

Deep Sequence Modelling

Transformers

Transfer Learning

Elli Valla

PhD student and junior researcher at

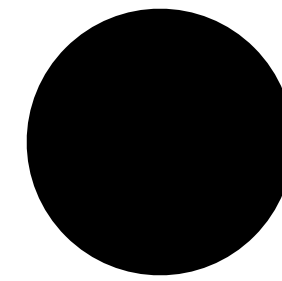
Department of Software Science
TalTech University

elli.valla@taltech.ee

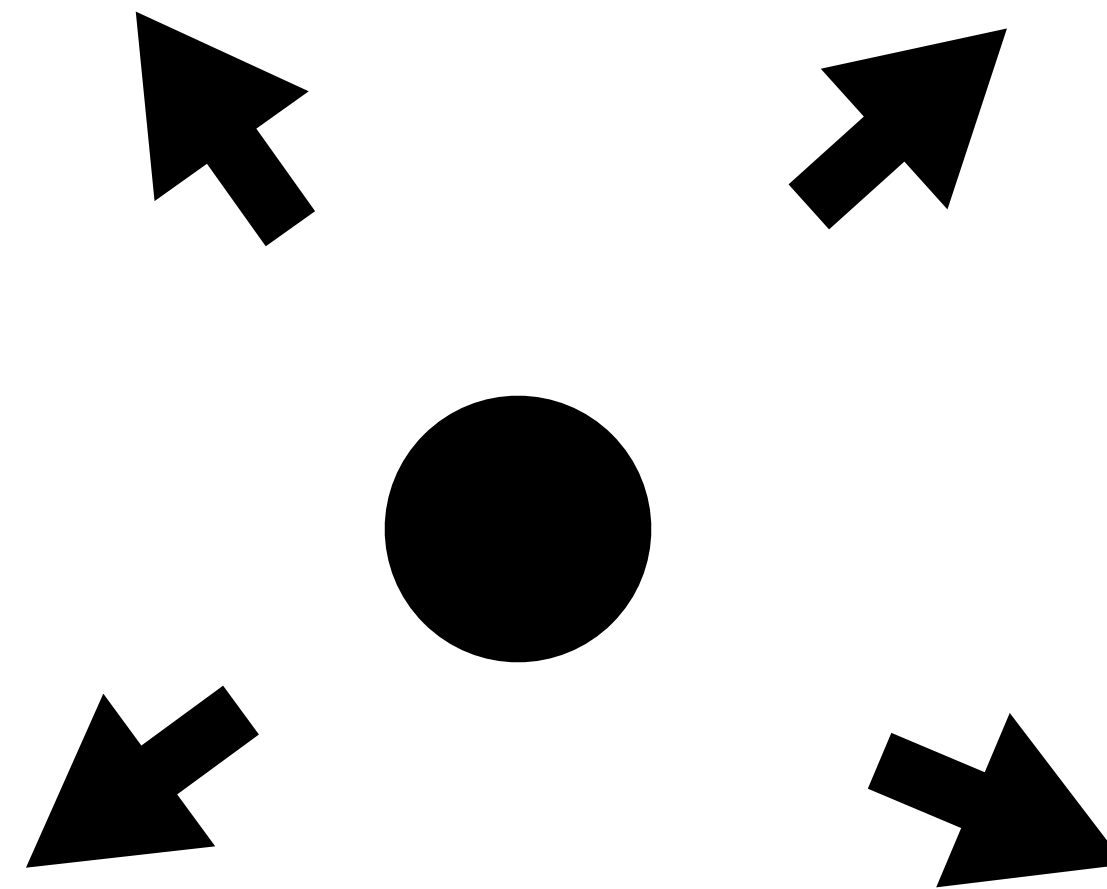
Plan for today:

1. Recurrent Neural Networks
2. Transformers
 - 2.1 Positional Encoding
 - 2.2 Self-attention
3. Pre-trained Models (transfer learning)

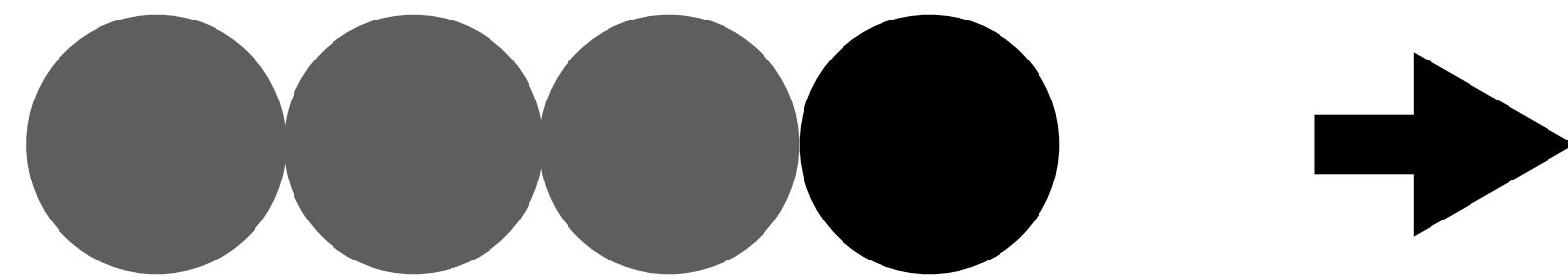
Where does the ball go next?



Where does the ball go next?



Where does the ball go next?



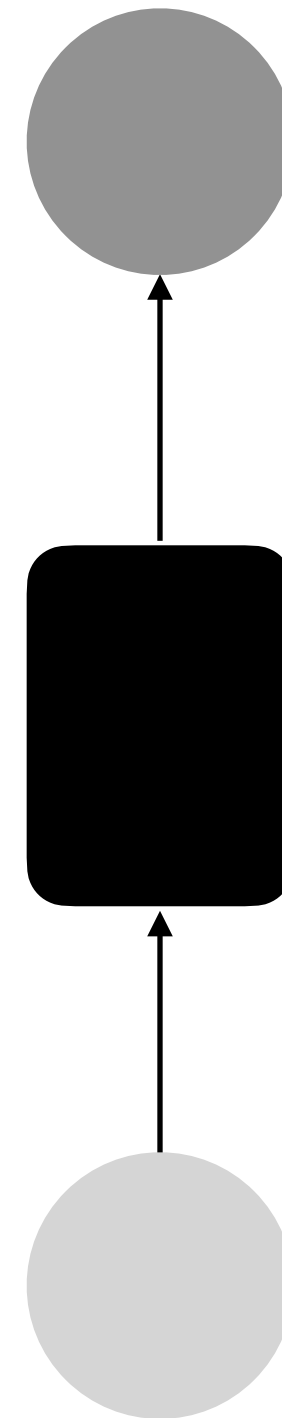
Sequential data is all around us



EXXT.DE iShares NASDAQ-100 (DE)		120,64 +0,60
S&P 500 S&P 500		4 175 -120,92
AAPL Apple Inc.		156,80 -6,08

