

A Mathematical Introduction to

The Transformer Architecture

Aaron Mazel-Gee

Attention Is All You Need

2017

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

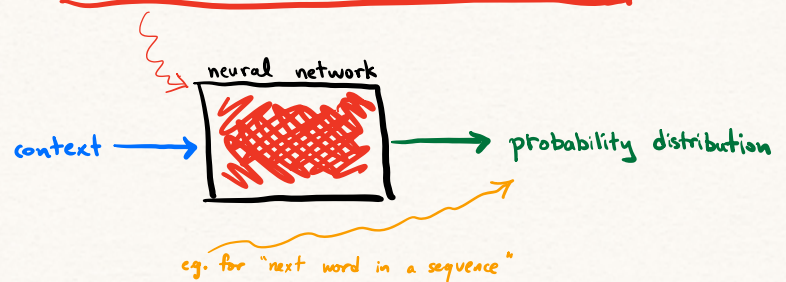
Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

transformer architecture



International Workshop on AI and _____

$P(\text{Topology}) = 0.6$

$P(\text{Ethics}) = 0.2$

$P(\text{Tofu}) = 0.05$

* main ingredient in LLMs, e.g. ChatGPT
GPT = Generative Pretrained Transformer

context

What is the capital of Denmark?

before

What is the capital of Norway?

What is the capital of Finland?

* also: "fine-tuning"

after

The capital of Denmark is Copenhagen.

Vocabulary := set of possible words (or more generally, "tokens")

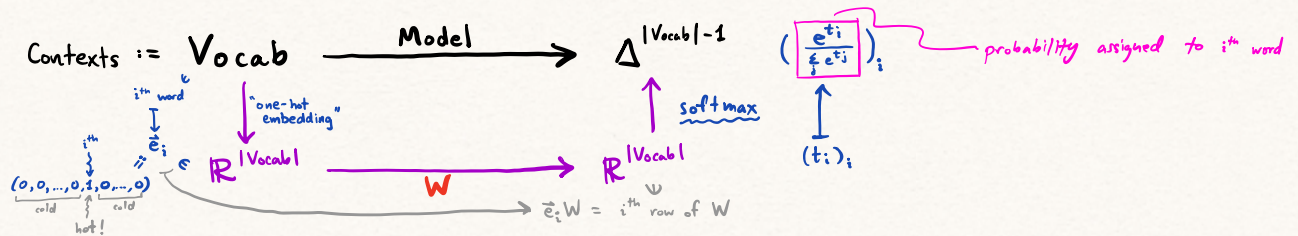
Common sub-word pieces

context, e.g. International Workshop on AI and $\in \text{Vocab}^5$

↳ in general, in Vocab^L for some $L \leq L_{\max}$ (max. poss. context length)

probability distribution, e.g. $\begin{cases} P(\text{Topology}) = 0.6 \\ P(\text{Ethics}) = 0.2 \\ P(\text{Tofu}) = 0.05 \\ \vdots \end{cases} \in \Delta^{|\text{Vocab}|-1} \subset \mathbb{R}^{|\text{Vocab}|}$
 $\Delta^{|\text{Vocab}|-1} := \{(t_i)_i : \sum_i t_i = 1, t_i \geq 0 \forall i\}$

★ a simple neural network, bigram model:



↳ to determine entries of W , maximize likelihood of training data

likelihood (International Workshop on AI and Topology) = $P(\text{Int}^4 | \text{start of message}) \cdot P(\text{Workshop} | \text{Int}^4) \cdot P(\text{on} | \text{Workshop}) \cdot \dots$
 where $P(x | y) := x^{\text{th}}$ coordinate of $\text{Model}(y)$ (also in Vocab...)

★ Guess generalizing to n-gram model: Contexts := $\text{Vocab}^{n-1} \xrightarrow{\text{one-hot}} \mathbb{R}^{(|\text{Vocab}|^{n-1})} \xrightarrow{W} \mathbb{R}^{|\text{Vocab}|} \xrightarrow{\text{softmax}} \Delta^{|\text{Vocab}|-1}$

Problem 1 (sparse data): most elts of Contexts are vanishingly rare in training data.

