

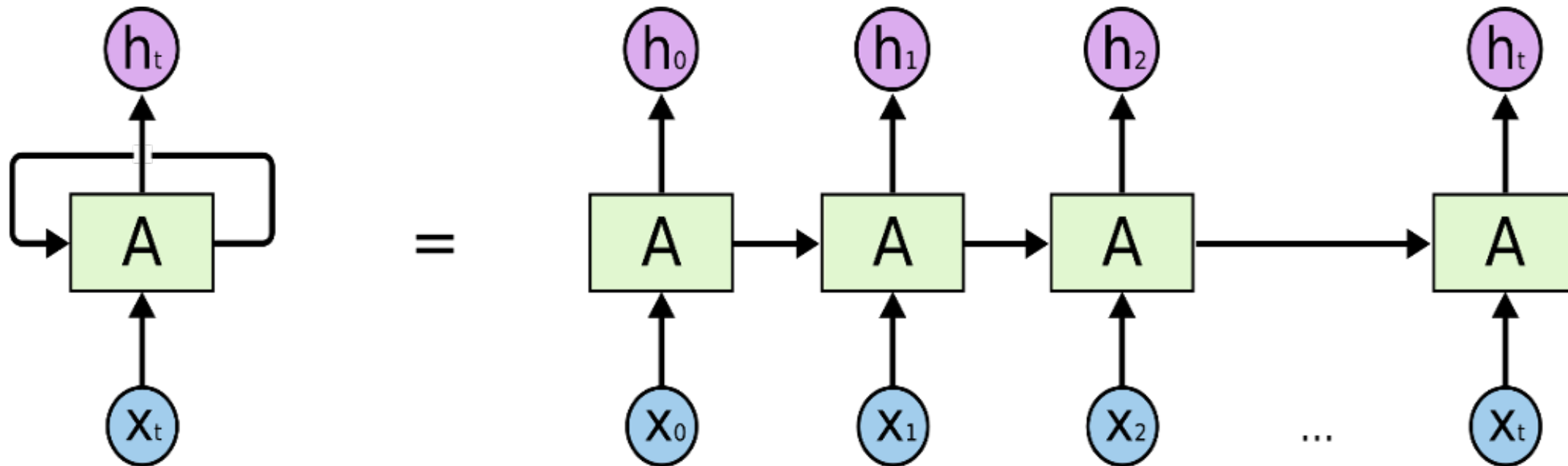
95-865 Unstructured Data Analytics

Last recitation: More on transformers
& Hugging Face 🤗

Slides by George H. Chen+Yubo

Recurrent Neural Network (RNN)

(Elman, 1990)



