# TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

## The Transformer

# The Transformer

Attention is All You Need, Vaswani et al., June 2017

The transformer has now essentially replaced RNNs in natural language applications.

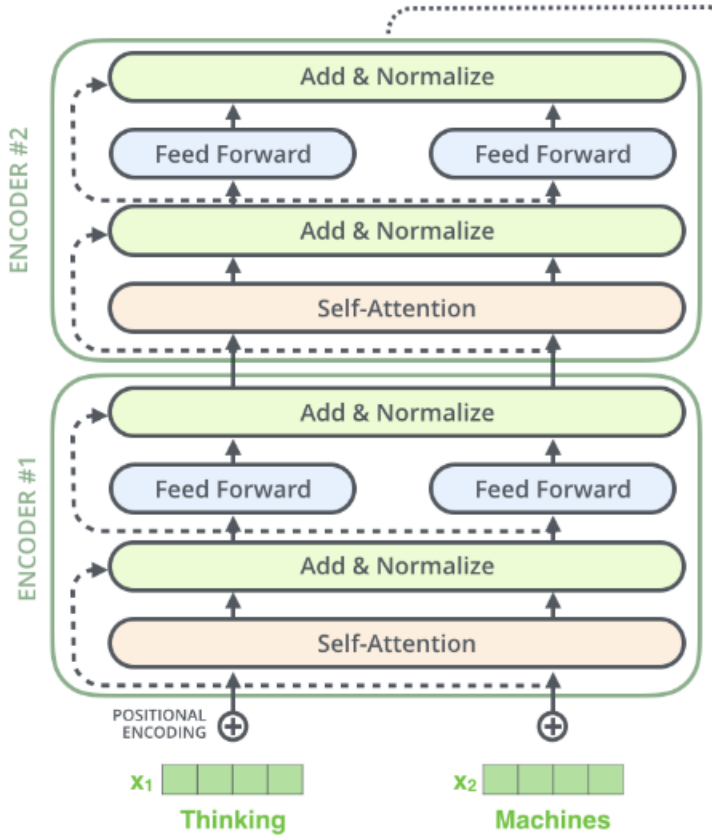
The recent progress on natural language understanding (GLUE) is based on general language modeling using transformers.

# The Transformer

Unlike RNNs, transformers run in parallel time in proportional to the layering depth independent of the length of the input sequence.

Transformers also do a better job of allowing information early in the sequence to be used later in the sequence — they have better memory when used as a langauge model.

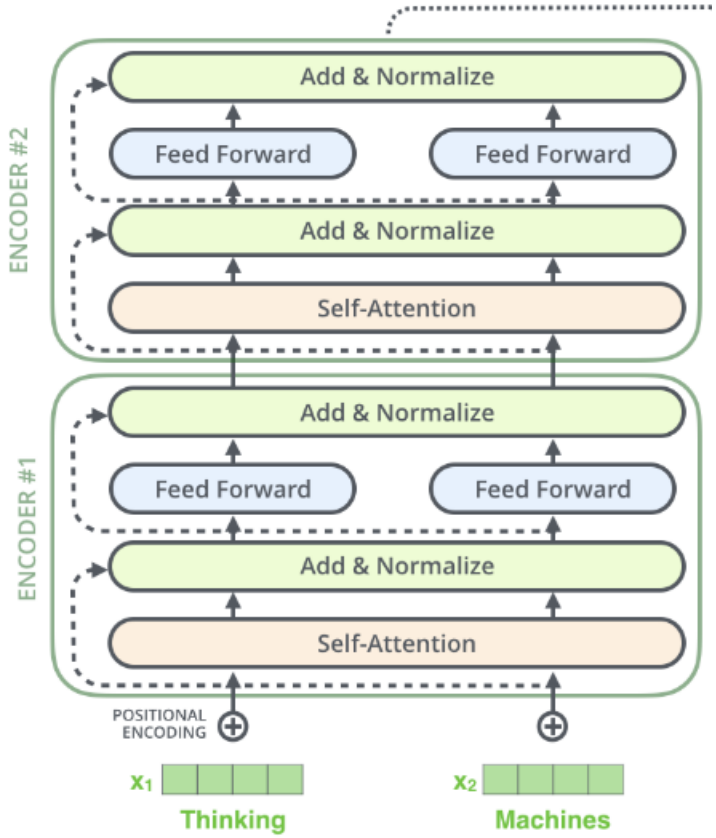
# The Transformer




Jay Alammar's blog

All layers run in $O(1)$ parallel time independent of text length.

# The Transformer




Jay Alammar's blog

Layers are stacked with residual connections.

# A Self-Attention Layer

Given an $h_{\text{in}}[T, J]$ we will construct $h_{\text{out}}[T, J]$ ("in" and "out" refer to layering, not translation).

We first construct a head-specific self-attention $\alpha[k, t_1, t_2]$ — the attention position $t_1$ is giving to position $t_2$ for "head" $k$.

The motivation for different heads is that there are different relationships between words such as "refers to" for pronouns, or "subject of" and "object of" for verbs. But the meaning of each head is not specified and emerges from training.