Problem 2 (combinatorial explosion): W has $|\text{Vocab}|^n$ many entries - waaay too many.

★ solⁿ: embeddings. For $d_{\text{emb}} \ll |\text{Vocab}|^{n-1}$, instead learn a function

Emb(mom) - Emb(woman) $\approx$ Emb(uncle) - Emb(aunt), etc. etc.

Vocab $\xrightarrow{\text{Emb}} \mathbb{R}^{d_{\text{emb}}}$ to optimize the model

Contexts := $\text{Vocab}^{n-1} \xrightarrow{\text{Emb}^{n-1}} (\mathbb{R}^{d_{\text{emb}}})^{n-1} \xrightarrow{\text{concatenate}} \mathbb{R}^{d_{\text{emb}} \cdot (n-1)} \xrightarrow{W} \mathbb{R}^{|\text{Vocab}|} \xrightarrow{\text{softmax}} \Delta^{|\text{Vocab}|-1}$

★ Fancier: more layers, nonlinearity, etc.

multilayer perceptron:

$\mathbb{R}^{d_0} \xrightarrow{(-)W_0 + \vec{b}_0} \mathbb{R}^{d_1} \xrightarrow{\text{ReLU}} \mathbb{R}^{d_1} \xrightarrow{(-)W_1 + \vec{b}_1} \mathbb{R}^{d_2} \xrightarrow{\text{ReLU}} \dots \xrightarrow{\text{ReLU}} \mathbb{R}^{d_L}$

applied coordinate-wise

$\text{ReLU}(x) := \max\{x, 0\}$

Generative Transformer Architecture

① Positional Encoding

Replace $Vocab^L \xrightarrow{Emb^L} (\mathbb{R}^{d_{emb}})^L$
 $(x_1, \dots, x_L) \mapsto (Emb(x_1), \dots, Emb(x_L))$
 with
 $(x_1, \dots, x_L) \mapsto (Emb(x_1) + Pos_1, \dots, Emb(x_L) + Pos_L)$
 for learned vectors $Pos_i \in \mathbb{R}^{d_{emb}}$ for $1 \leq i \leq L_{max}$.
 or specific periodic choices: $Pos_i = (\sin(\frac{i}{10,000 \cdot 2/d_{emb}}), \cos(\frac{i}{10,000 \cdot 2/d_{emb}}), \sin(\frac{i}{10,000 \cdot 1/d_{emb}}), \cos(\frac{i}{10,000 \cdot 1/d_{emb}}), \dots, \sin(\frac{i}{10,000 \cdot 2/d_{emb}}), \cos(\frac{i}{10,000 \cdot 2/d_{emb}}))$