The Full Transformer Neural Net

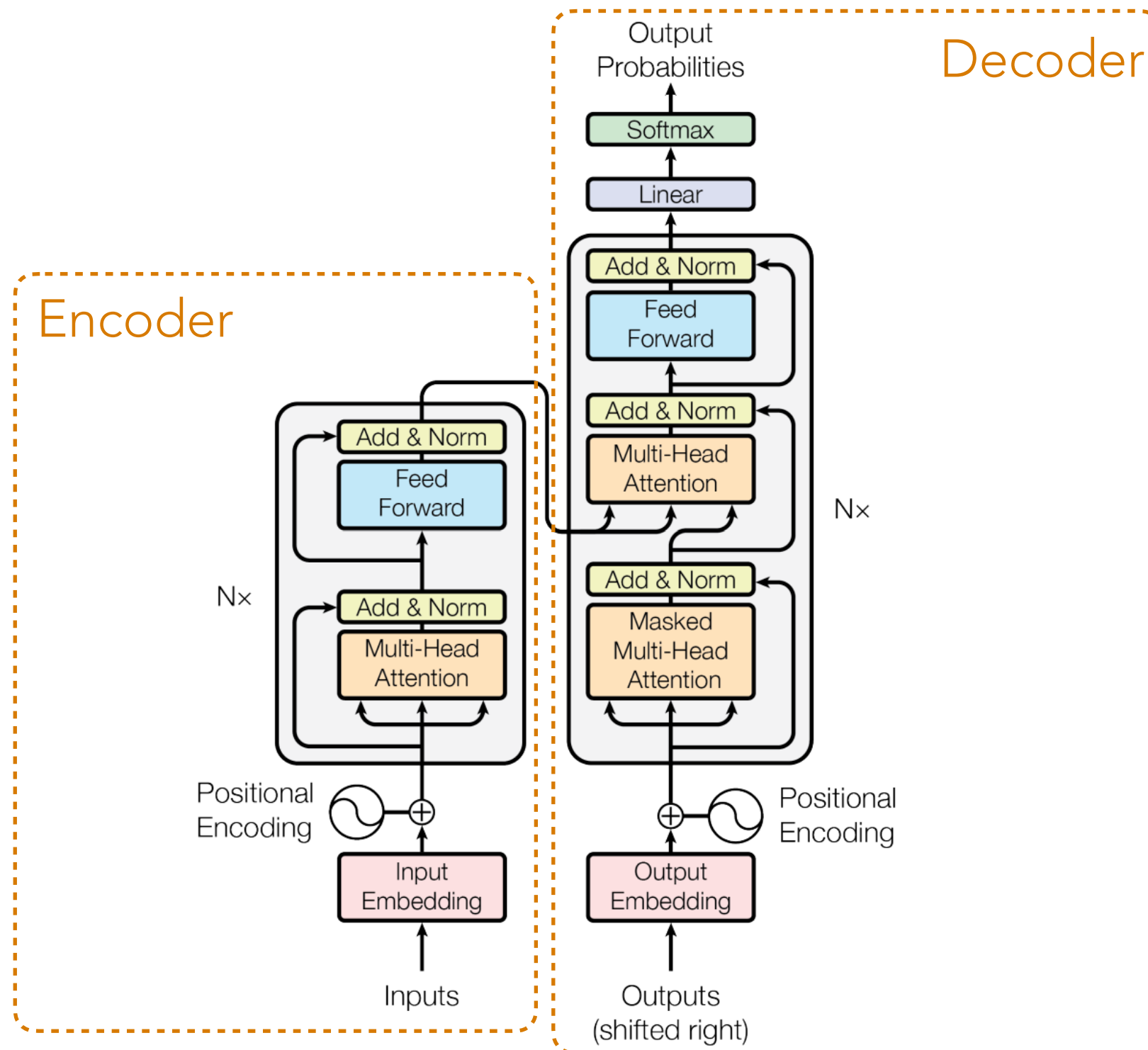


Figure 1: The Transformer - model architecture.

Vaswani et al. "Attention is All You Need". NeurIPS 2017.

The Full Transformer Neural Net

The original full transformer was used for translating between languages

Encoder sees input text (e.g., English)