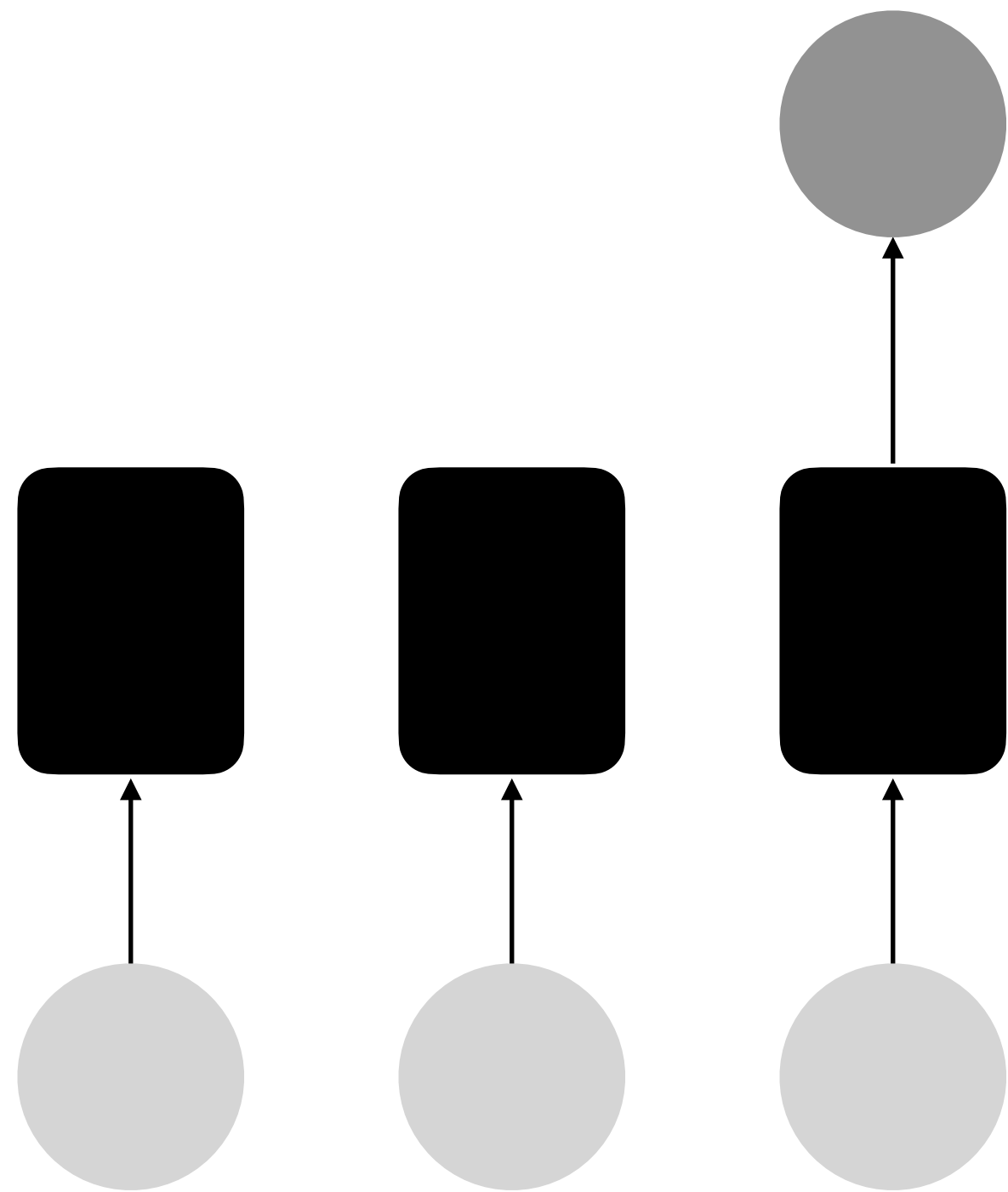
Sequential data is all around us

Sequence Modelling Types



Binary classification

Sequence Modelling Types



Sentiment classification

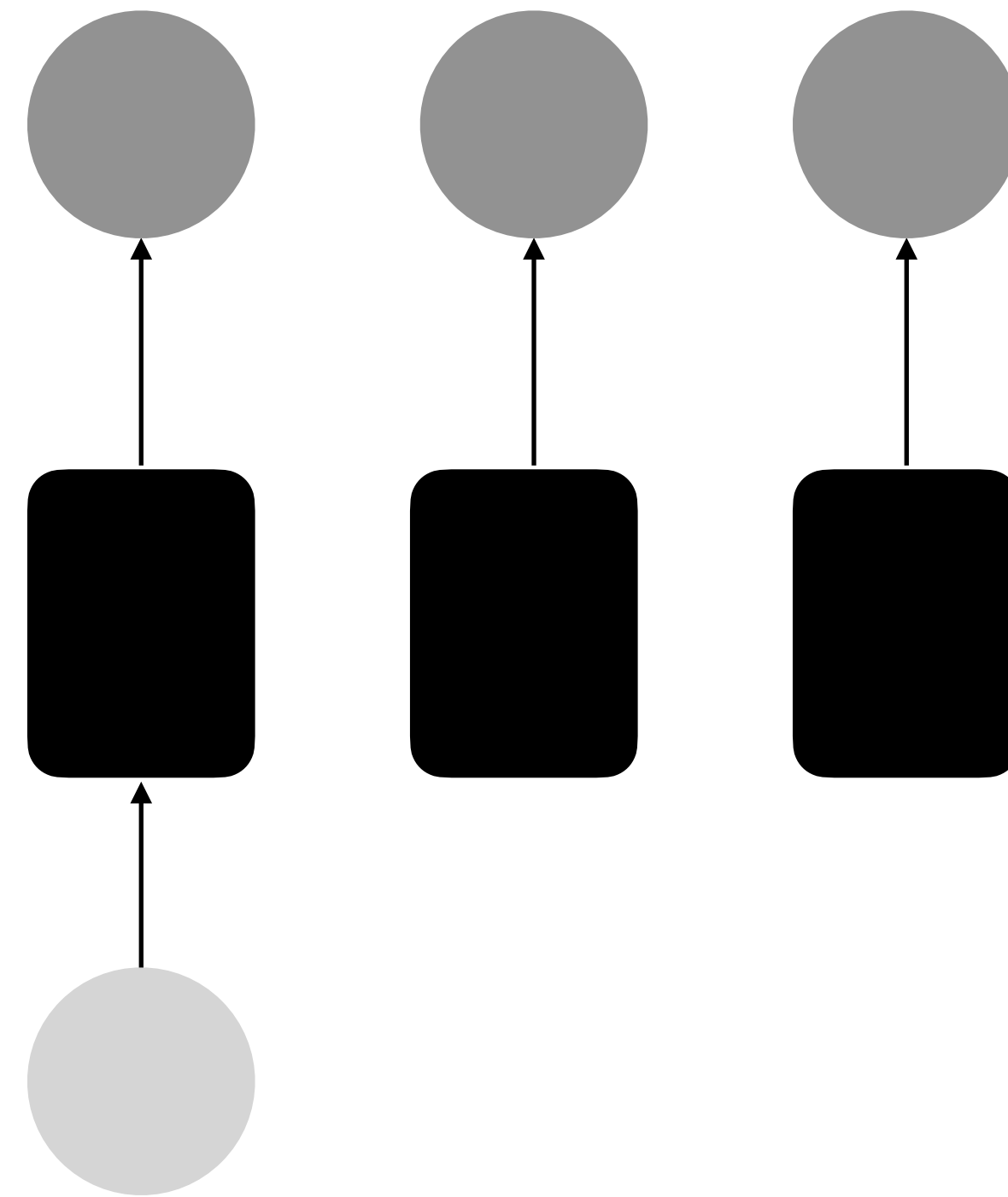
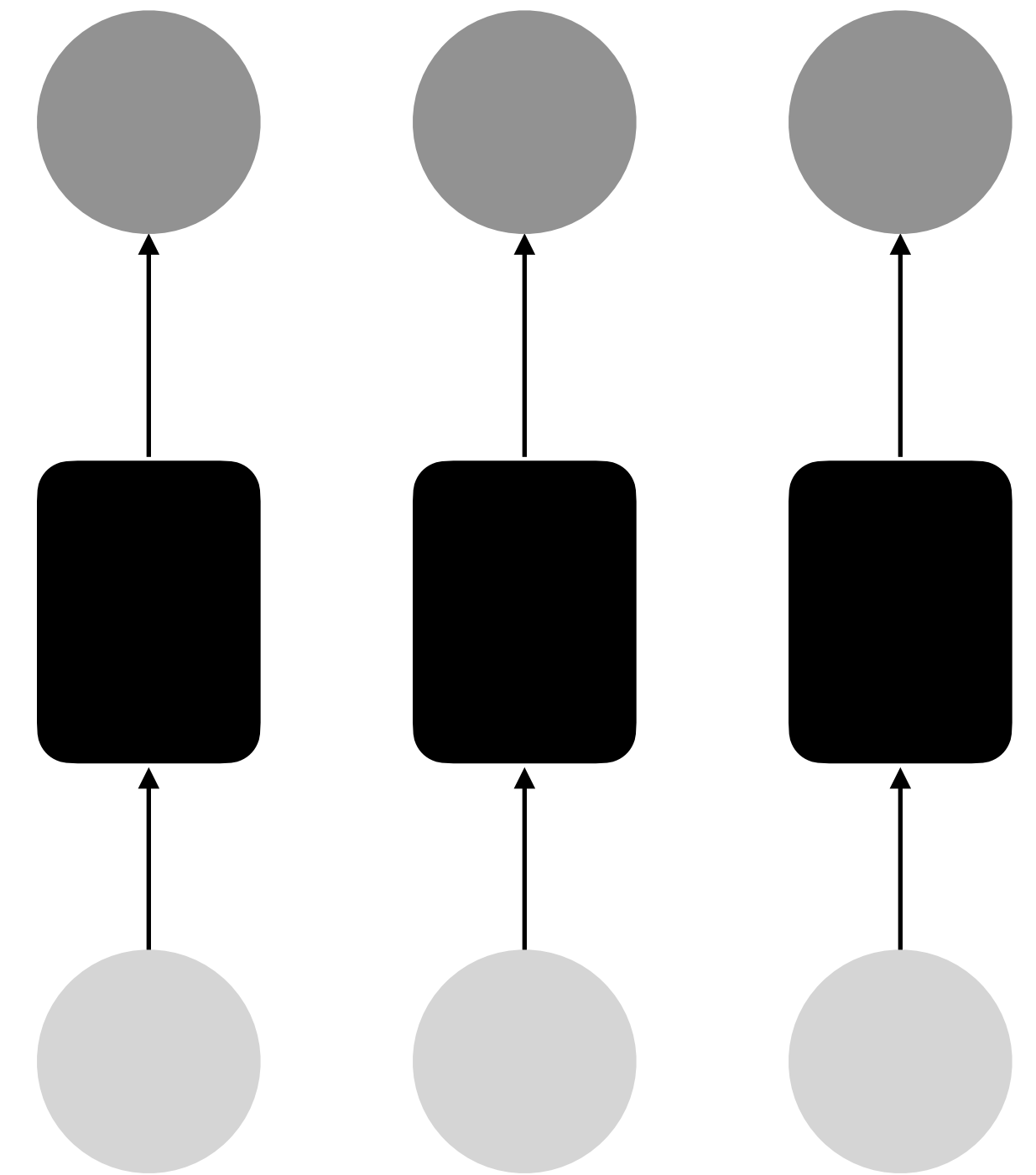
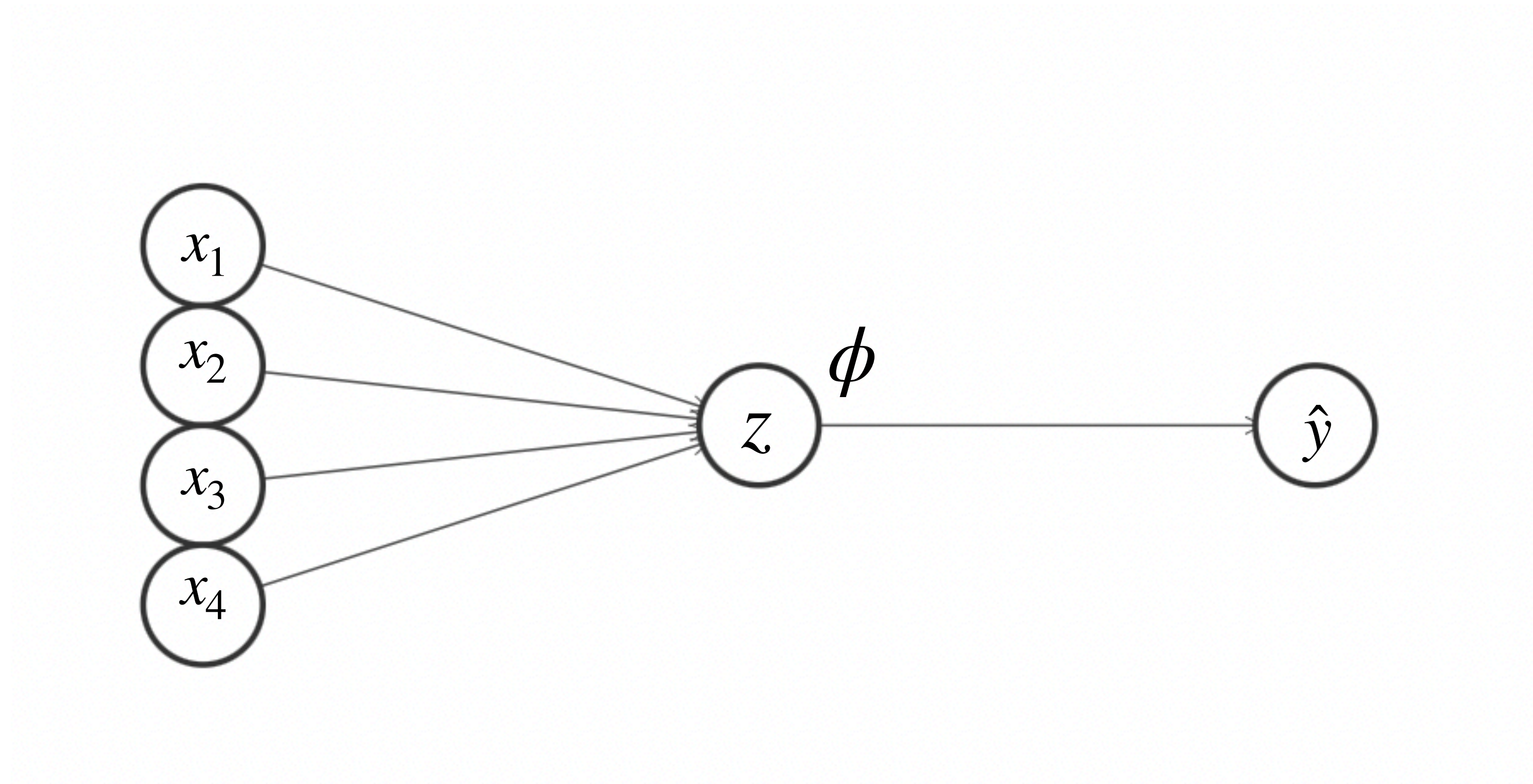