② Add

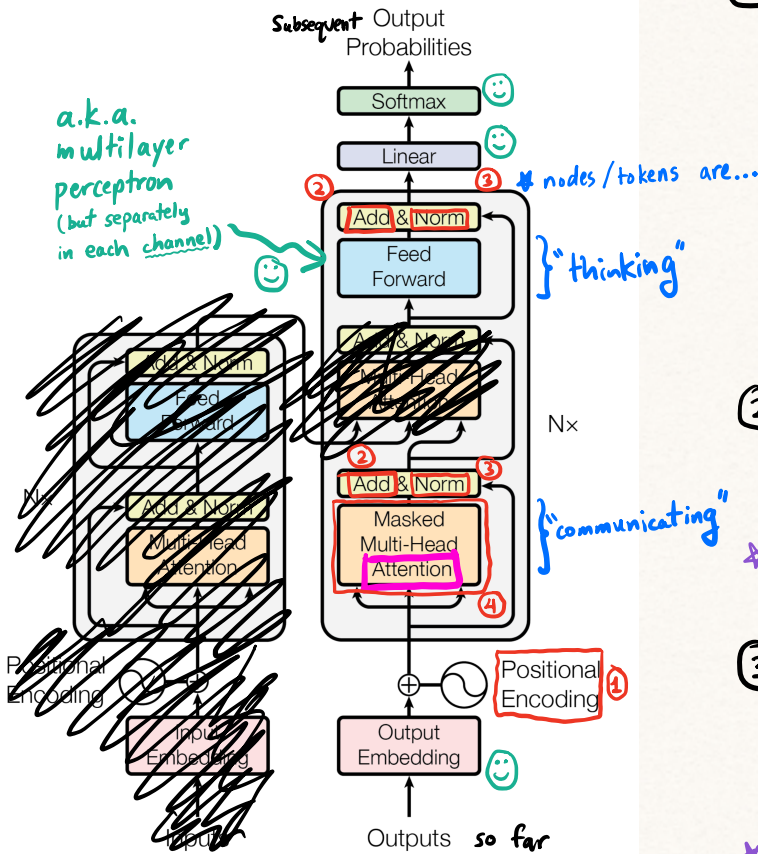
Replace $\begin{matrix} F(x) \\ \uparrow \\ F \\ \uparrow \\ x \end{matrix}$ with $\begin{matrix} F(x) \\ \uparrow \\ F \\ \uparrow \\ x \end{matrix} \oplus x$. "residual connection"

* better for learning: for F "small", perturb id instead of 0.
 (so $\frac{\partial \text{loss}}{\partial \text{parameter}} \approx 1$ instead of 0 at initialization)

③ Norm

"layer normalization"

Replace $y := F(x) + x \in \mathbb{R}^d$ with
 $\hat{y} := (\frac{y_i - \text{mean}_i(y_i)}{\text{stdev}_i(y_i)} \cdot \hat{x}_i + \hat{\mu}_i) \in \mathbb{R}^d$ "gain" "bias"
 for learned $\hat{x}, \hat{\mu} \in \mathbb{R}^d$ initialized as $\hat{x}_i = 1, \hat{\mu}_i = 0$.
 * better for depth: otherwise, very sensitive to initialization.
 (e.g. stdev's may vanish or explode through many layers)

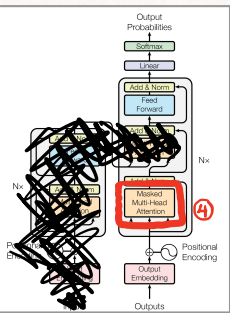


AIAYN, Figure 1

④ Masked Multi-Head Attention

"scaled dot-product"

(a) Single-Head Attention: learn linear maps $\mathbb{R}^{d_{in}} \xrightarrow{\text{Query}} \mathbb{R}^{d_{kt}}$



for any $(x_1, \dots, x_L) \in (\mathbb{R}^{d_{in}})^L$, for each $i \in \{1, \dots, L\}$,
 get relevance scores $(r_{i,1}, \dots, r_{i,L}) := \frac{1}{\sqrt{d_{kt}}} (Q(x_i) \cdot K(x_1), \dots, Q(x_i) \cdot K(x_L))$
 (or use more general (learned) bilinear form)

probability distribution

$$(p_{i,1}, \dots, p_{i,L}) := \text{softmax}(r_{i,1}, \dots, r_{i,L}) = \left(\frac{e^{r_{i,1}}}{\sum_j e^{r_{i,j}}}, \dots, \frac{e^{r_{i,L}}}{\sum_j e^{r_{i,j}}} \right)$$

get attention head

$$(\mathbb{R}^{d_{in}})^L \xrightarrow{\text{Head}_{a,k,v}} (\mathbb{R}^{d_{out}})^L$$

$$(x_i)_i \mapsto \left(\sum_j p_{i,j} \cdot V(x_j) \right)_i$$

(b) Masked: reset relevance scores $r_{i,j} = -\infty$ (so $p_{i,j} = 0$) for $j > i$

"no communication from the future" ("auto-regressive")