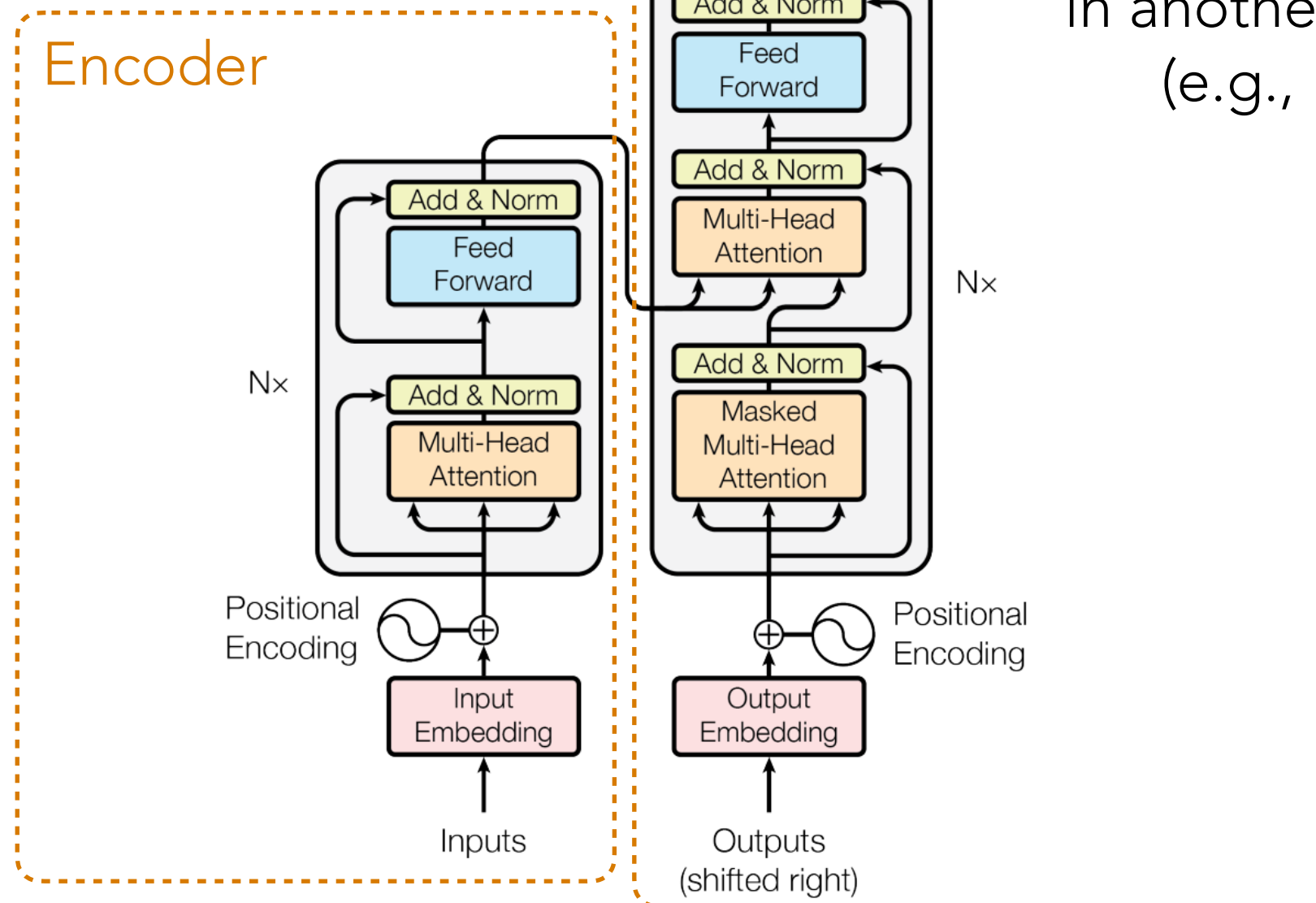


Figure 1: The Transformer - model architecture.

Vaswani et al. "Attention is All You Need". NeurIPS 2017.

(Flashback)

There are a few implementation details
that I won't go over in lecture

Basically, it turns out that when neural nets get very deep,
training can be more difficult without some now-standard tricks
(these tricks work with *many* neural net architectures, not just GPTs)