Image captioning

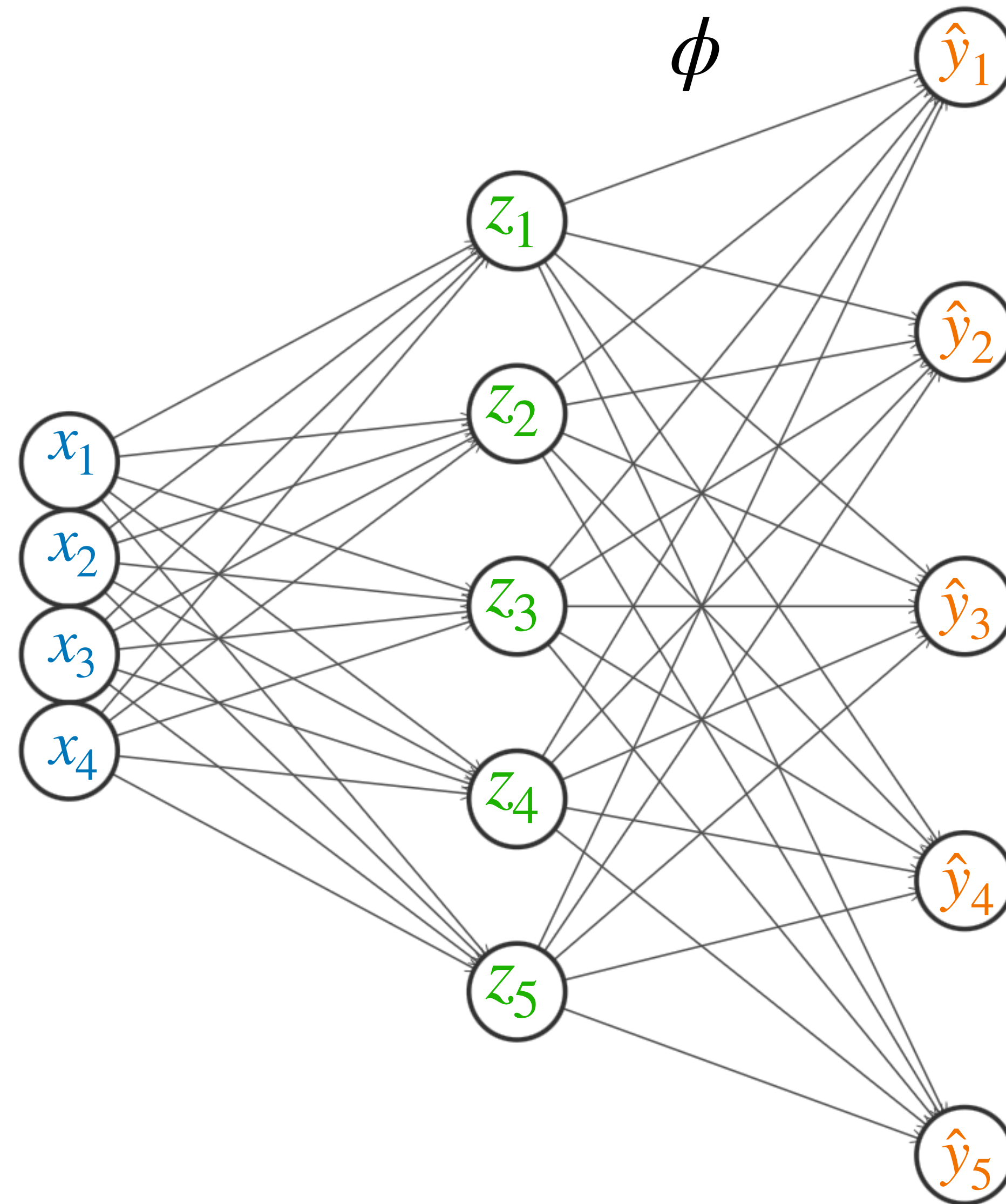


Machine translation

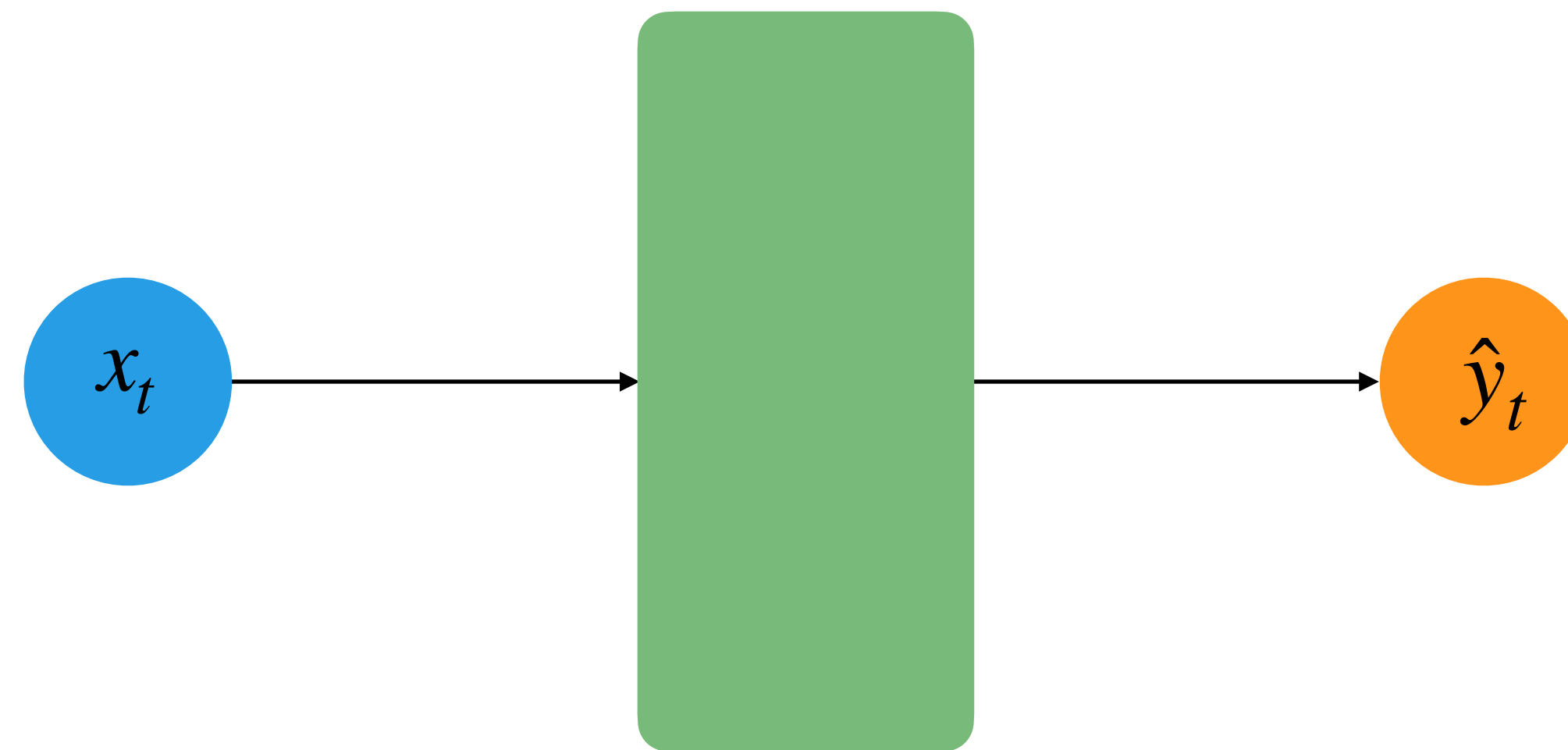
The Perceptron



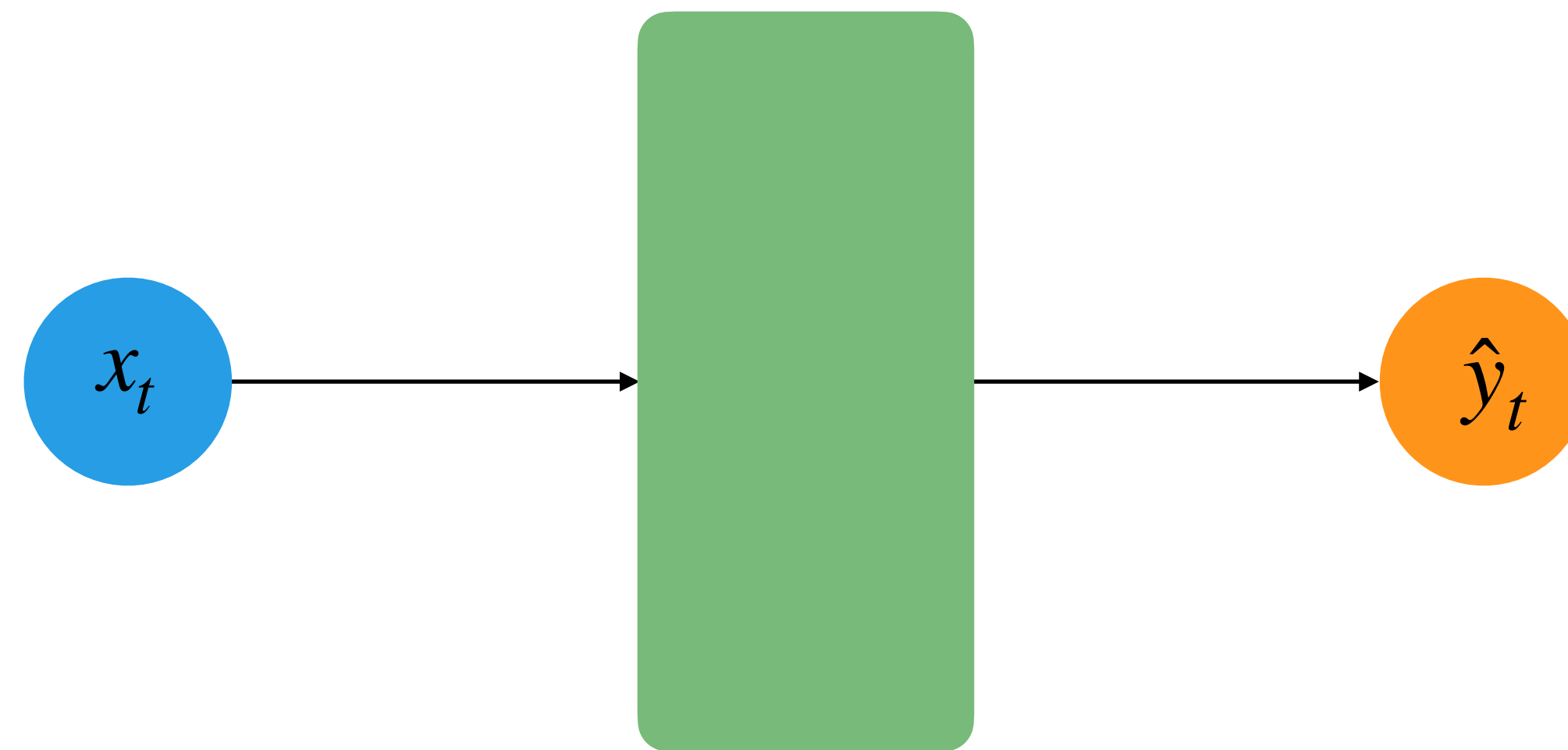
Feed Forward Neural Network



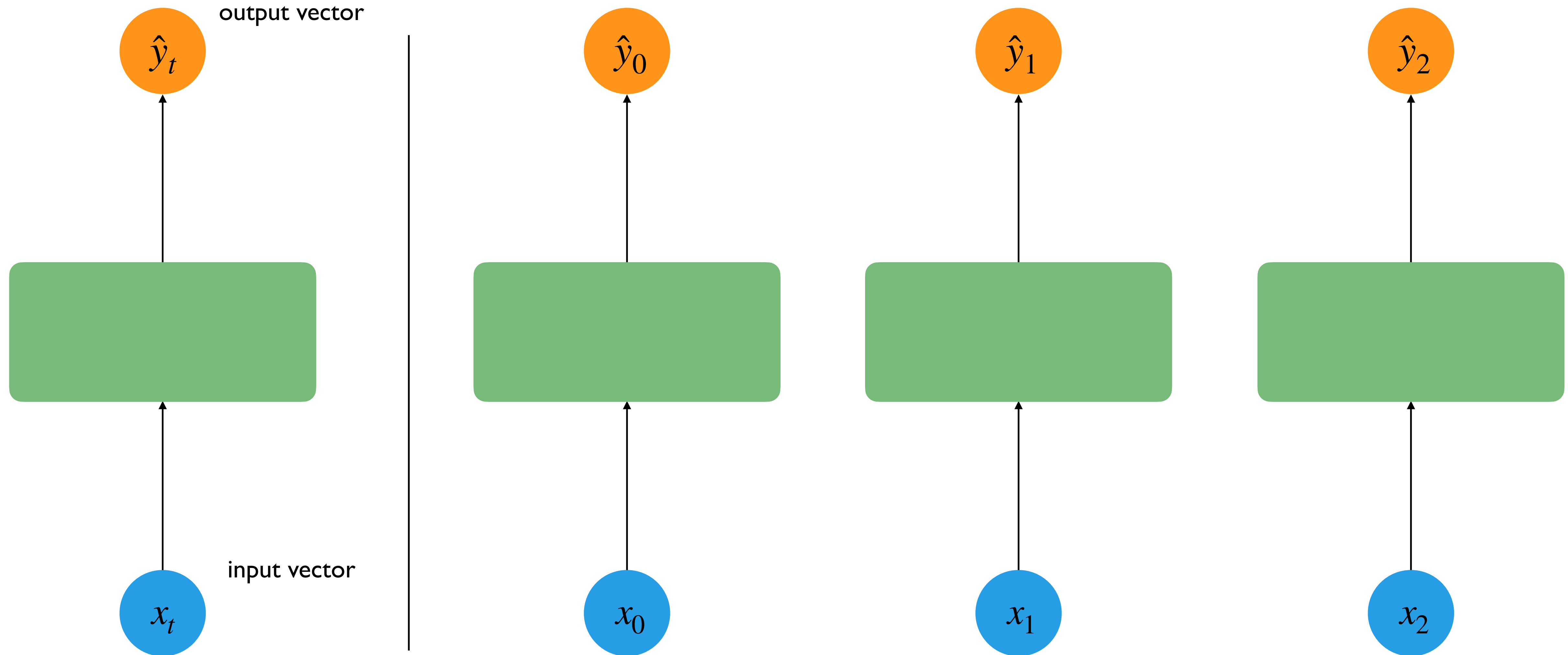
