Introduction to Neural Networks

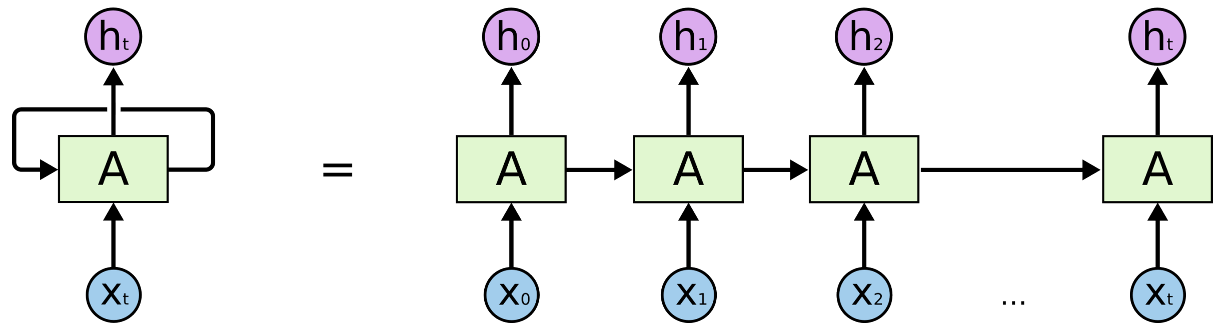
Transformers and the Attention Mechanism

Francis Engelmann

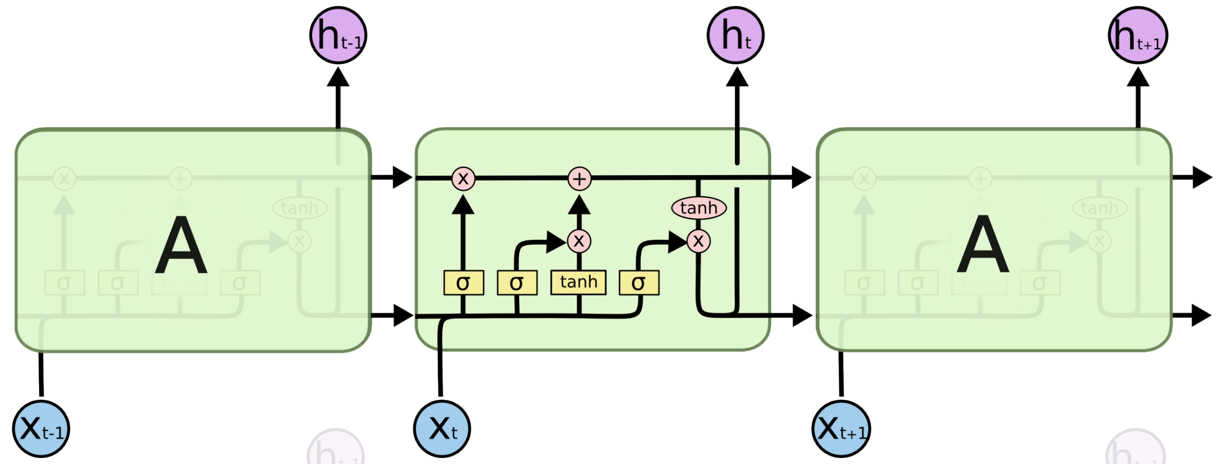
7 June 2024

Last Week: Recurrent Neural Networks (Recap)