- LayerNorm
- Residual connections
- Dropout

You're not expected to
know these technical details

Also, there are some standard strategies for initializing GPT training

Decoder-Only Transformer

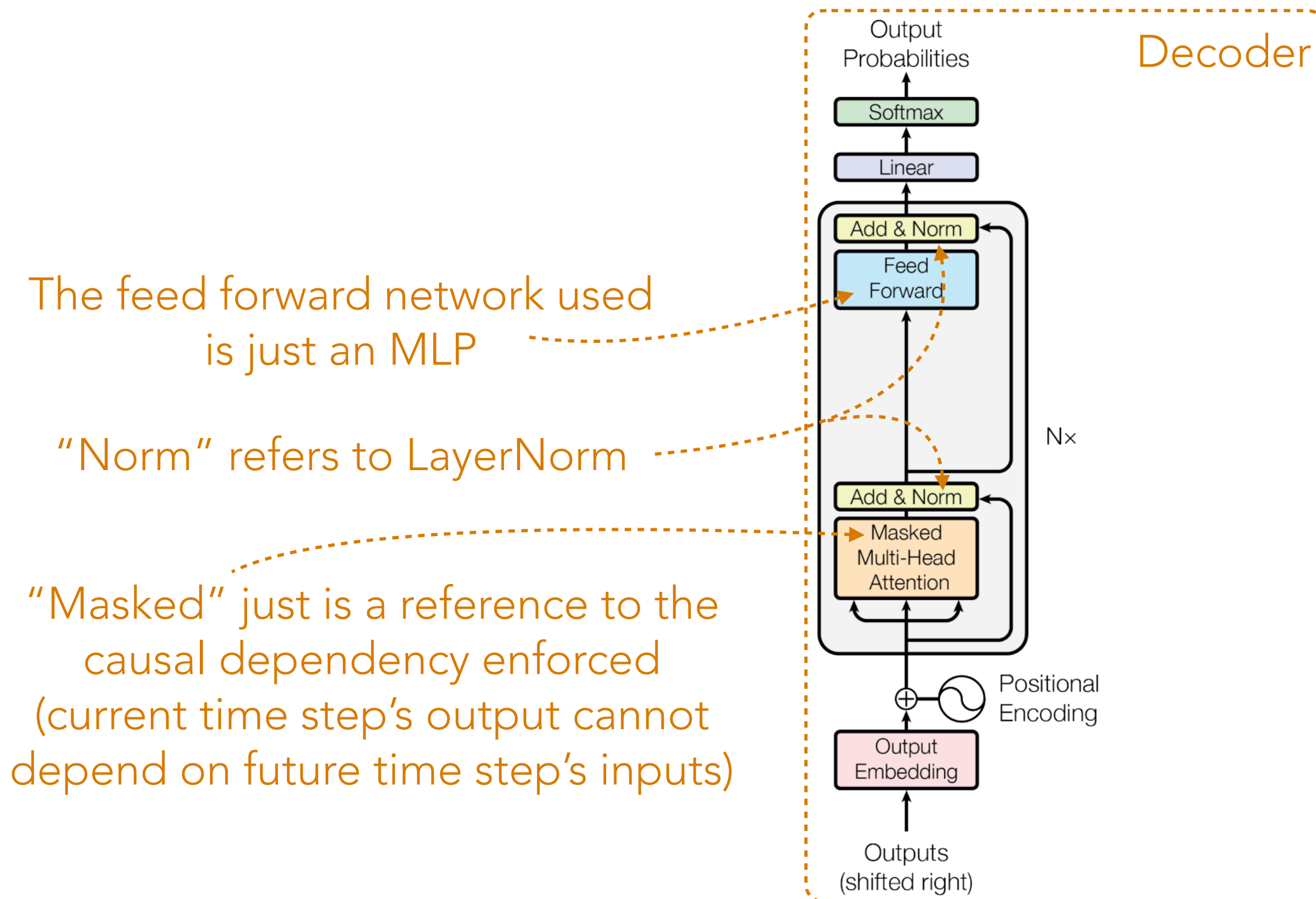


Figure 1: The Transformer - model architecture.

Vaswani et al. "Attention is All You Need". NeurIPS 2017.