Adding a notion of time

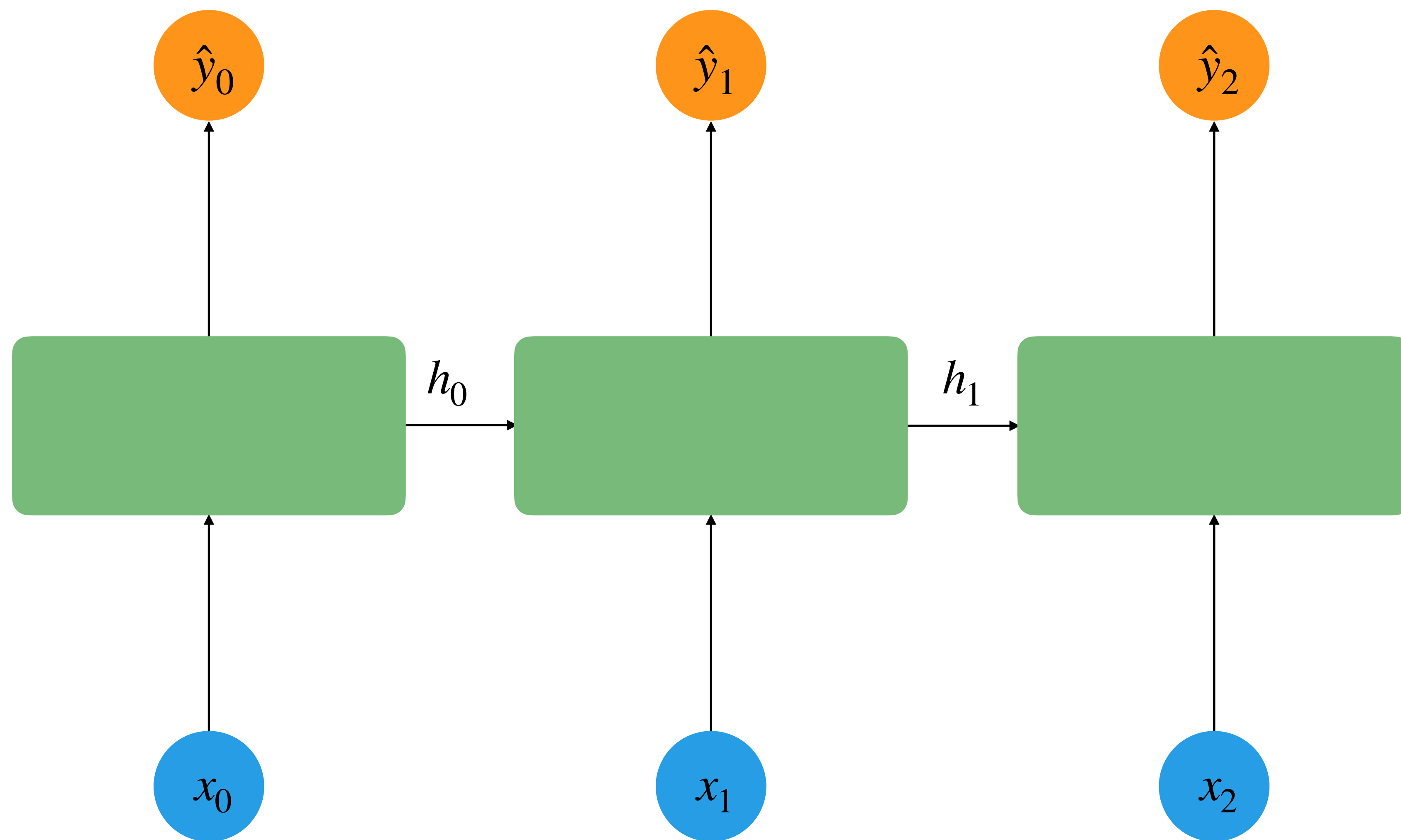
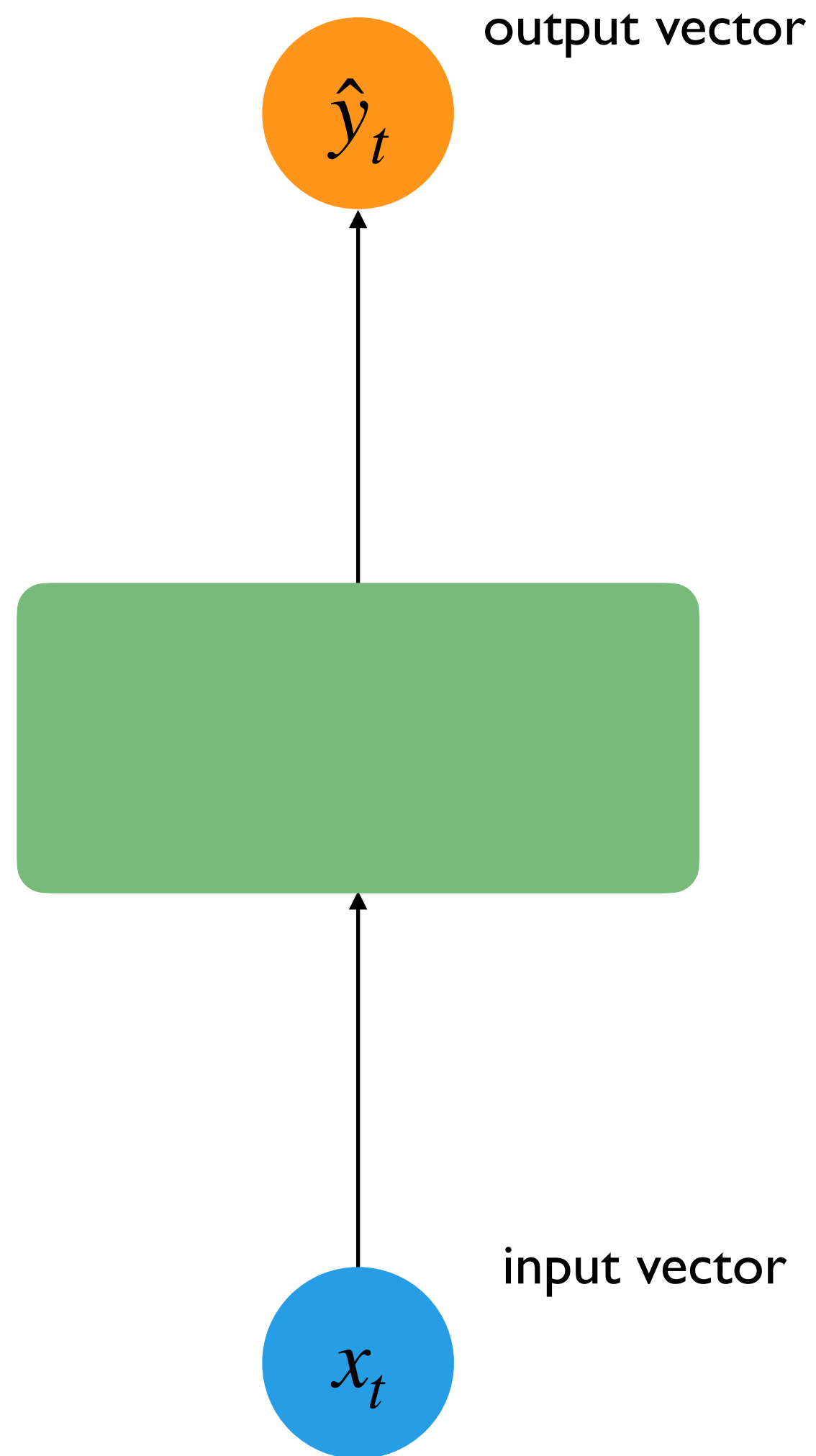


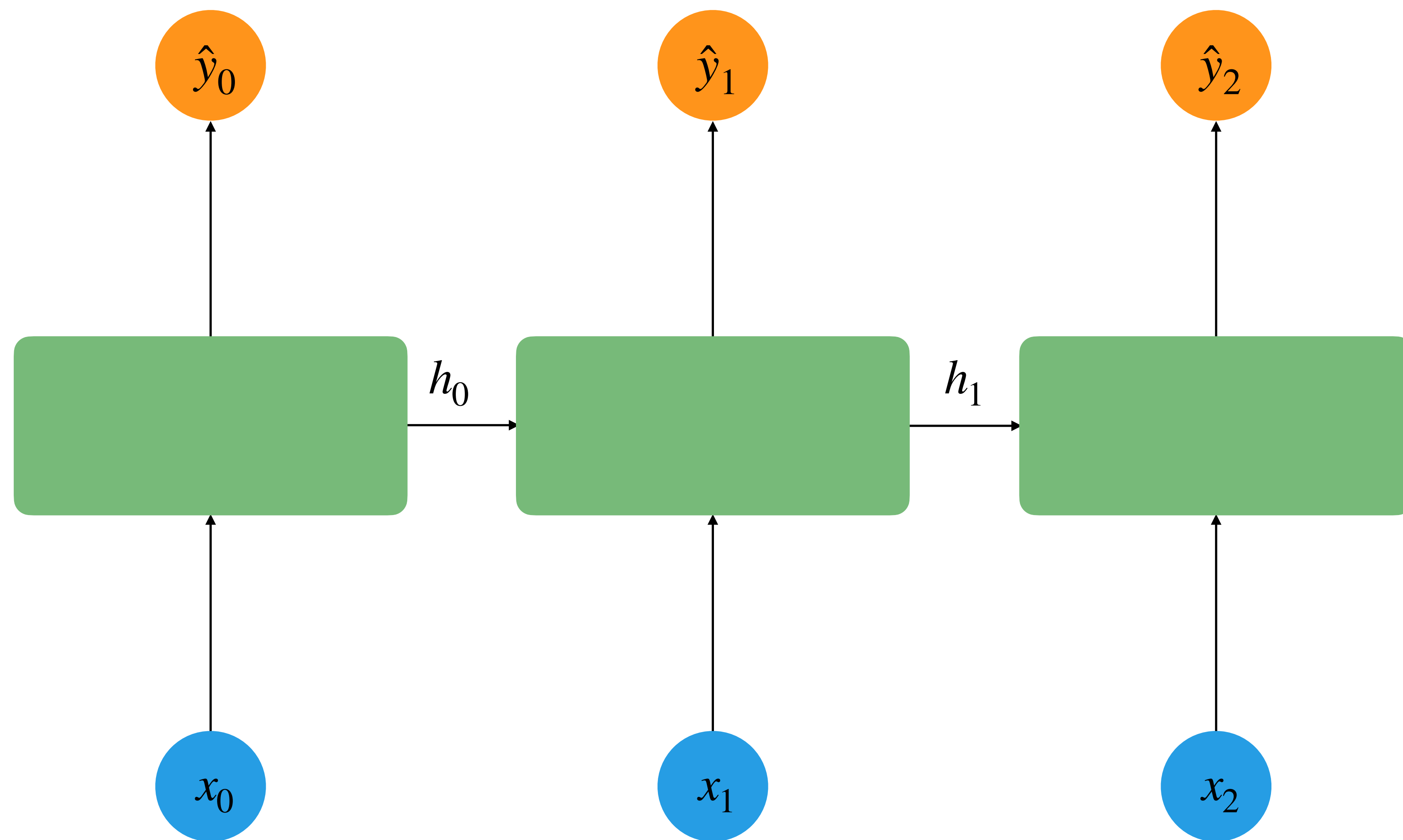
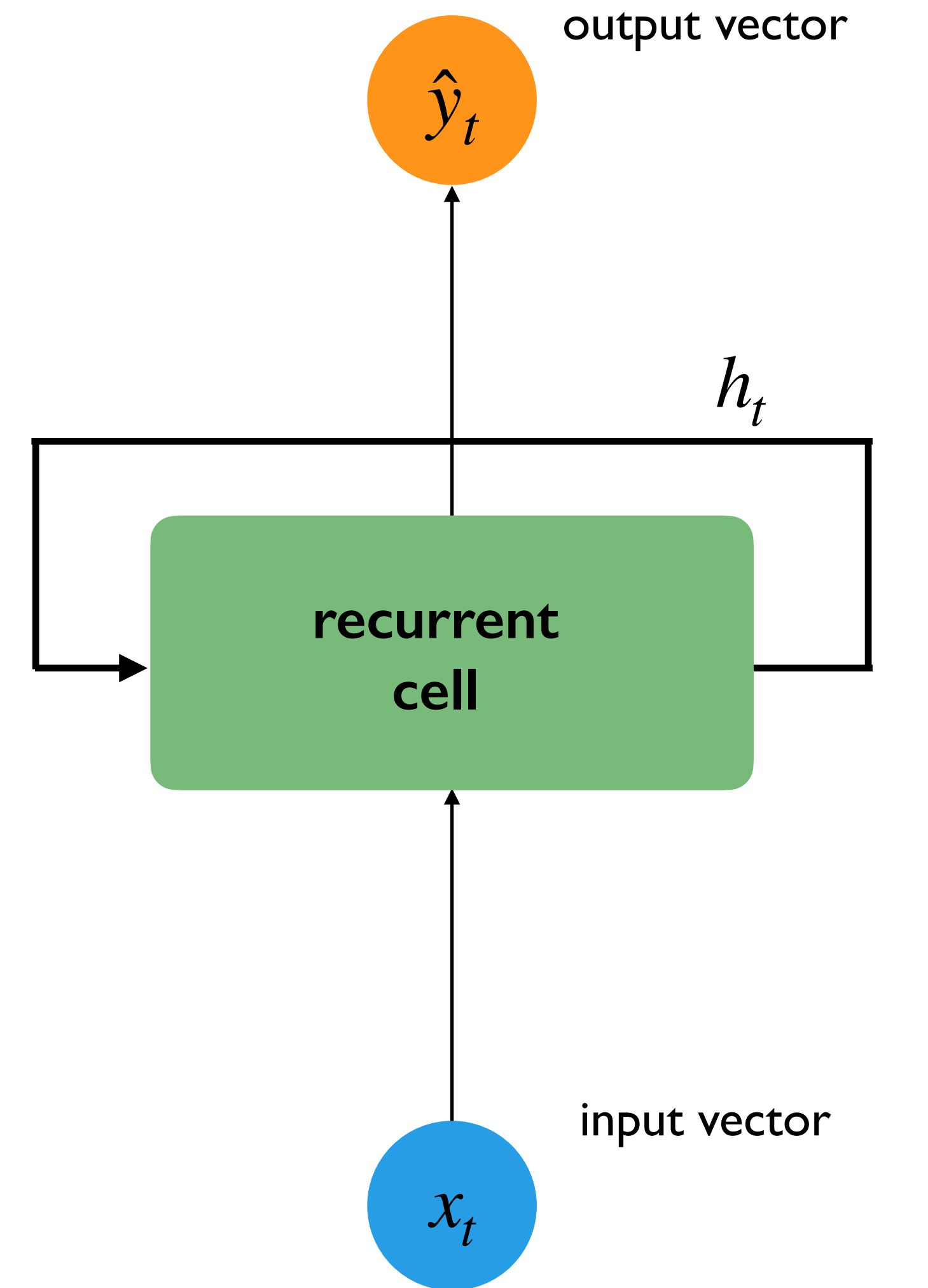
Adding a notion of time