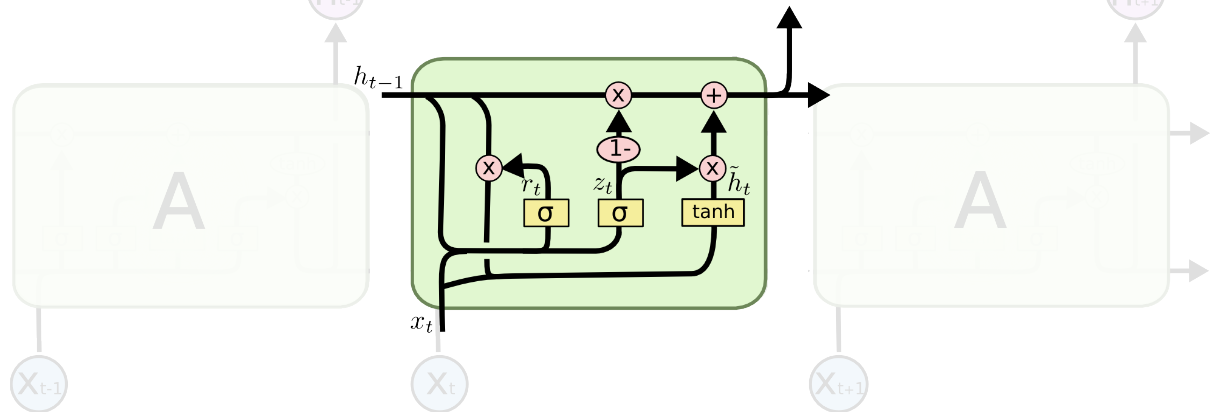
- Recurrent Neural Networks (RNN)



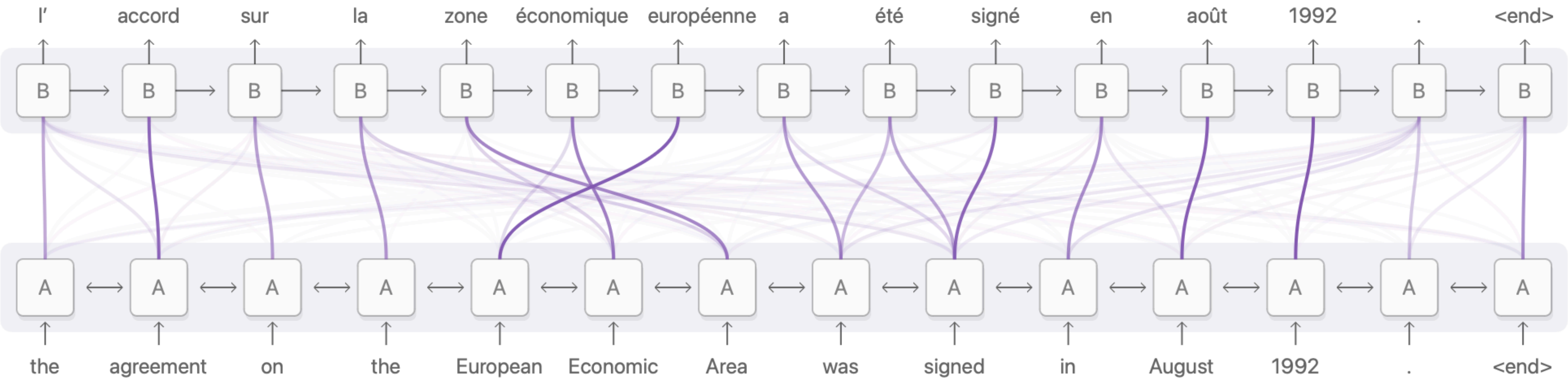
- Long Short Term Memory (LSTM)



- Gated Recurrent Unit (GRU)