Decoder-Only Transformer

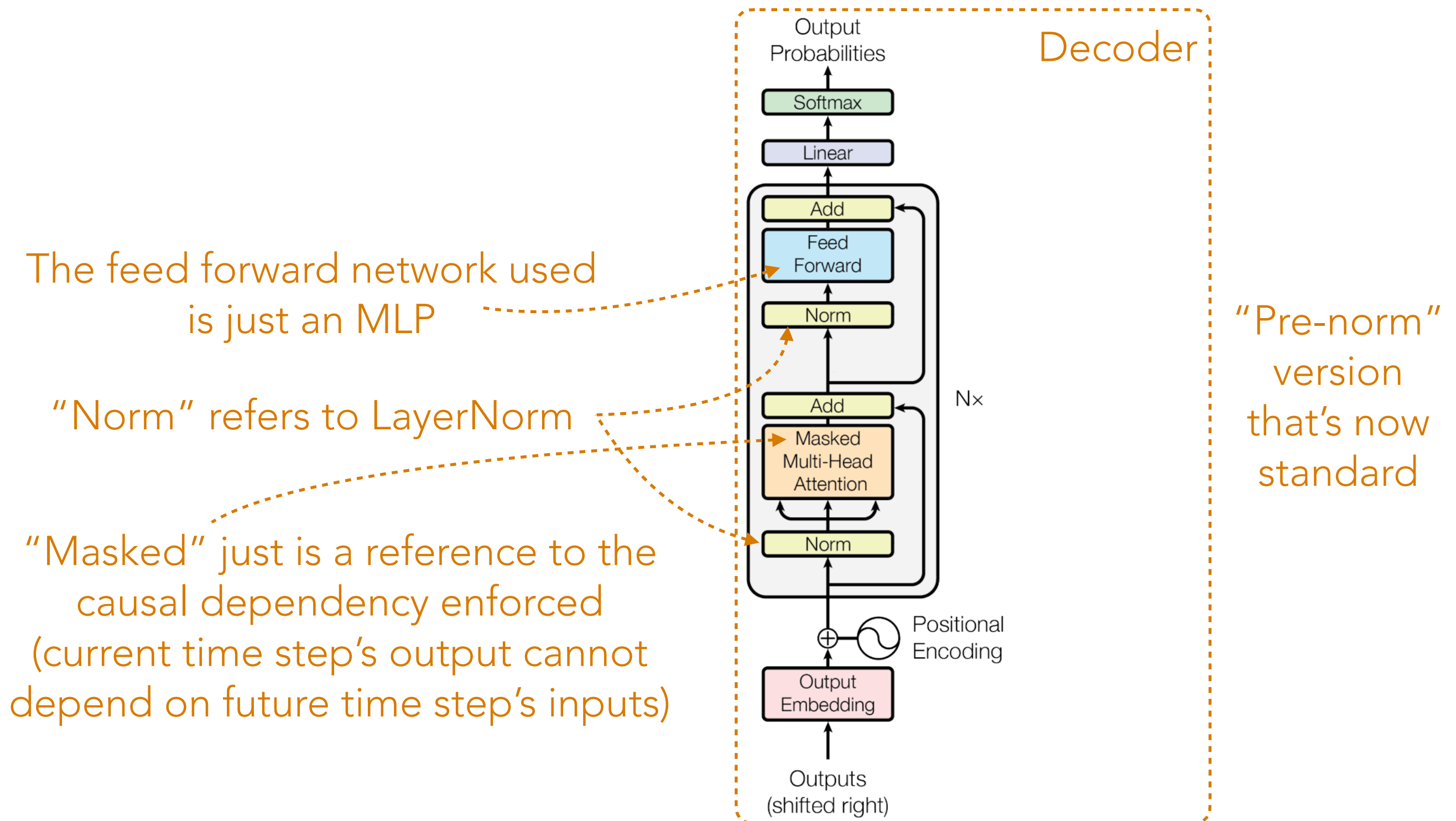


Figure 1: The Transformer - model architecture.

Vaswani et al. "Attention is All You Need". NeurIPS 2017.

