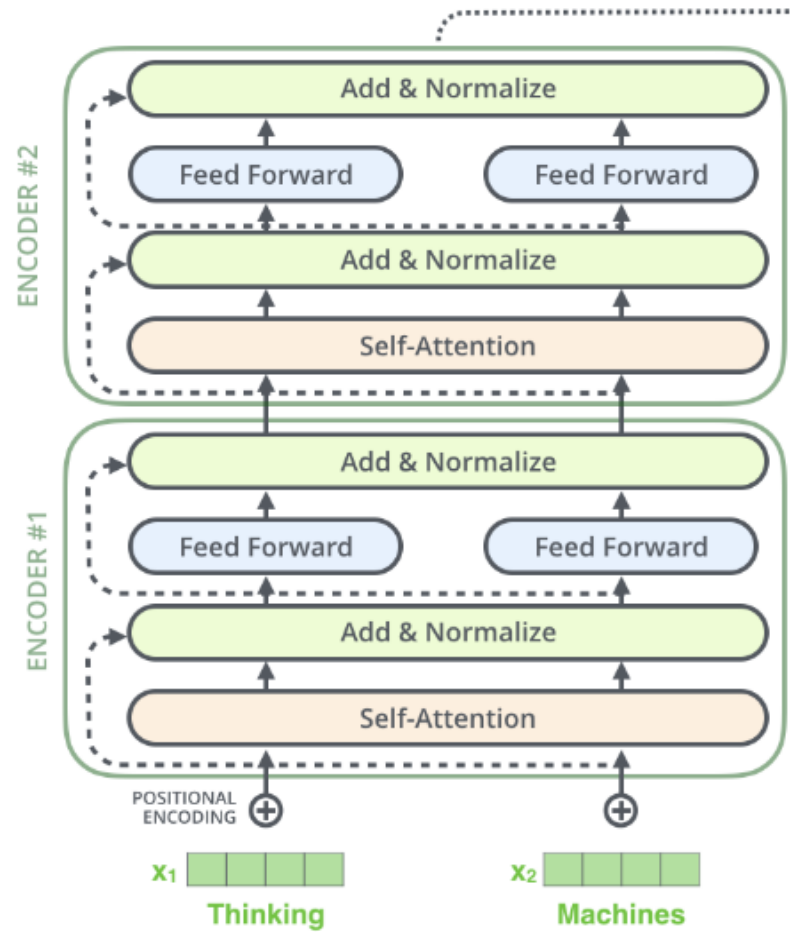


# **TTIC 31230, Fundamentals of Deep Learning**

David McAllester, Autumn 2020

## **The Transformer Part II**

# Transformer Layers



Jay Alammar's blog

## Layer Normalization

The transformer uses layer normalization rather than batch normalization.

$$\mu_\ell = \frac{1}{TJ} \sum_{t,j} L_\ell[t, j]$$

$$\sigma_\ell = \sqrt{\frac{1}{TJ} \sum_{t,j} (L_\ell[t, j] - \mu_\ell)^2}$$

$$\tilde{L}_{\ell+1}[t, j] = \text{ReLU} \left( \frac{A_{\ell+1}[j]}{\sigma_\ell} (L_\ell[t, j] - \mu_\ell) + B_{\ell+1}[j] \right)$$

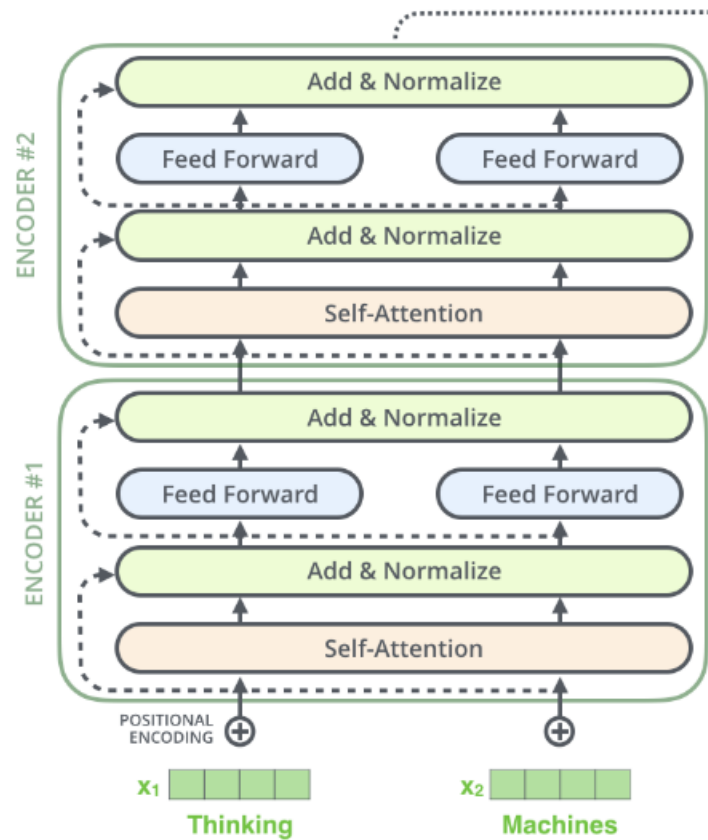
## Feed-Forward Layers

The feed-forward layers apply a two-level multi-layer perceptron (MLP) to the vector at each time position independently.