## Computing the Self Attention

For each head $k$ and position $t$ we compute a key vector and a query vector with dimension $U$ typically smaller than dimension $J$.

$$\text{Query}[k, t, u] = W^Q[k, u, J] h_{\text{in}}[t, J]$$

$$\text{Key}[k, t, u] = W^K[k, u, J] h_{\text{in}}[t, J]$$

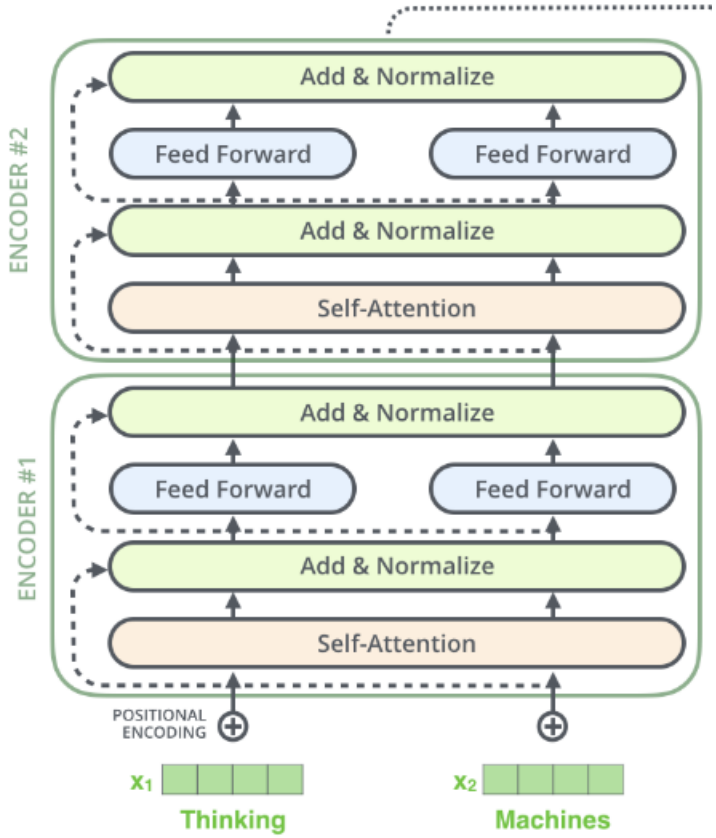
$$\alpha[k, t_1, t_2] = \underset{t_2}{\text{softmax}}\ \text{Query}[k, t_1, U]\text{Key}[k, t_2, U]$$

## Computing the Output

We require $I = J/K$.

$$\begin{aligned}
\text{Value}[k, t, i] &= W^V[k, i, J] h_{\text{in}}[t, J] \\
\tilde{h}_{\text{out}}[k, t, i] &= \alpha[k, t, T_2] \text{Value}[k, T_2, i] \\
h_{\text{out}}[t, J] &= \tilde{h}_{\text{out}}[1, t, I]; \cdots ; \tilde{h}_{\text{out}}[K, t, I]
\end{aligned}$$

# The Transformer




Jay Alammar's blog

Position encodings are inserted at the bottom.

## Encoding Positional Information

At the input layer we augment the word embeddings with position information. For example:

$$h[0, t, J] = e[w[t], I]; e^{i\omega t}; e^{i2\omega t}; e^{i4\omega t} \cdots; e^{i2^k \omega t}$$

In modern versions a position encoding is trained for each position in the text.

## ELMO: Language Modeling

To do language modeling we fix $\alpha[k, t_1, t_2] = 0$ for $t_2 > t_1$.

We can then predict the word $w_t$ as

$$P(w_t | w_1, \ldots, w_{t-1}) = \underset{w_t}{\text{softmax}} \; e[w_t, I] h_{\text{top}}[t-1, I]$$

where $h_{\text{top}}$ is the top level hidden vector of the transformer.

## Machine Translation

Translation is just a conditional language model.

We take the input English sentence followed by a special token and then generate the output from the transformer language model.

## Continuing from a Prompt

GPT-2 from Open AI (1.5 billion parameters, June 2018)

Continue from:

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

## The Predicted Continuation

Continuation excerpted from a single response, the response selected from 10 tries.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.
Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when ... Pérez and his friends were astonished to see the unicorn herd. ... While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example,

that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina ... some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. ... However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA.

...

## Continuing from a Prompt

GPT-3 from Open AI (175 billion parameters, July 2020, ArXiv 2005.14165)

Continue from: Feeling unproductive? Maybe you should stop overthinking.

To In order to get something done, maybe we need to think less. Seems counter-intuitive, but I believe sometimes our thoughts can get in the way of the creative process. We can work better at times when we "tune out" the external world and focus on what's in front of us.

I've been thinking about this lately, so I thought it would be good to write an article about it.

So what exactly does this mean? Well, for starters, let's start with some definitions.

Definition #1: Creative Thinking (CT) is the act of using your mind to come up with ideas that are not already present in reality or that have never existed before. It is a form of mental activity that involves the use of logic and reason.

Definition #2: Over-Thinking (OT) is the act of trying to come up with ideas that have already been thought through by someone else. OT usually results in ideas that are impractical, impossible, or even stupid.

...

## Few Shot Learning

GPT-3 can be prompted with a sequence of German-to-English translation pairs ending in a German sentence without the English translation.

GPT-3 continues with the English Trnanslition of the given German.

The translaion is not close to state of the art but still startling.

Various other tasks can be done with few-shot examples in this way. For example question answering, unscrambling words and three digit arithmetic.

## Few Shot Learners

A very recent paper (Sept. 15, 2020, ArXiv 2009.07118) has shown that the few-shot learning accomplishments of GPT-3 can be done with far fewer parameters.

GPT-3 has 175 billion parameters. The same few-shot performance can be achieved with 223 million parameters — three orders of magnitude fewer.

The new paper uses a method they call PET (Pattern Exploiting Training) involving "distillation" (student-teacher network-to-network training).

# END