Encoder-Only Transformer

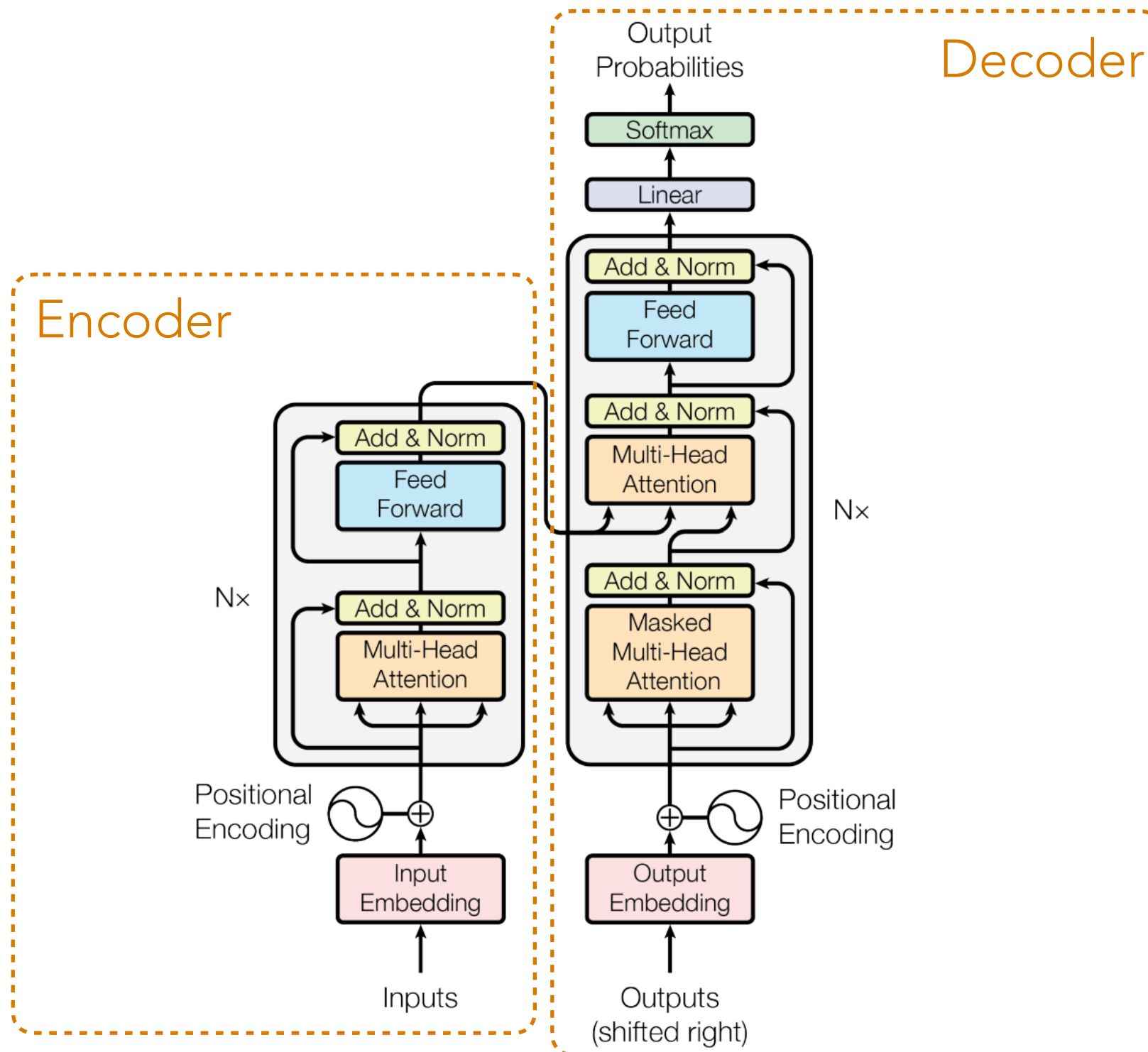
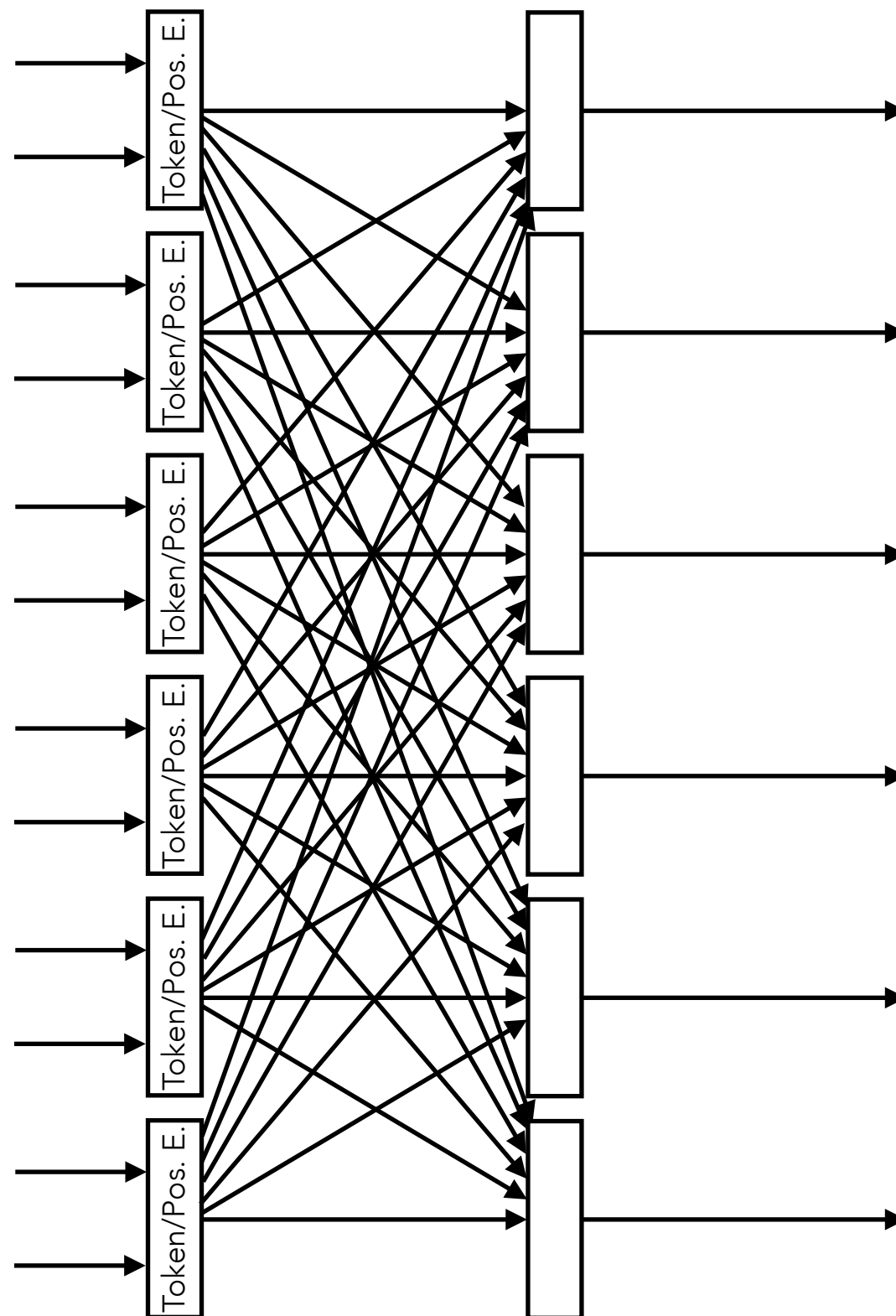


Figure 1: The Transformer - model architecture.