$$h_{\ell+1}[t, i] = \text{ReLU}(W_{\ell+1}^{\text{FF1}}[i, J] L_{\ell}[t, J] - B_{\ell+1}^{\text{FF1}}[i])$$

$$L_{\ell+1}[t, j] = W_{\ell+1}^{\text{FF2}}[j, I] h_{\ell+1}[t, I] - B_{\ell+1}^{\text{FF2}}[j]$$

# The Transformer



Jay Alammar's blog

## Encoding Positional Information

At the input layer we concatenate the word embeddings with position embeddings.

$$L_0[t, J] = e^W[w[t], I^W]; e^T[t, I^P]$$

Note that without the position embeddings the model has no position information — the input has to be treated as a bag of words.

These days both word embedding and position embeddings are trained as model parameters.

## Language Modeling

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

We can then predict the word  $w_t$  without looking into the future with

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

where  $L_{\text{top}}$  is the top level hidden vector of the Transformer.

# Machine Translation

Translation is just a conditional language model.

This can be done with separate models for the encoder and decoder (an encoder model and a decoder model)

Or it can be done with a **decoder-only** model in which We take the input English sentence followed by a special token and then generate the output from the Transformer language model. In this case causal masking is done on the input sentence as well as the generation of the translation.



## 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.

...

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

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

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

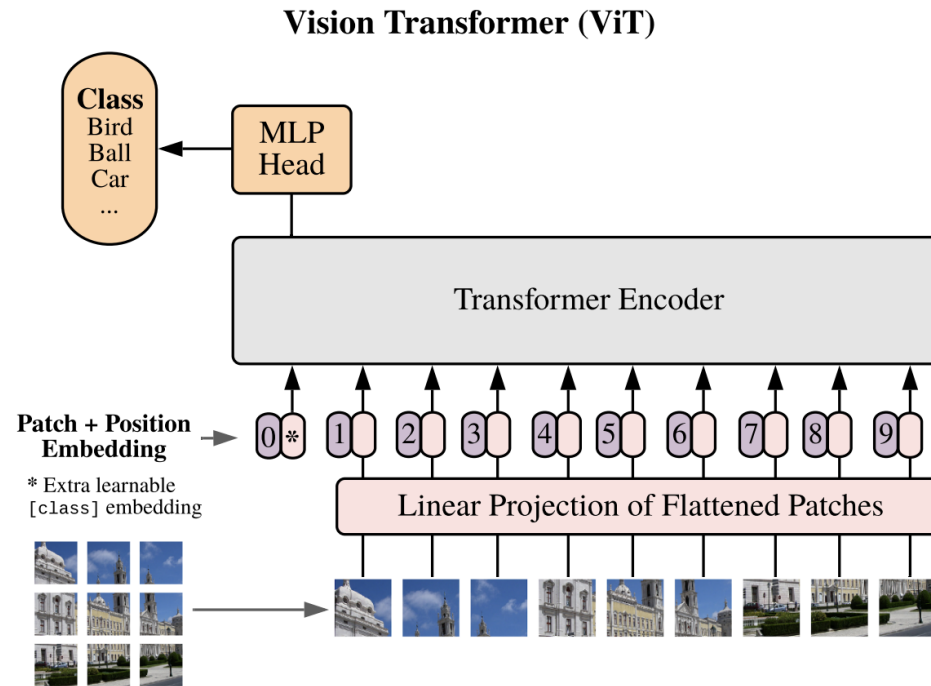
in reality or that have never existed before. It is a form of mental activity that involves the use of logic and reason.

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.

...

# Visual Transformers

**An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale** Alexey Dosovitskiy et al., Google Brain, October 2020.



## 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 Translation of the given German.

The translation 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.

# Chain of Thought Prompting

Chain of Thought Prompting Elicits Reasoning in Large Language Models, Wei et al, Google Brain, January 2022.

## Standard Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. ❌

## Chain of Thought Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅



## Step by Step Prompting, June 2022

### Large Language Models are Zero-Shot Reasoners

Kojima et al, University of Tokyo, June 2022

#### (c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

---

(Output) 8 **X**

## Step by Step Prompting, June 2022

### Large Language Models are Zero-Shot Reasoners

Kojima et al, University of Tokyo, June 2022

#### (d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

---

*(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

**END**