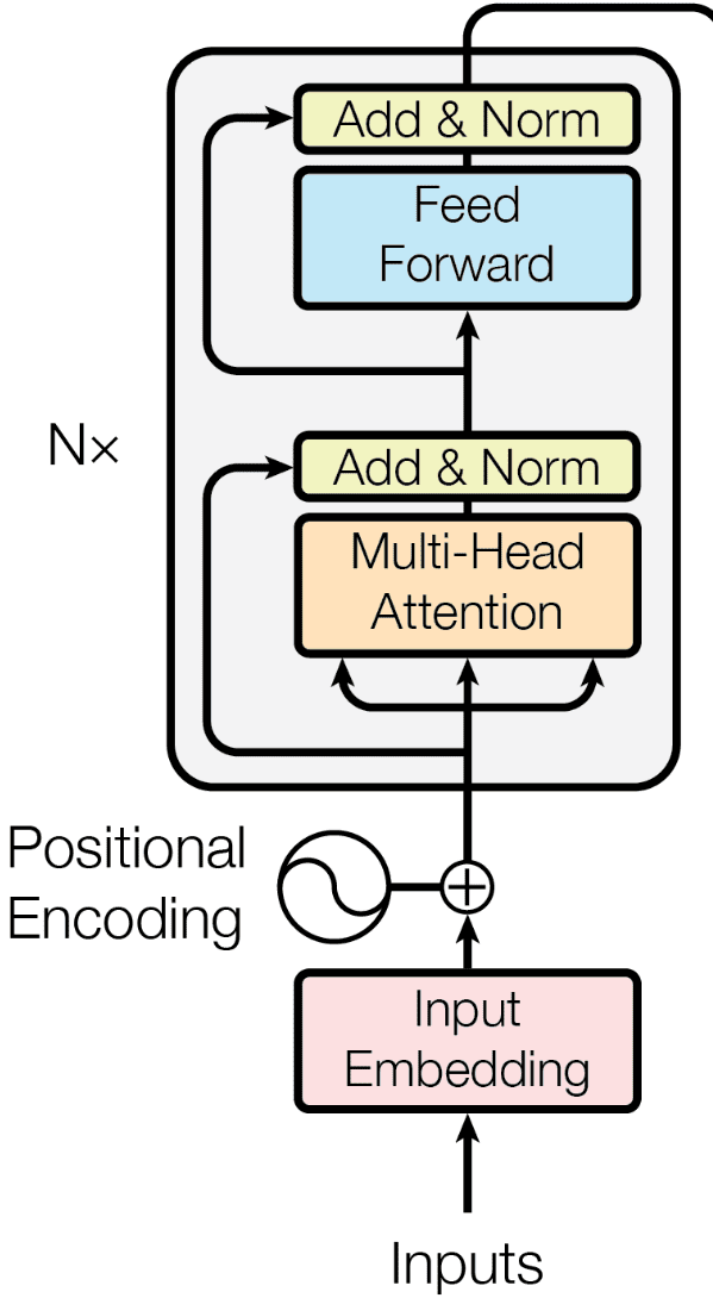
- Multiple parallel attention computation
- Helps with extracting diverse features