BERT (2018)



no causal dependence

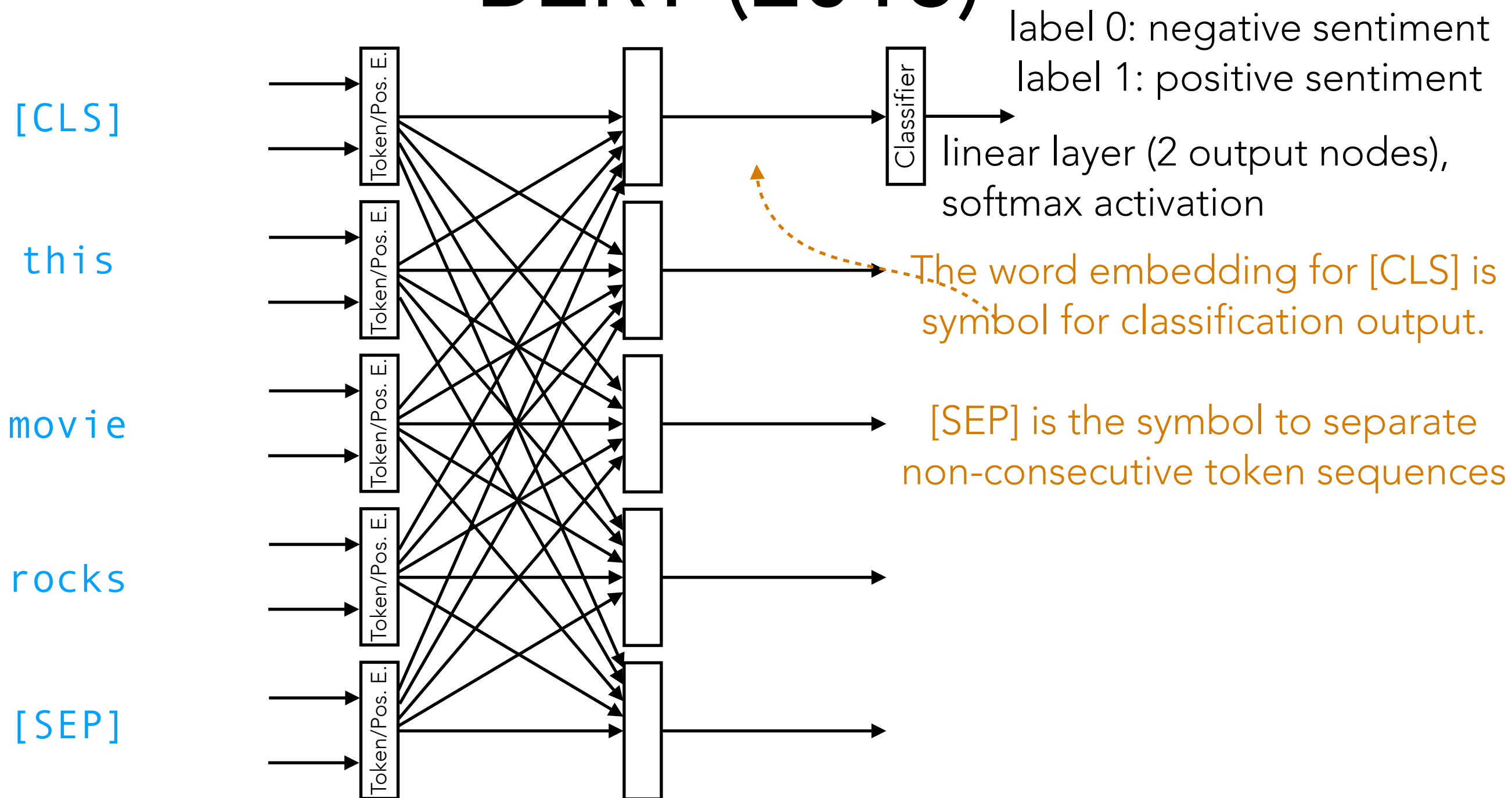
The prediction at any time step depends on the input at all time steps

This lack of causal dependence is also sometimes referred to as "bidirectional"

A transformer layer like this without a causal constraint is sometimes called an "encoder-only" transformer layer