- **Today**: Attention and Transformer



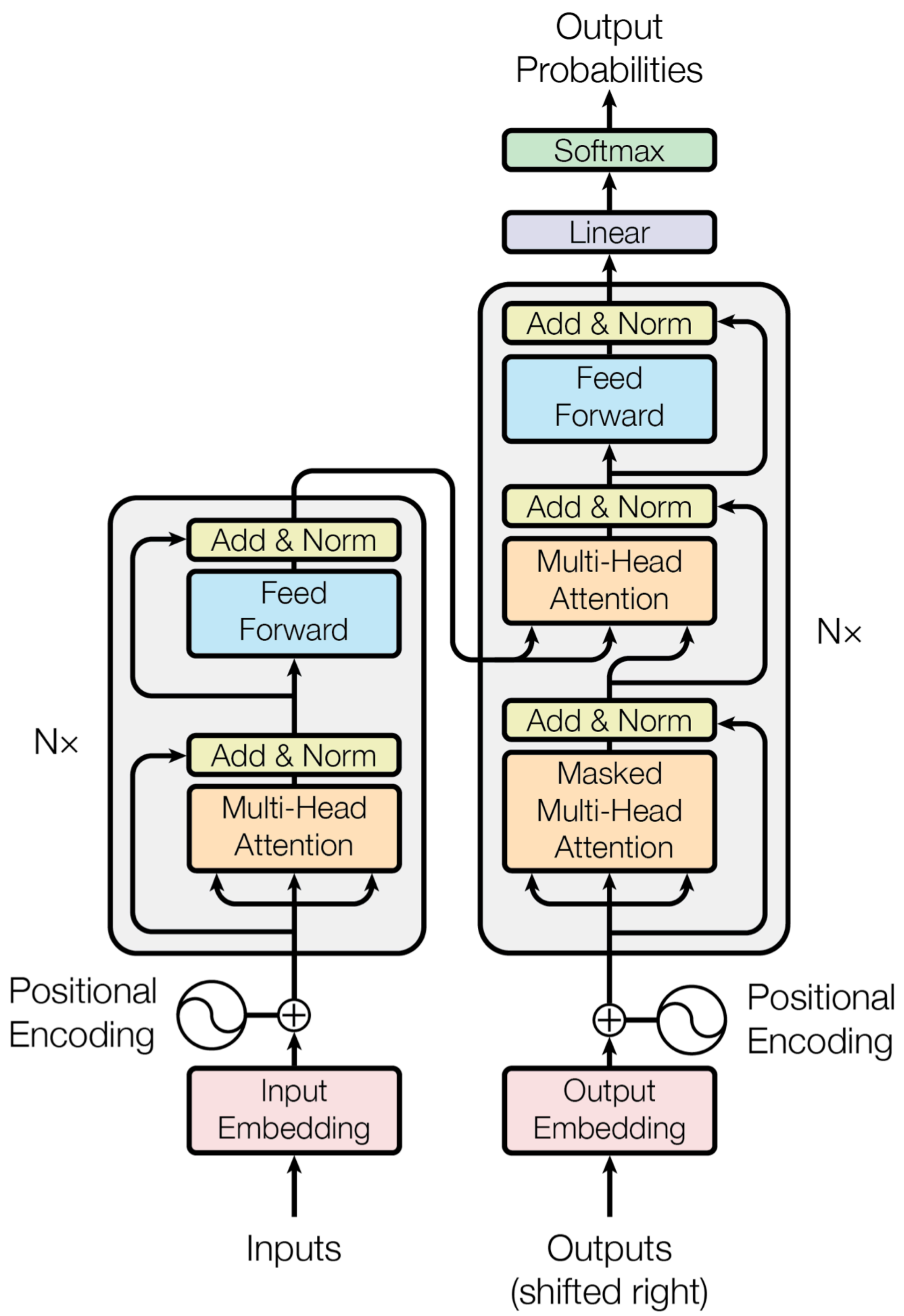
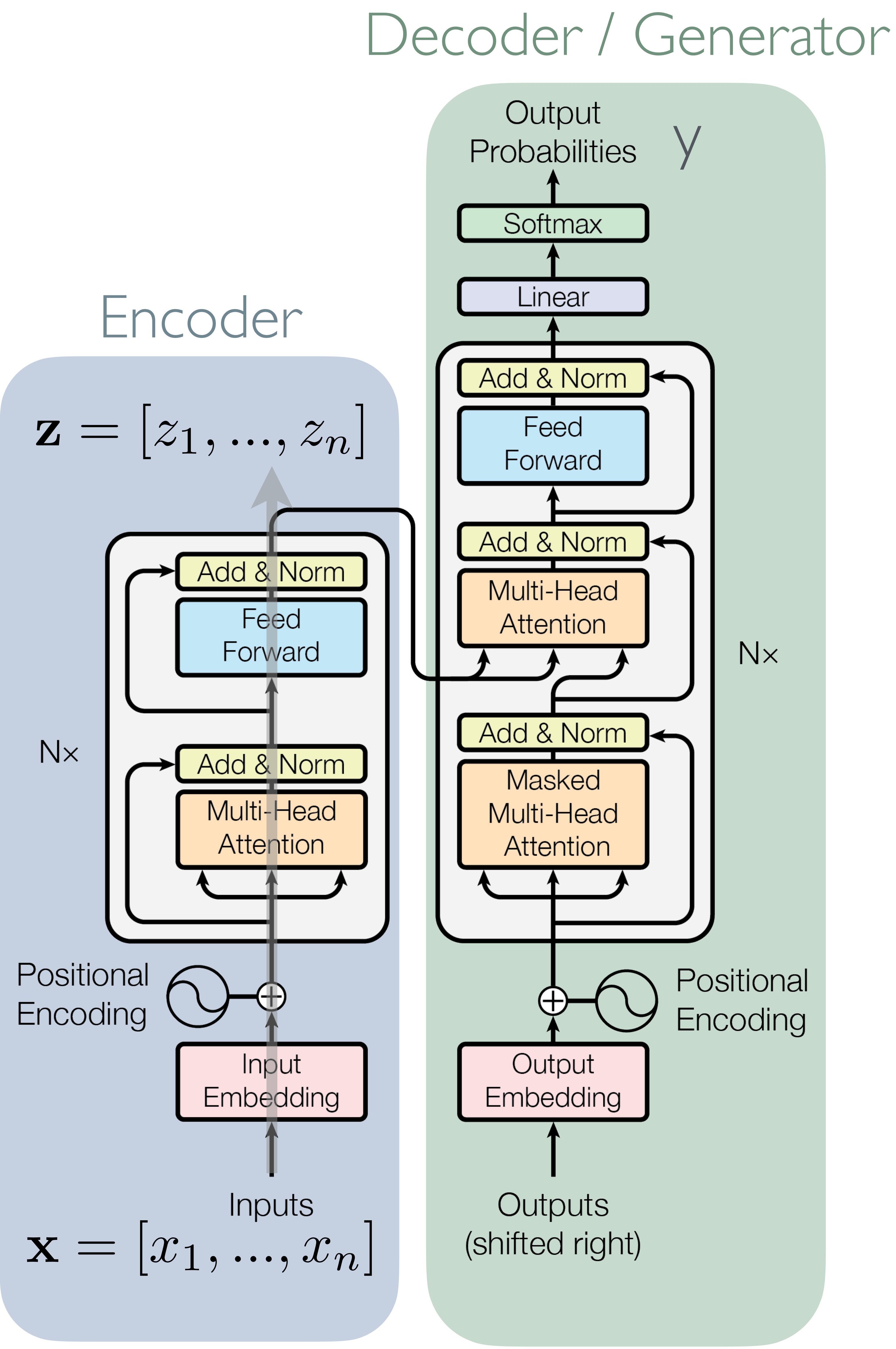


Image credit: Chris Olah, Vaswani et al.

The Transformer

Model Architecture

Encoder:
maps input sequence of tokens $\mathbf{x}$ to sequence of learned representations $\mathbf{z}$



Decoder:

given z, the decoder generates output sequence (y_1, ..., y_m) auto-regressively predicts one element at a time

add examples of the generated output steps

Image credit: Vaswani et al.

Attention

Image credit: Vaswani et al.

Computational Cost vs RNNs

Summary and Outlook

- Transformer model
- Attentation
- **Next week:**
  - Linear Auto-encoders (PCA)
  - Variational Autoencoders

References & Further Reading

Slides & Code

- Code example: MinGPT from Andrey Karpathy https://github.com/karpathy/minGPT?tab=readme-ov-file
- RNN/LSTM http://colah.github.io/posts/2015-08-Understanding-LSTMs/