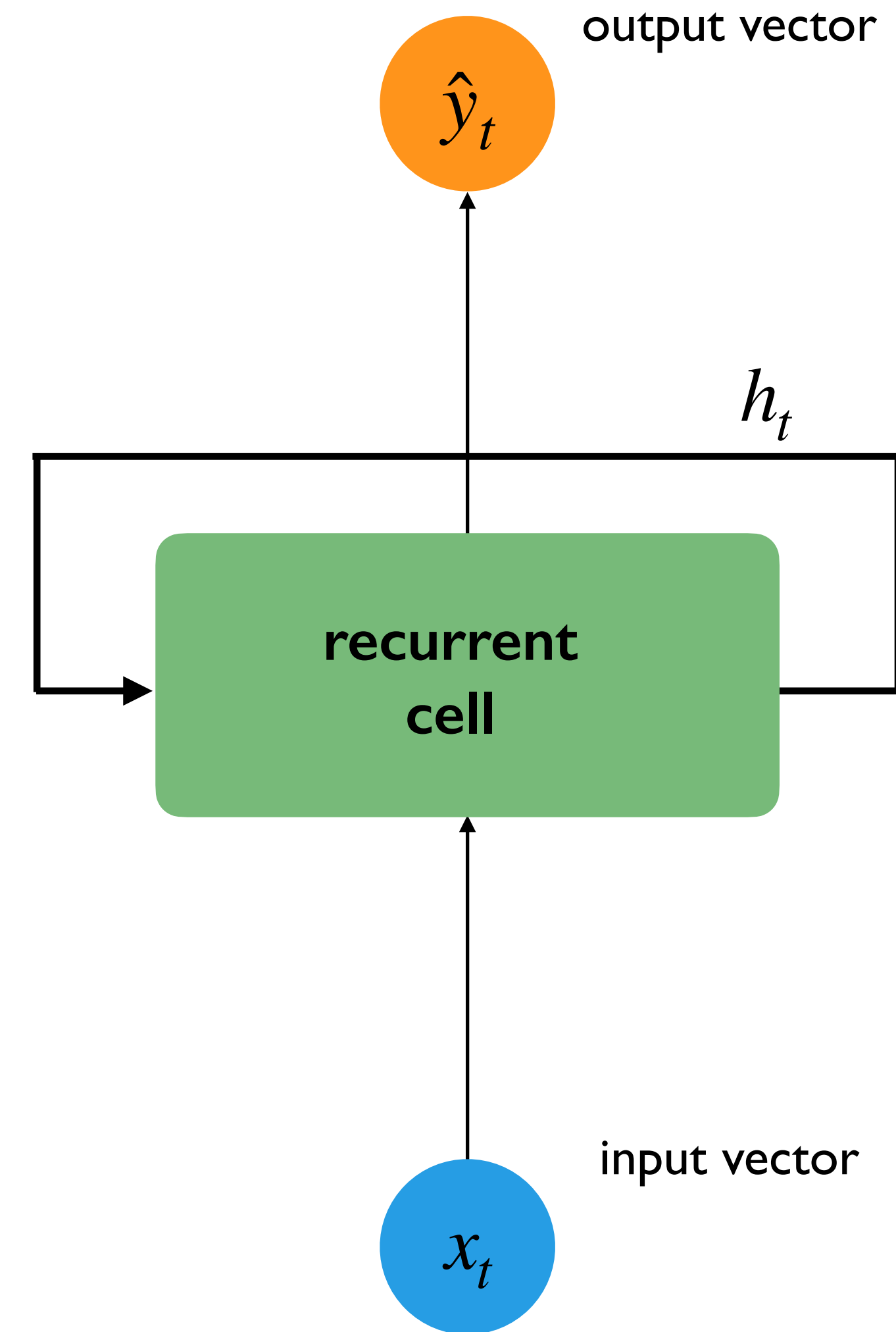
How can we capture inter-dependence?







Recurrent Neural Networks



cell state function with weights input old state

$$h_t = f_w(x_t, h_{t-1})$$

Output Vector

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

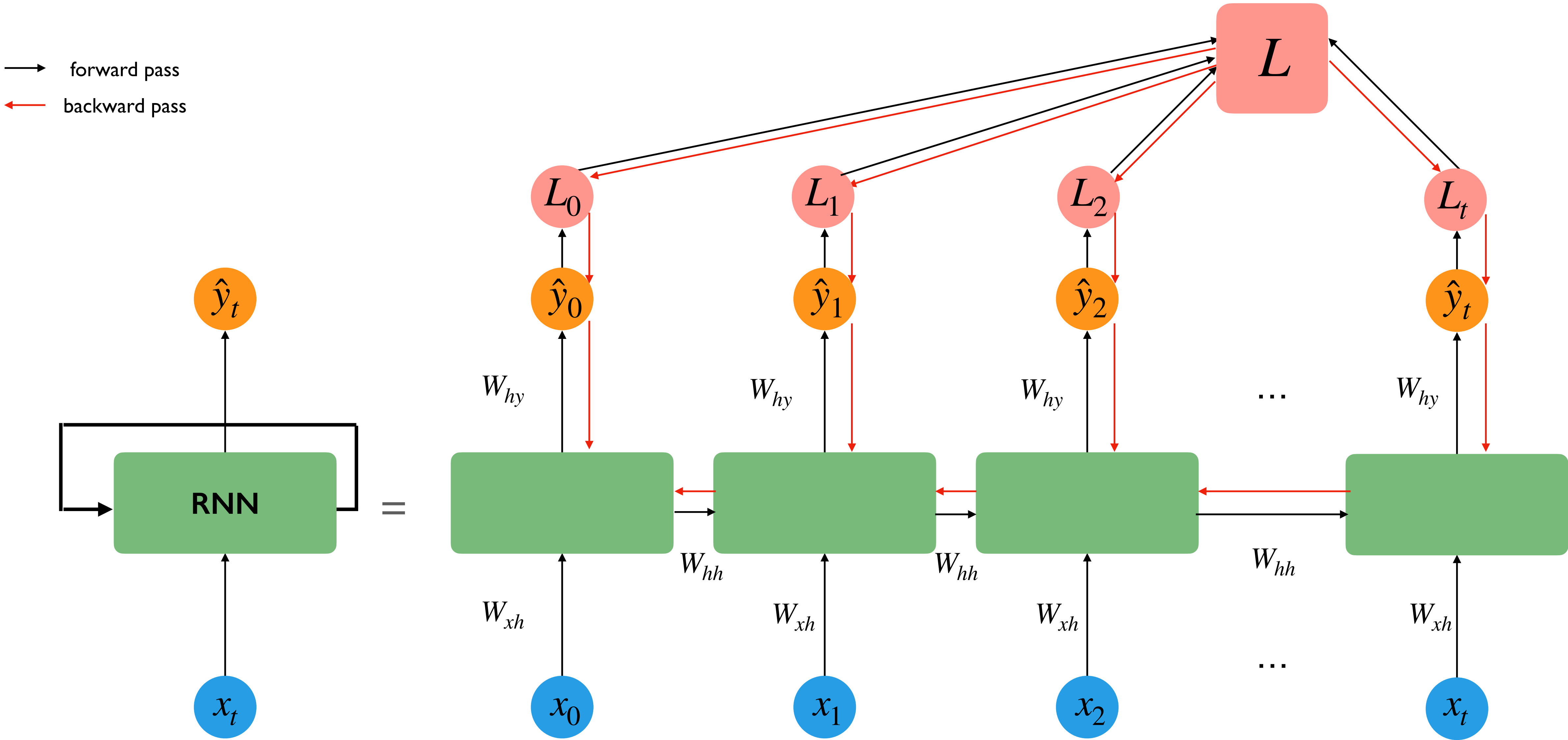
Update Hidden State

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

Input Vector

x_t

Backpropagation Through Time (BPTT)



RNN Limitations

A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of natural language processing (NLP) and computer vision (CV).

LSTM, GRU

No long-term memory.
Slow. No parallelisation.

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
lukaszkaizer@google.com

Ilia Polosukhin* ‡
illia.polosukhin@gmail.com

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

What part of the input should I focus on?

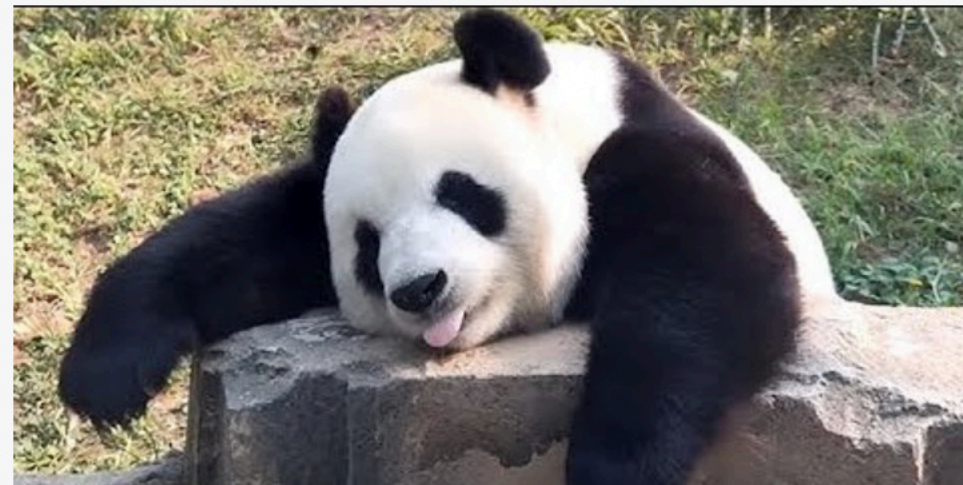


What part of the input should I focus on?





deep learning



Panda Funny Moment Videos Compilation


2.9M views • 1 year ago

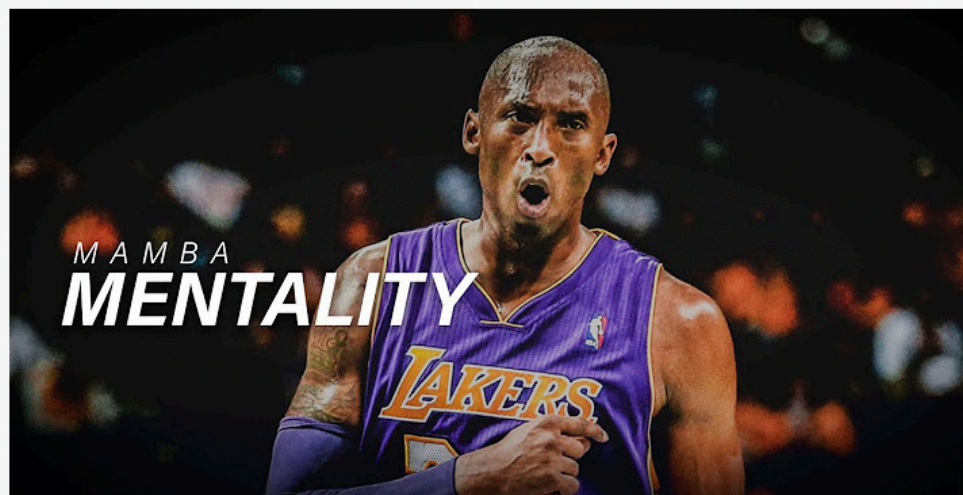
 Panda Story



MIT Introduction to Deep Learning | 6.S191


178K views • 1 month ago

 Alexander Amini



Mamba Mentality - Kobe Bryant (Motivational Video)

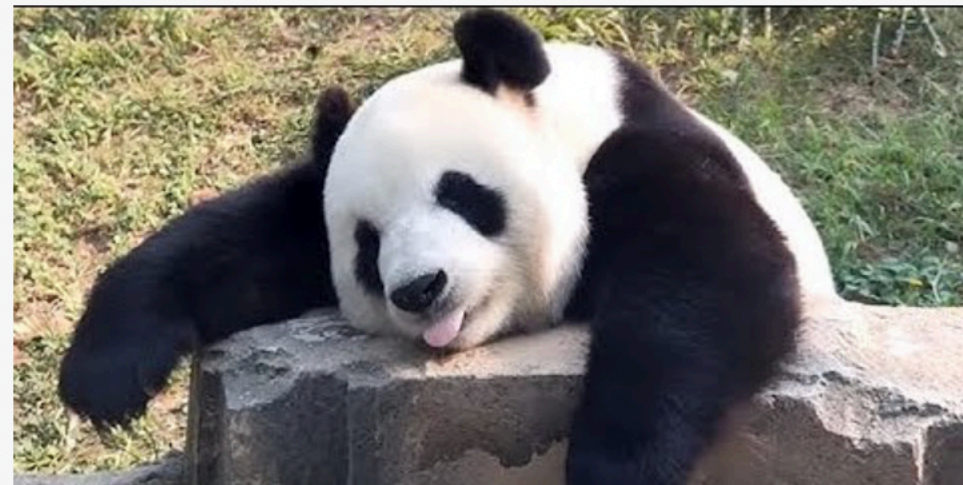
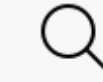
4.4M views • 2 years ago