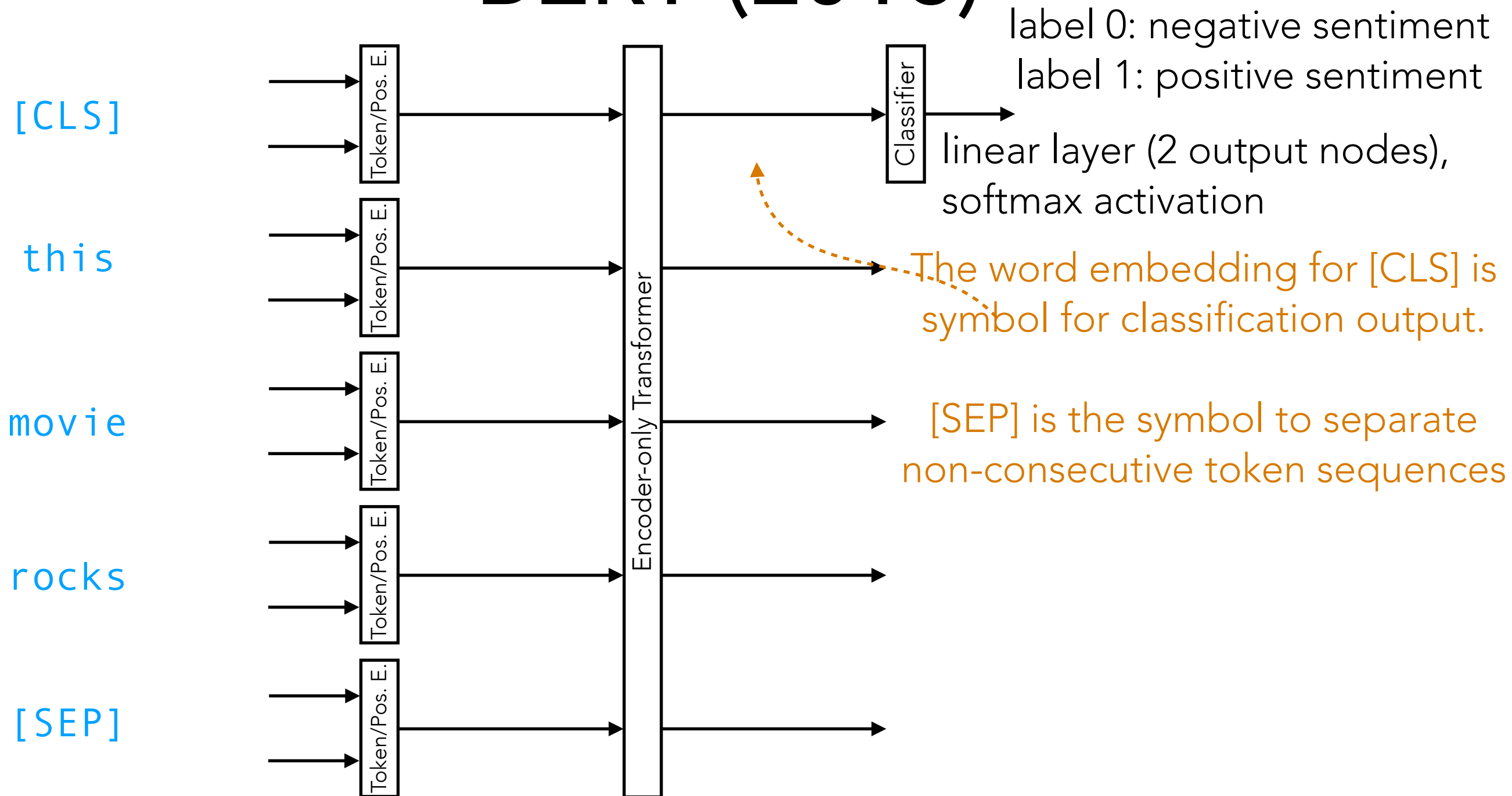
BERT is short for Bidirectional Encoder Representations from Transformers

BERT (2018)



BERT actually adds an initial "[CLS]" token and an ending "[SEP]" token (these are called **special tokens**)

BERT (2018)



BERT actually adds an initial "[CLS]" token and an ending "[SEP]" token (these are called **special tokens**)

Sentiment Analysis: Transformer Fine-tuning Without Using an RNN

Demo

More on Hugging Face 🤗

Demo (uses an airline tweets dataset from Kaggle):

- How to load in a pre-trained Hugging Face sentiment analysis model and use it (without re-training nor fine-tuning)
- How to use Hugging Face's pipelines functionality
- Topic modeling using a BERT-based topic model (BERTopic)