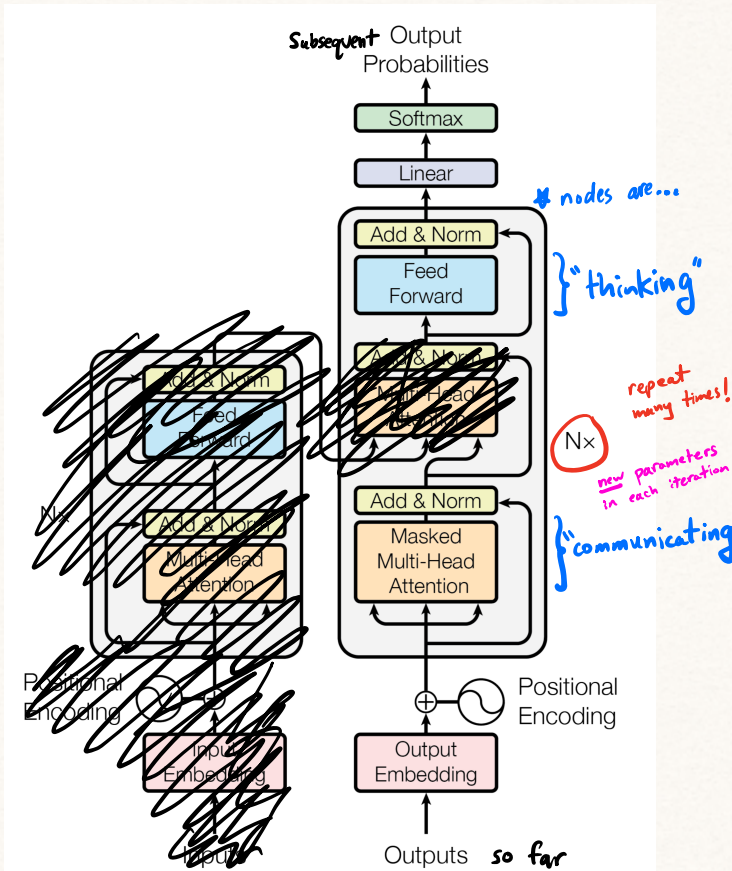


(or use any directed graph)

(c) Multi-Head: use n_{heads} heads $\leadsto (\mathbb{R}^{d_{in}})^L \xrightarrow{(\text{Head}_i)_i} ((\mathbb{R}^{d_{out}})^L)^{n_{heads}} \xrightarrow{\text{concatenate}} (\mathbb{R}^{d_{out} \cdot n_{heads}})^L$
 different notions of "relevance" (e.g. subject $\leadsto$ verb, adjective $\leadsto$ noun, ...)

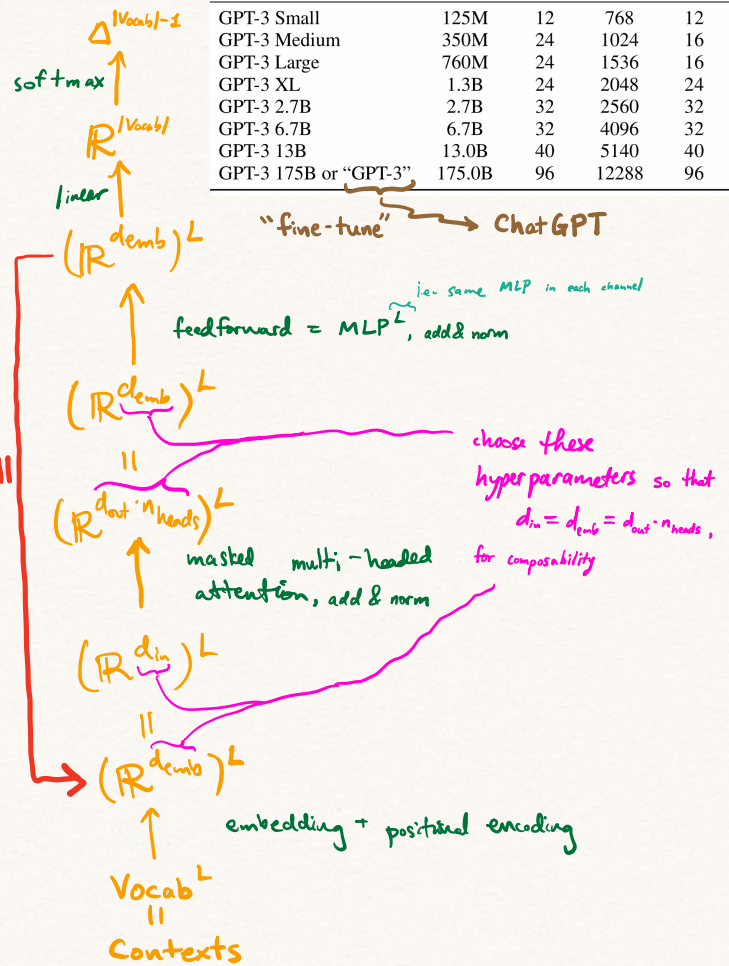
$\leadsto L$ distinct "channels of communication"