 Chispa Motivation ✓

Query (Q)



deep learning



Panda Funny Moment Videos Compilation


2.9M views • 1 year ago

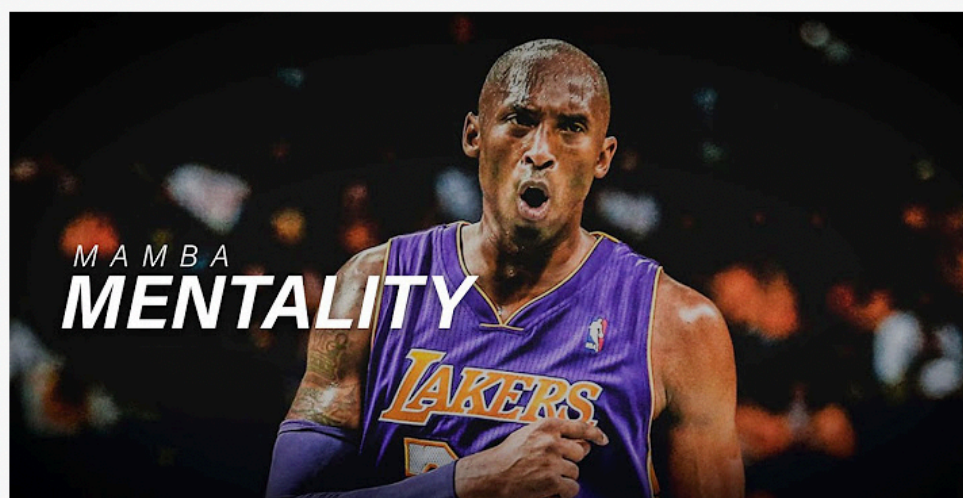
 Panda Story



MIT Introduction to Deep Learning | 6.S191


178K views • 1 month ago

 Alexander Amini



Mamba Mentality - Kobe Bryant (Motivational Video)

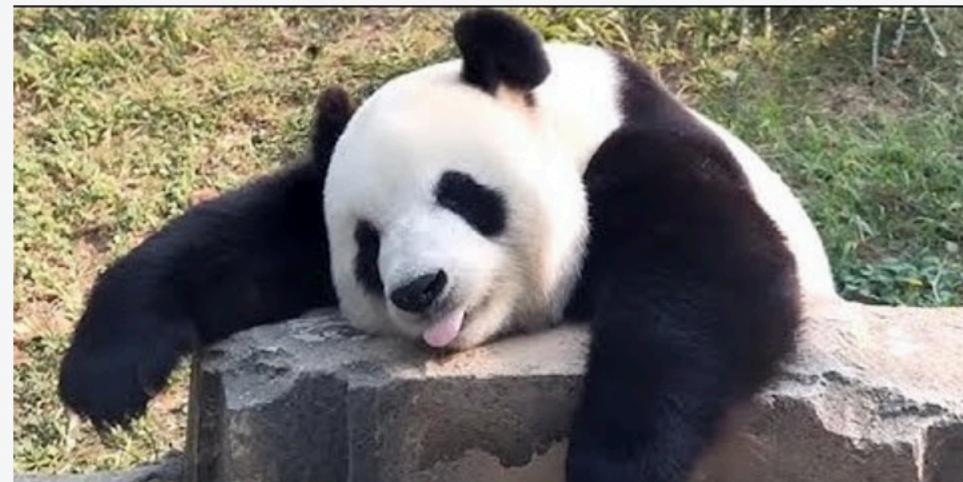
4.4M views • 2 years ago

 Chispa Motivation ✓

Query (Q)



deep learning



Panda Funny Moment Videos Compilation

2.9M views · 1 year ago

Panda Story

Key (K_1)

How similar is the key to the query?

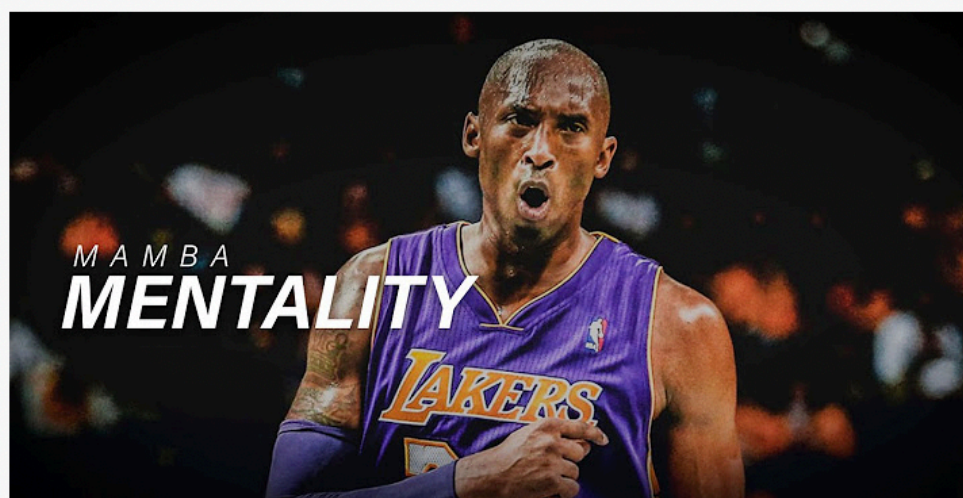


MIT Introduction to Deep Learning | 6.S191

178K views · 1 month ago

Alexander Amini

Key (K_2)



Mamba Mentality - Kobe Bryant (Motivational Video)

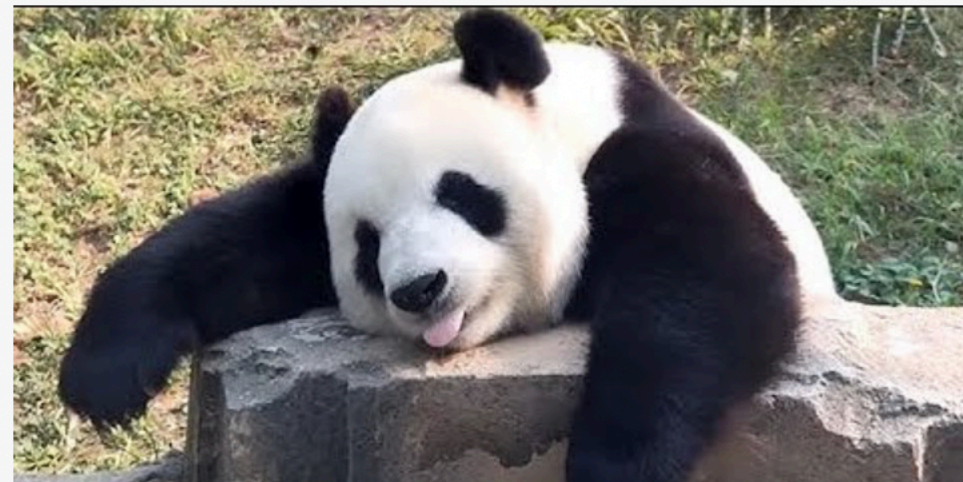
4.4M views · 2 years ago

Chispa Motivation

Key (K_3)



deep learning



Panda Funny Moment Videos Compilation

2.9M views · 1 year ago

Panda Story

Key (K_1)

Query (Q)

NOT SIMILAR



MIT Introduction to Deep Learning | 6.S191

178K views · 1 month ago

Alexander Amini

Key (K_2)

SIMILAR



Mamba Mentality - Kobe Bryant (Motivational Video)

4.4M views · 2 years ago

Chispa Motivation

Key (K_3)

NOT SIMILAR

Query (Q)



deep learning



Value (V)



MIT Introduction to Deep Learning | 6.S191

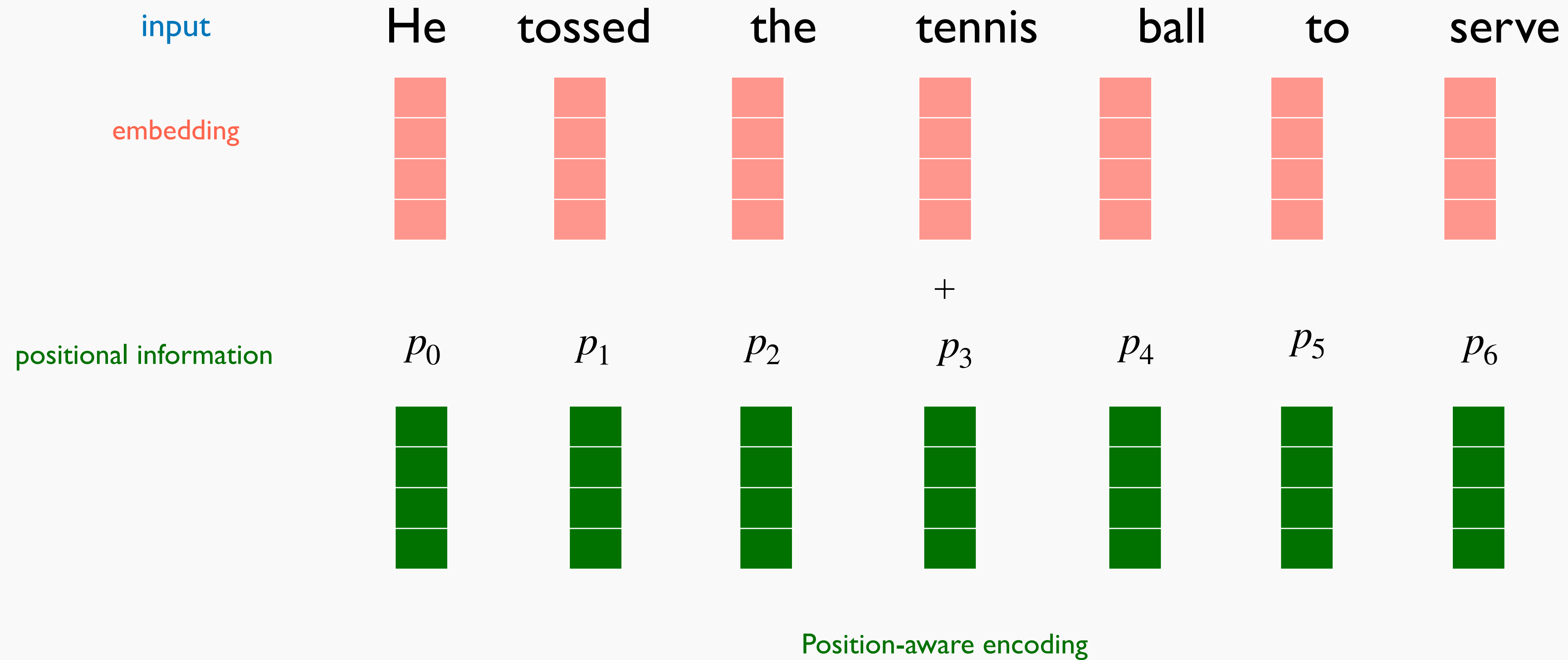
178K views · 1 month ago



Alexander Amini

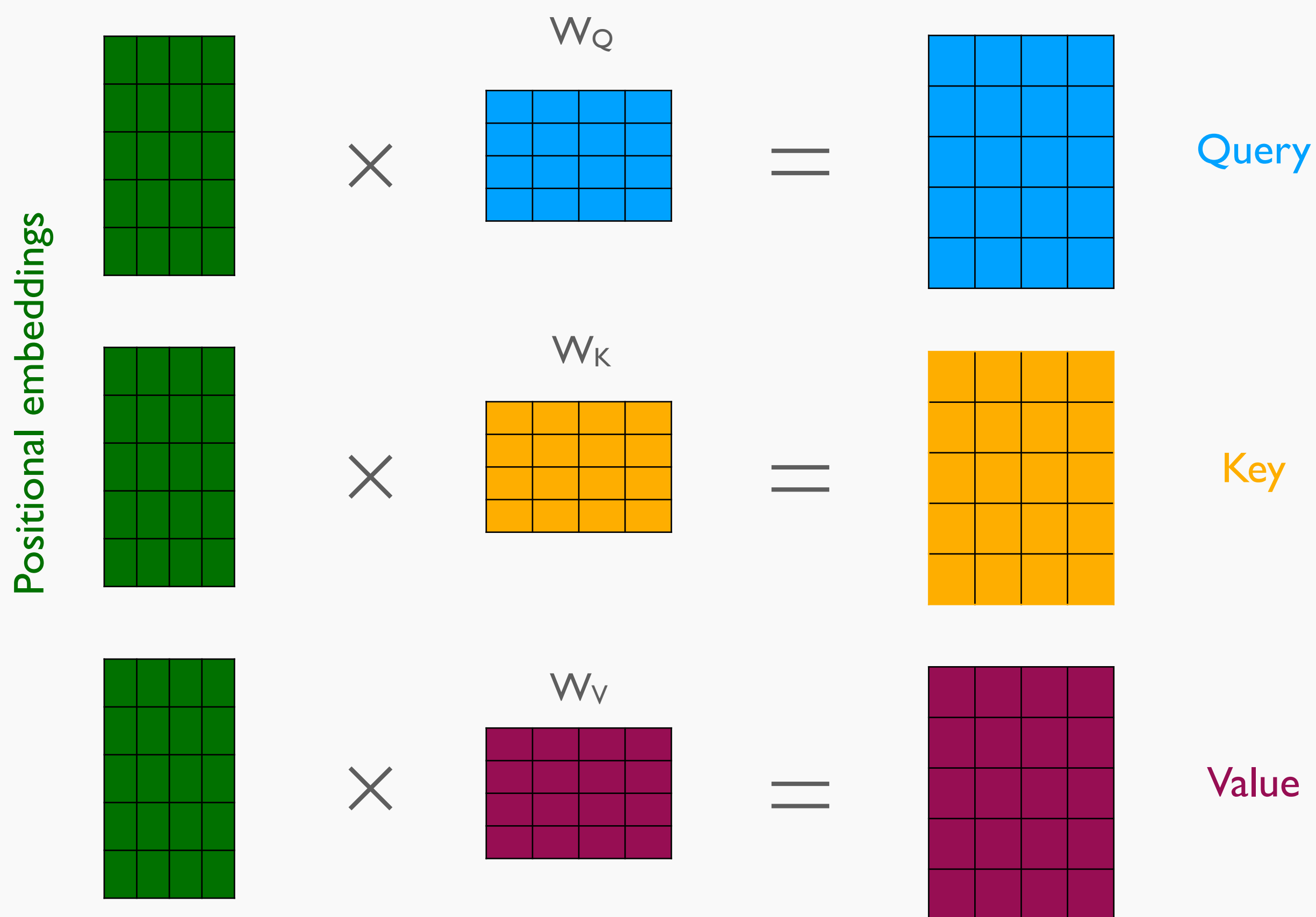
Key (K₂)