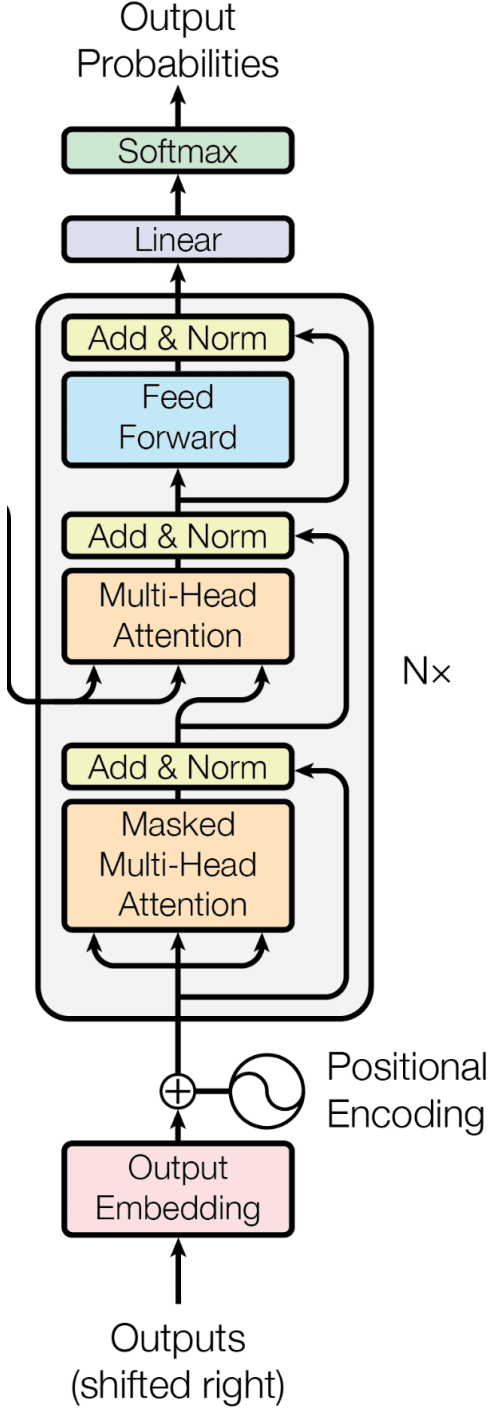
Encoder bigger picture

x + atten(x) is important for:

- Gradient flow
- Preserving information



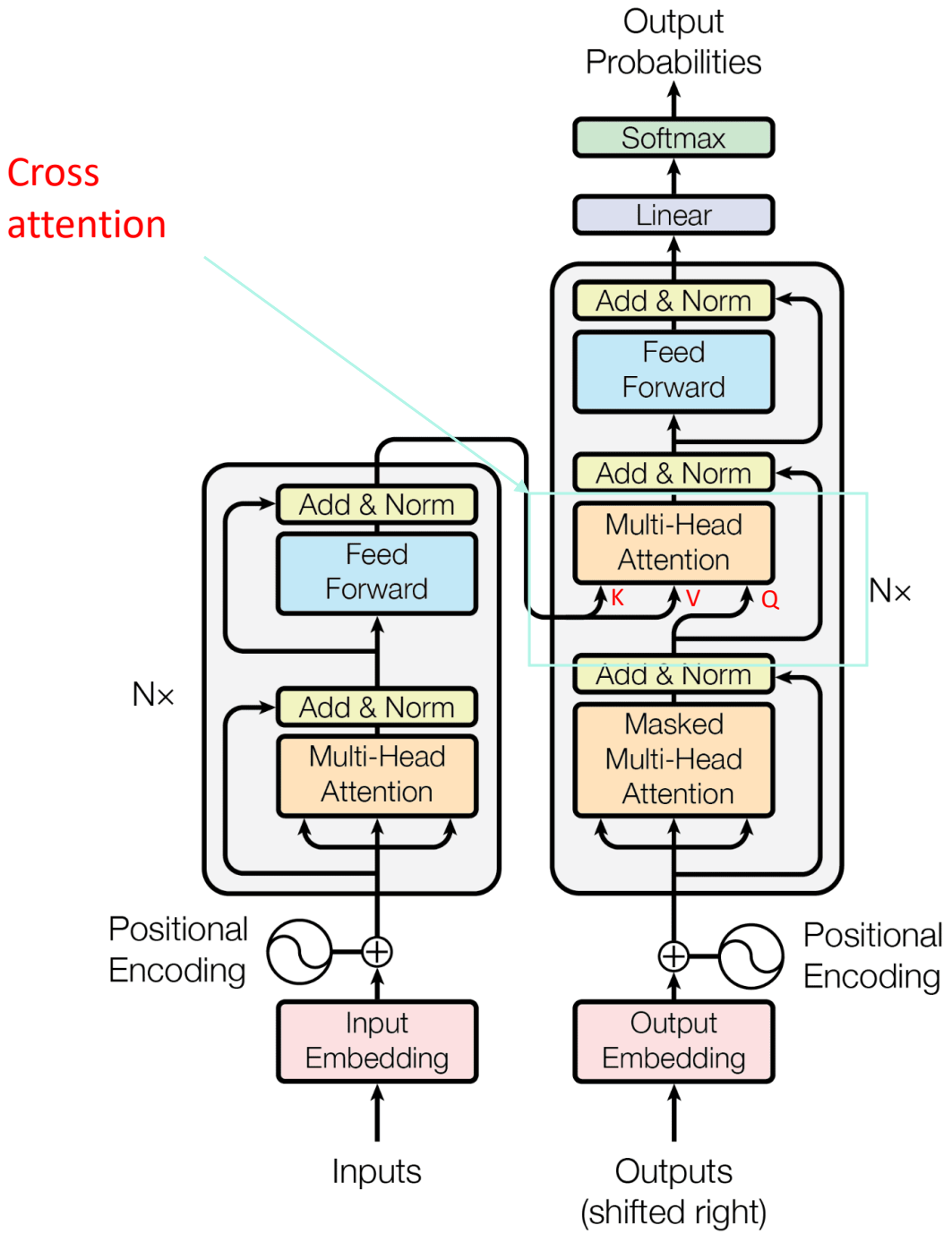
Decoder bigger picture



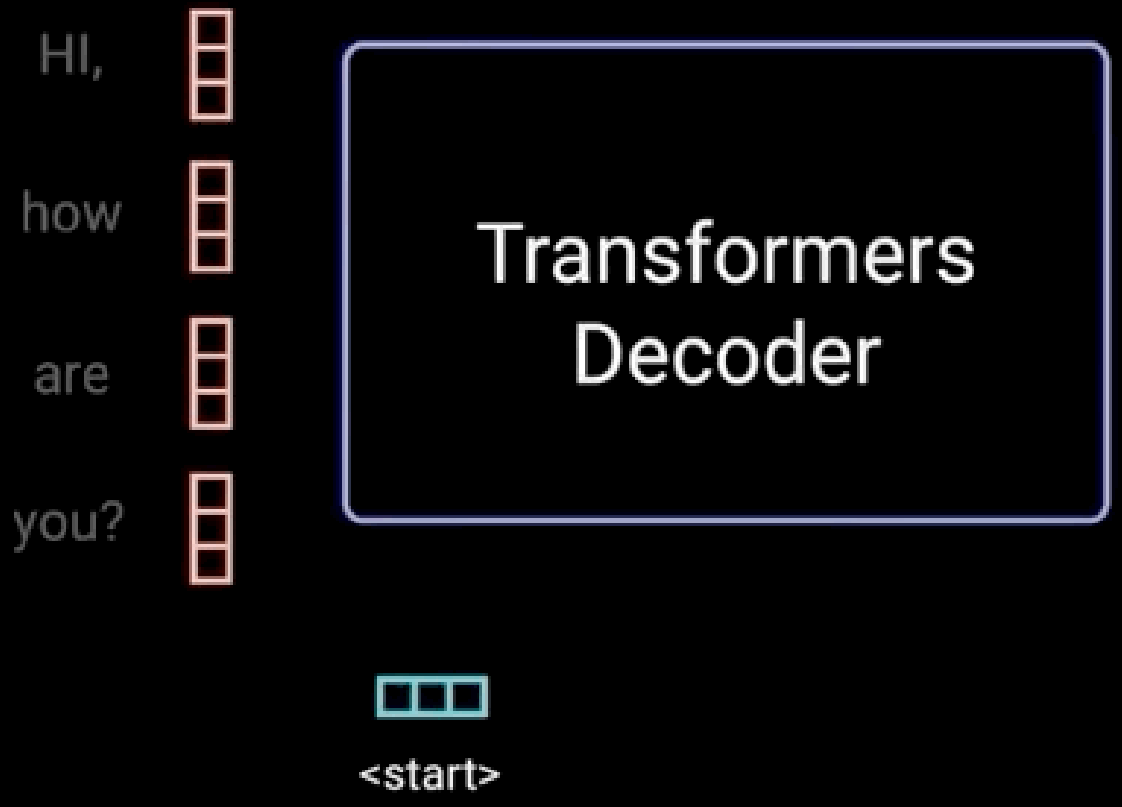
Transformers

Q -> Input

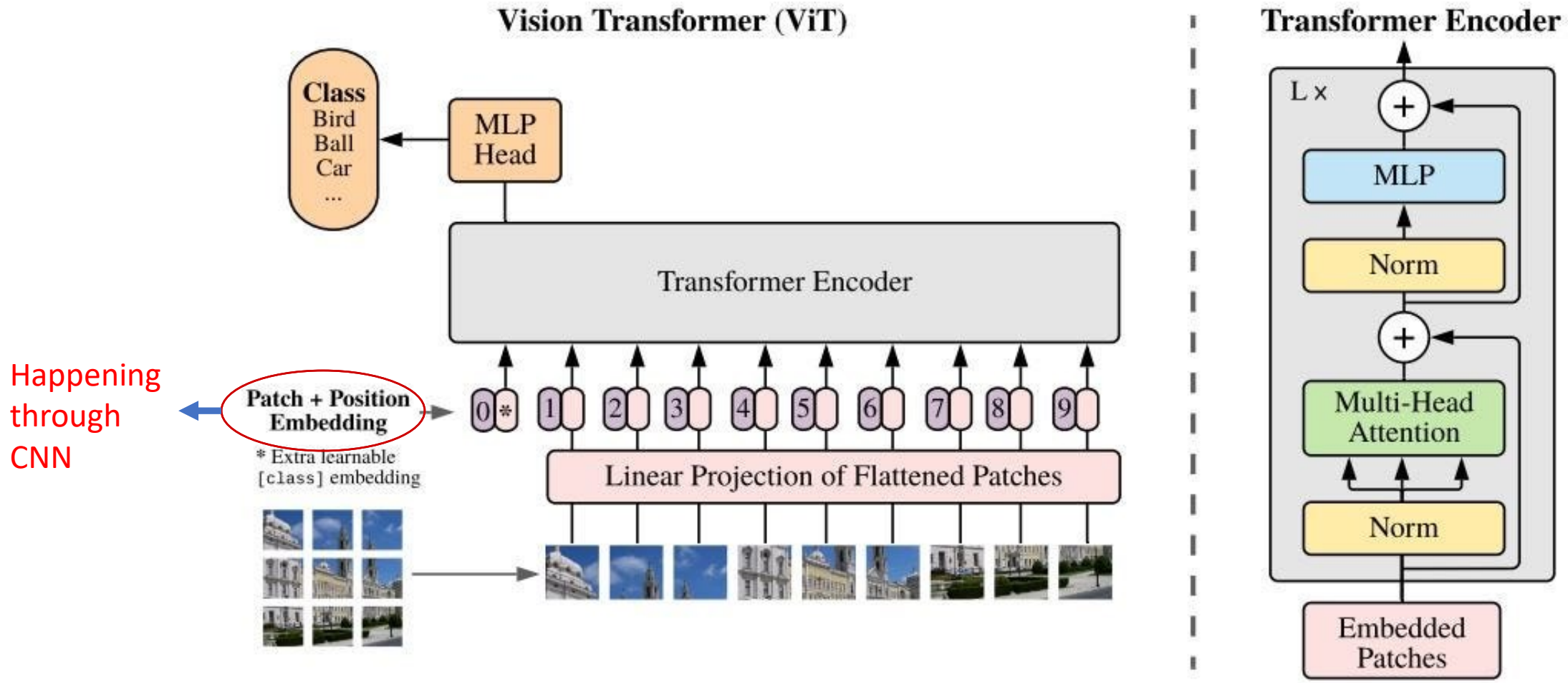
K, V -> Conditional information



Transformers Decoder

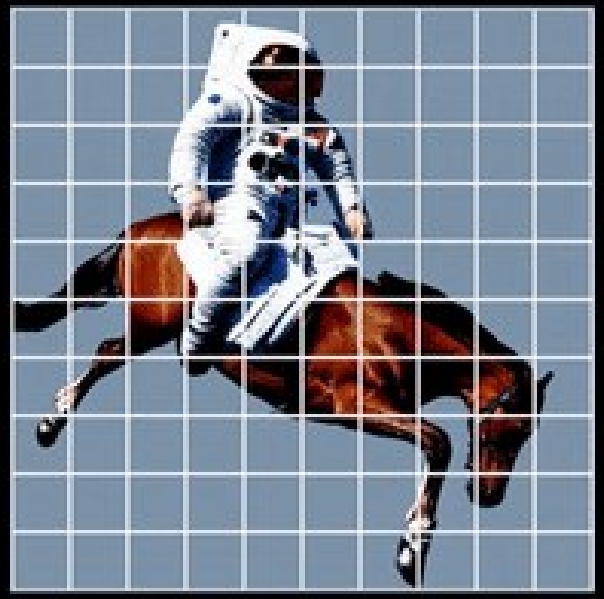


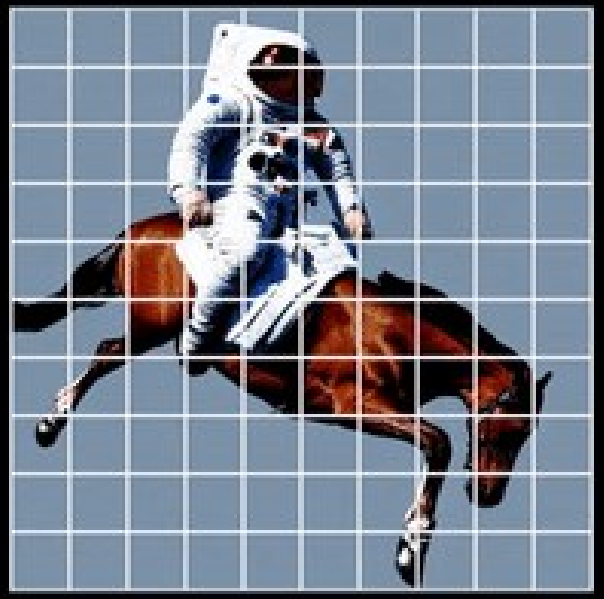
Vision transformers



Happening through CNN

Vision transformers





Summary

- How attention mechanism solve challenges of RNNs.
- What are key, query and value mean in attention and how to calculate attention masks.
- Encoder and decoder.
- Vision transformers.

Useful resources

- https://www.youtube.com/watch?v=ySEx_Bqxvvo&ab_channel=AlexanderAmini
- https://www.youtube.com/watch?v=kCc8FmEb1nY&pp=ygUIa2FycGF0aHk%3D
- https://www.youtube.com/watch?v=OyFJWRnt_AY&t=3634s&pp=ygUOcGFzY2FsIHBvdXBhcnQ%3D