Generative Transformer Architecture



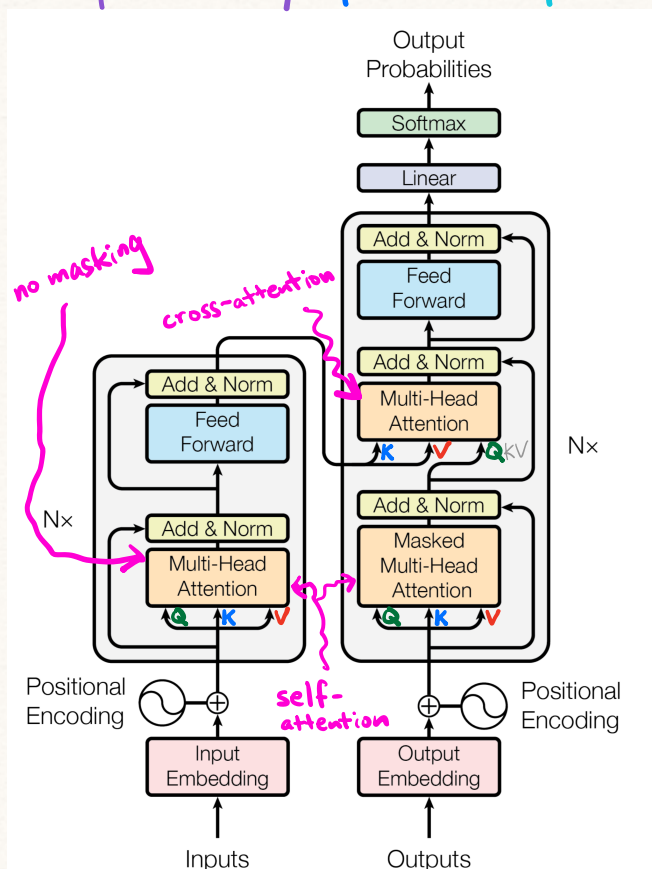
AIAYN, Figure 1

Model Name	n_{params}	m_{layers}	d_{model}	m_{heads}	d_{head}
GPT-3 Small	125M	12	768	12	64
GPT-3 Medium	350M	24	1024	16	64
GPT-3 Large	760M	24	1536	16	96
GPT-3 XL	1.3B	24	2048	24	128
GPT-3 2.7B	2.7B	32	2560	32	80
GPT-3 6.7B	6.7B	32	4096	32	128
GPT-3 13B	13.0B	40	5140	40	128
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128



General Transformer Architecture

encoder decoder



AIAYN, Figure 1

self-attention

decoder-only:

completion (e.g. ChatGPT)

International Workshop on AI and Topology

encoder-only:

classification (of e.g. sentiments or images)



cross-attention

encoder-decoder:

translation

Le singe est dans l'arbre.

The monkey is in the tree.

Le singe est dans l'arbre.

The monkey is in the tree.

Le singe est dans l'arbre.

The monkey is in the tree.

Thanks for listening!

slides available at:

<https://etale.site/ai.html>