Positional Encoding

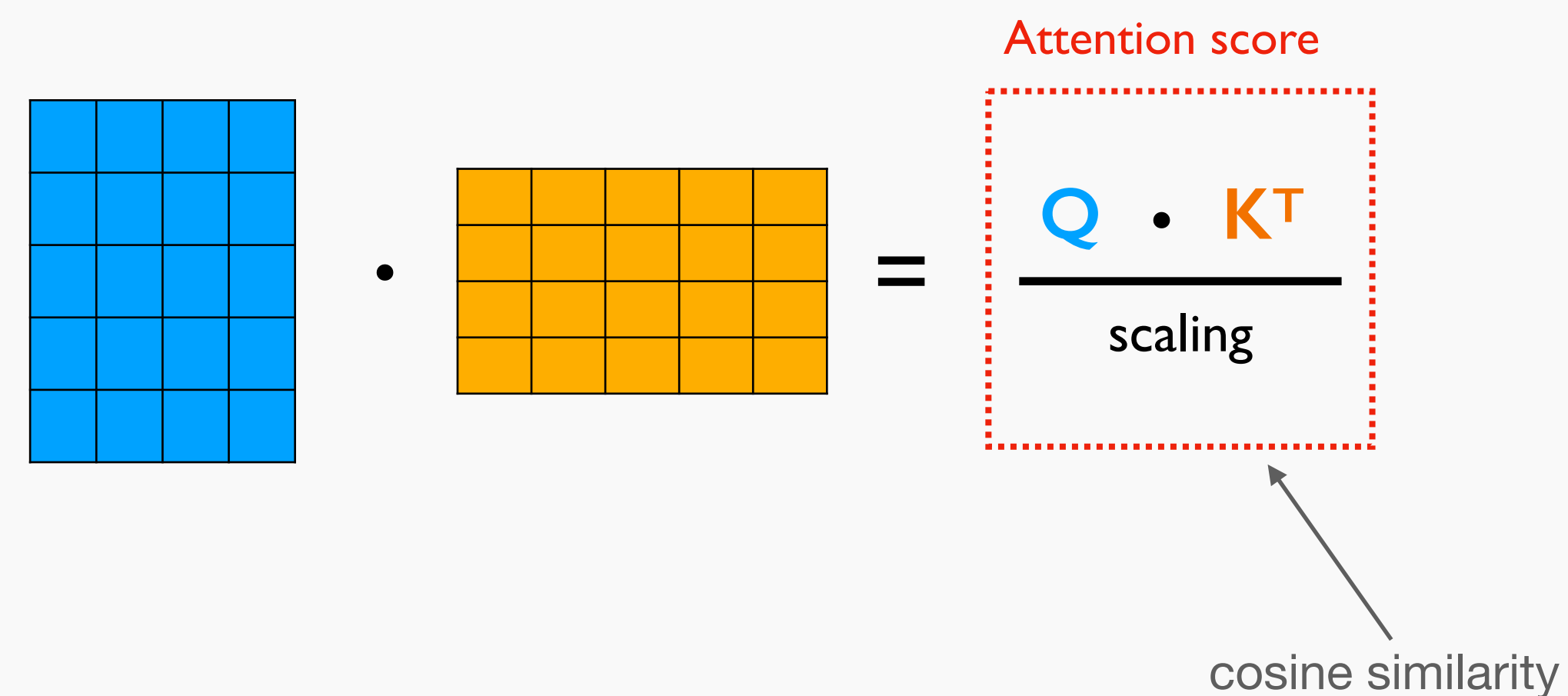


We need to preserve order without recurrence and without processing words individually.

1. Encode positional information.
2. Extract query, key and value.

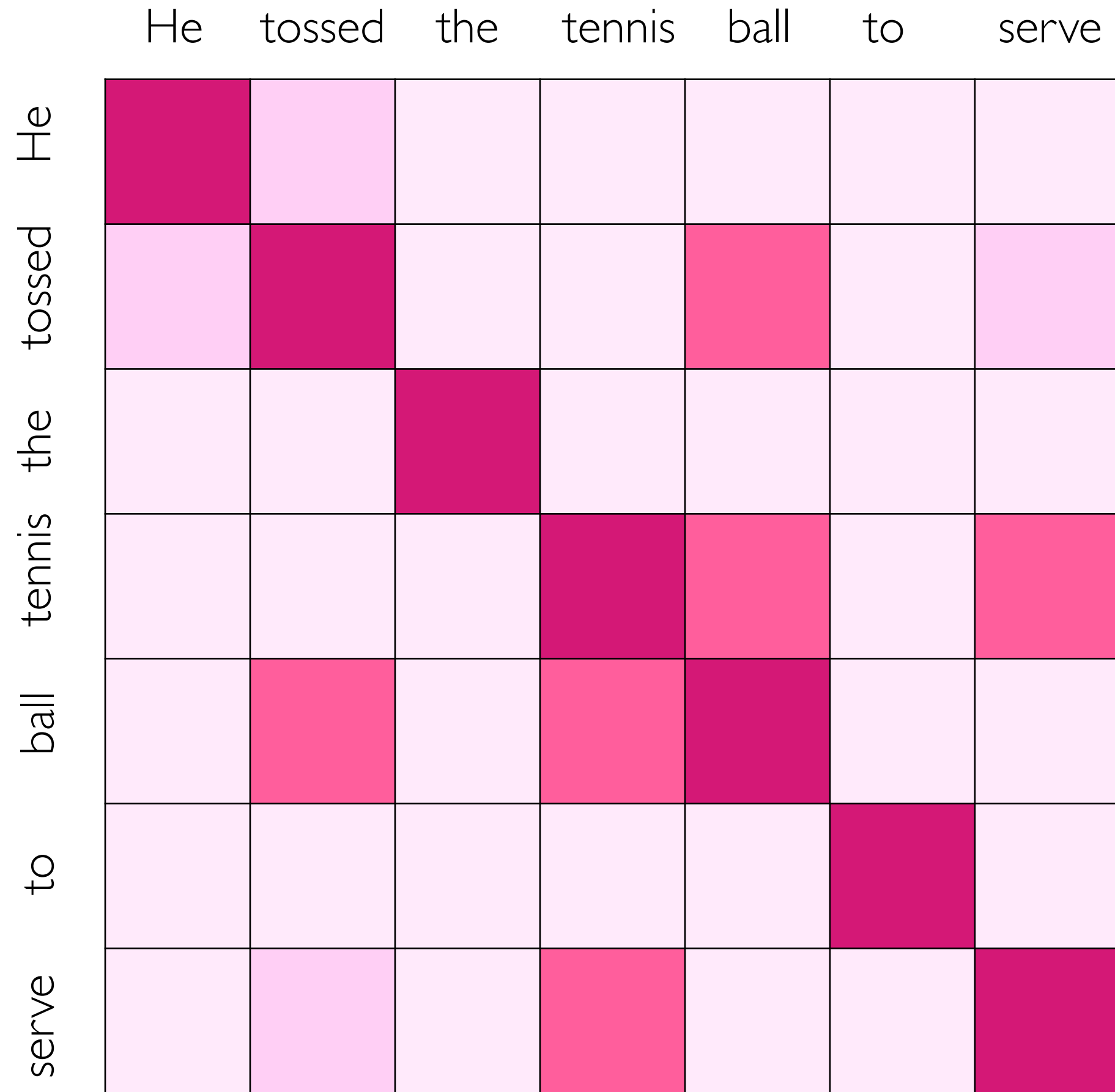


3. Compute attention weighting.



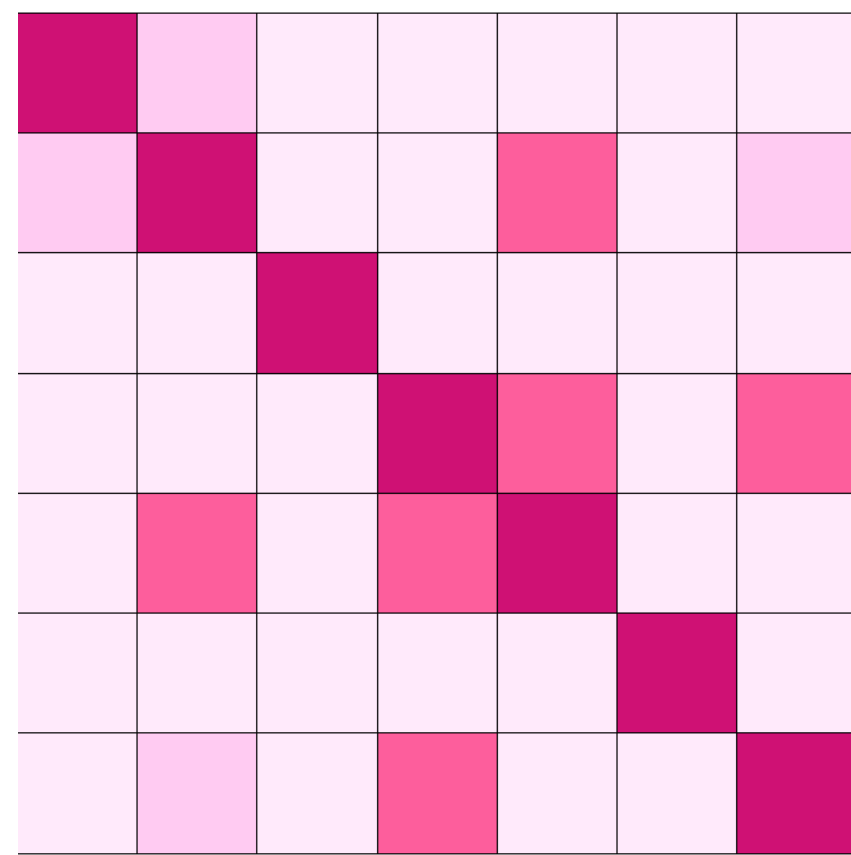
Attention weighting

$$\frac{Q \cdot K^T}{\text{scaling}} =$$



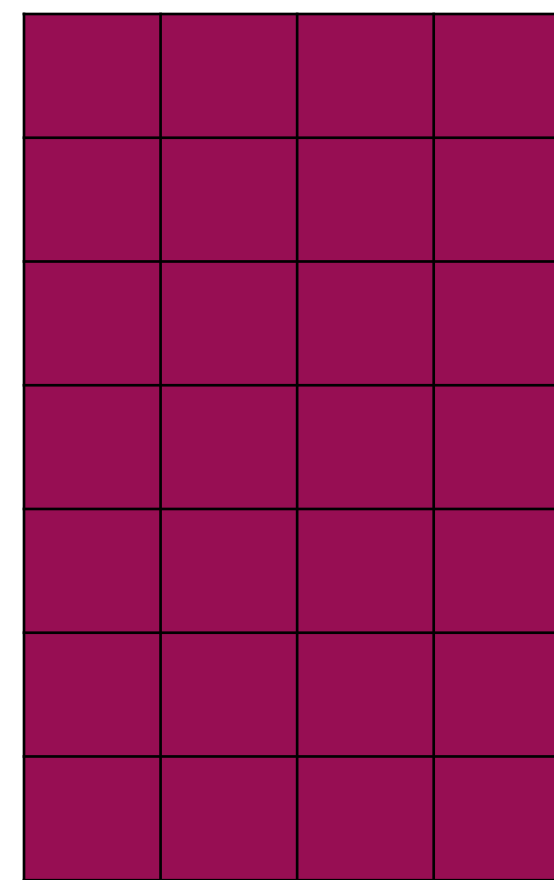
$$\text{softmax}(\text{matrix})$$

4. Extract features with high attention.



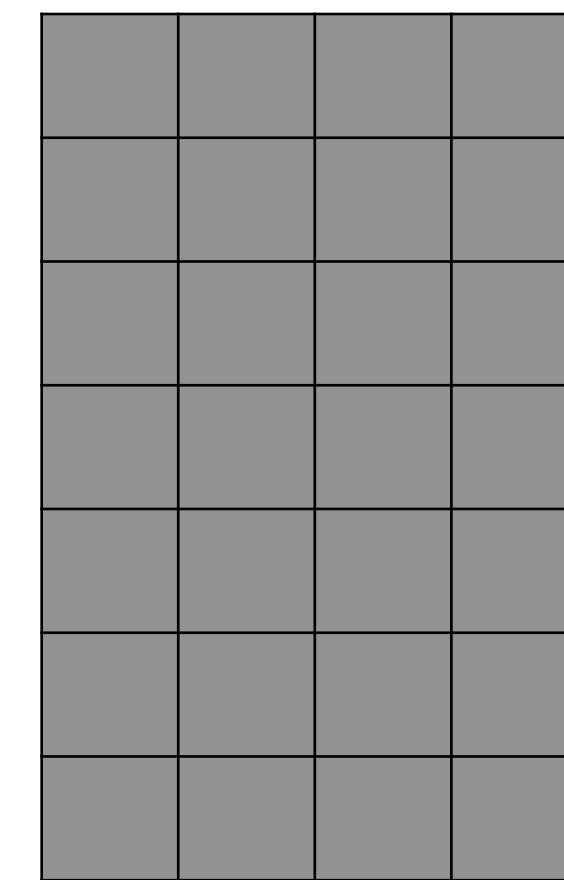
Attention Weighting

×



Value

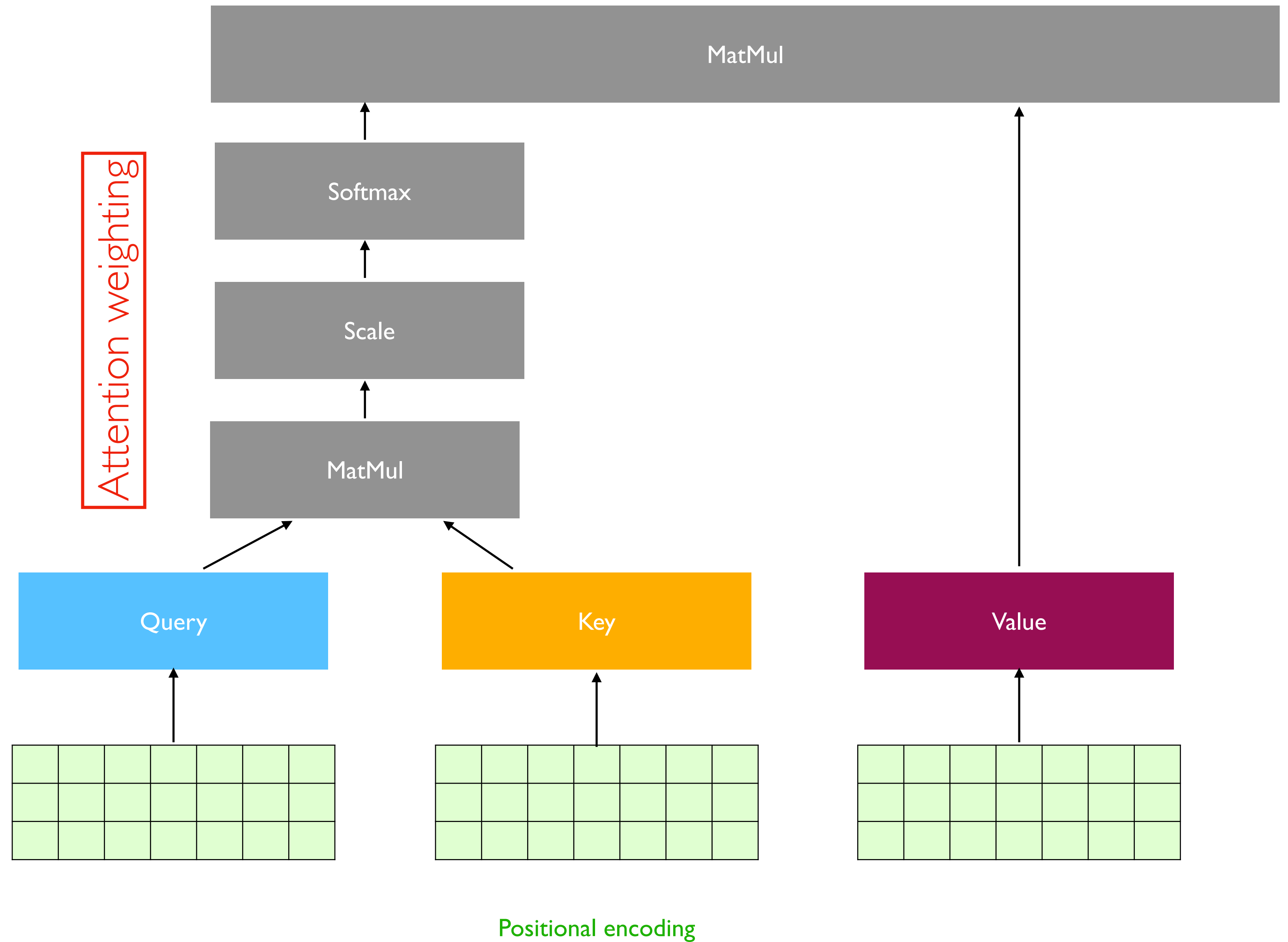
=



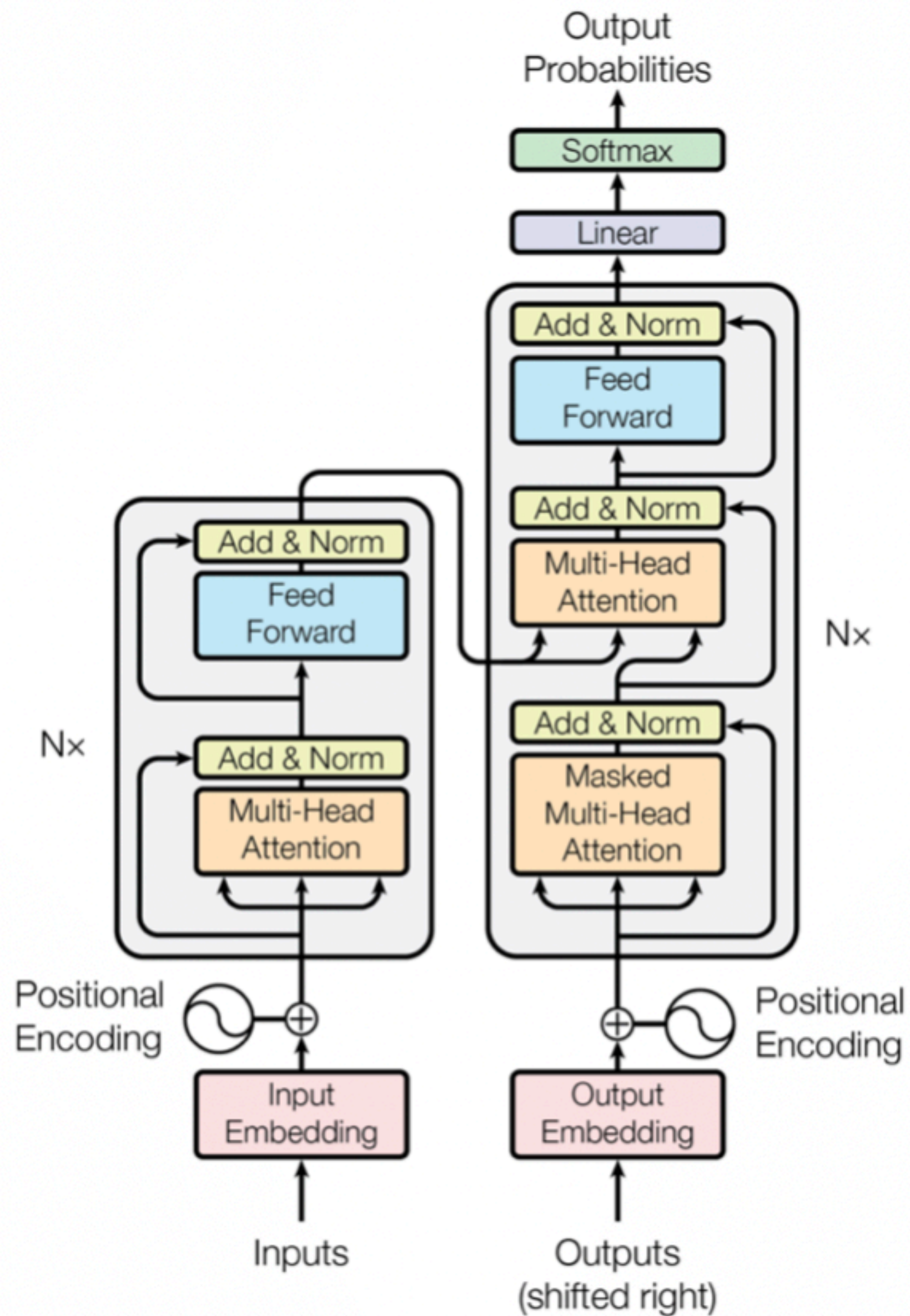
Output

$$\text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\text{scaling}}\right) \cdot \mathbf{V} = \mathbf{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$$

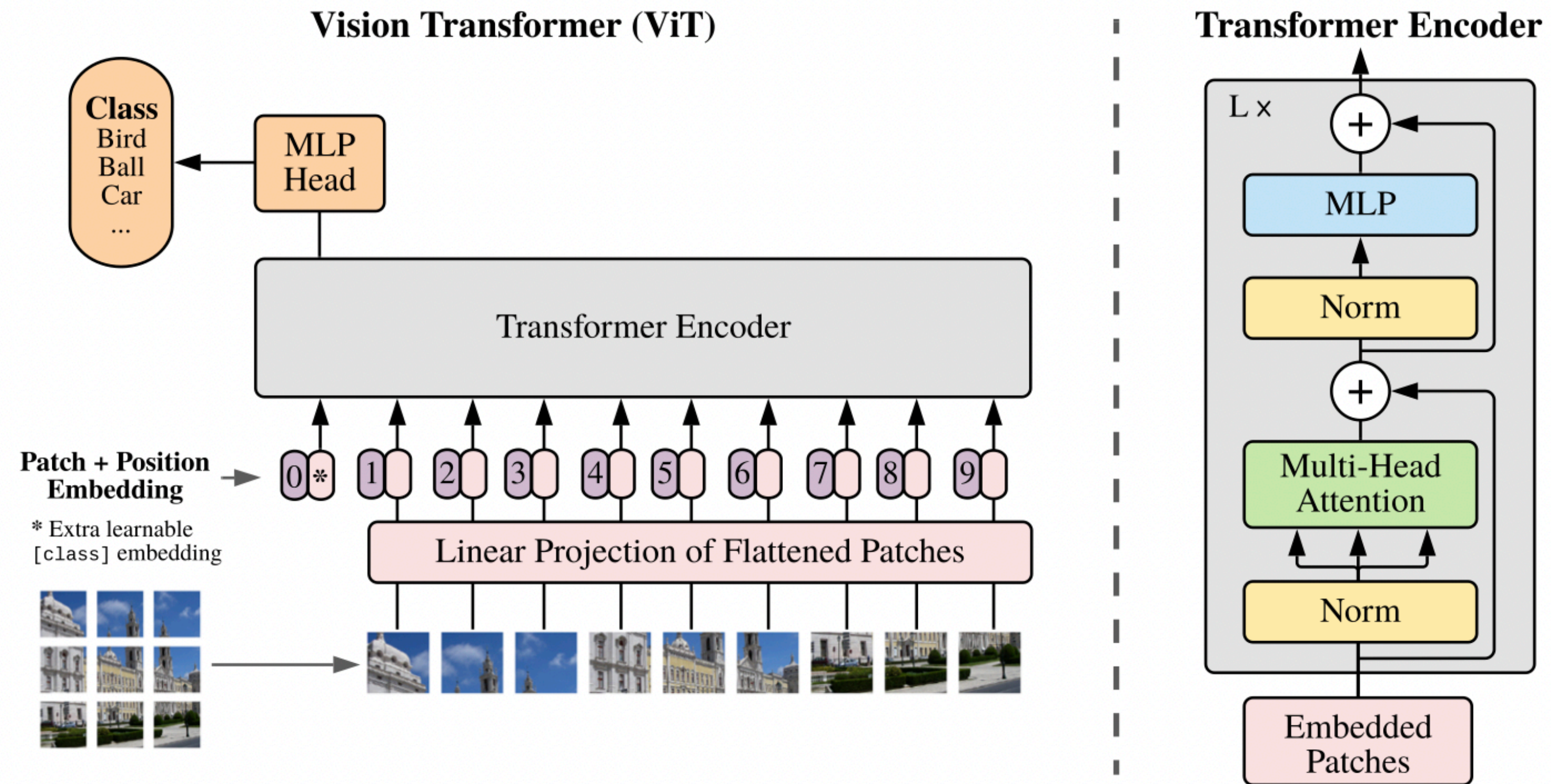
Self-attention head



Attention Is All You Need



An Image Is Worth 16x16 Words



Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Pre-trained models (transfer learning)



[TensorFlow Hub](#)



[HuggingFace Library](#)



[Keras Applications](#)



[PyTorch Hub](#)

BERT

Bidirectional Encoder Representations from Transformers

GPT-3

Generative Pre-trained Transformer

Transfer learning will be the next driver of machine learning's commercial success after supervised learning.

Andrew Ng

Thank you for your attention ;)