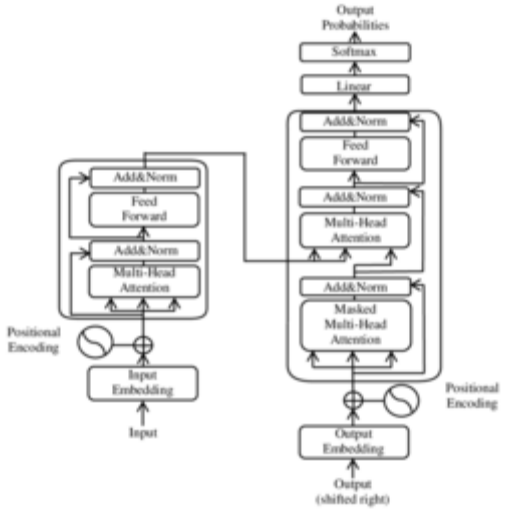
From Wikipedia, the free encyclopedia



An illustration of main components of the transformer model from the paper

"**Attention Is All You Need**"[1] is a 2017 landmark[2][3] research paper in machine learning authored by eight scientists working at Google. The paper introduced a new deep learning architecture known as the transformer, based on the attention mechanism proposed in 2014 by Bahdanau et al.[4] It is considered a foundational[5] paper in modern artificial intelligence, as the transformer approach has become the main architecture of large language models like those based on GPT.[6][7] At the time, the focus of the research was on improving Seq2seq techniques for machine translation, but the authors go further in the paper, foreseeing the technique's potential for other tasks like question answering and what is now known as multimodal Generative AI.[1]

The paper's title is a reference to the song "All You Need Is Love" by the Beatles.[8] The name "Transformer" was picked because Uszkoreit liked the sound of that word.[9]

An early design document was titled "Transformers: Iterative Self-Attention and Processing for Various Tasks", and included an illustration of six characters from the *Transformers* animated show. The team was named Team Transformer.[8]

Some early examples that the team tried their Transformer architecture on included English-to-German translation, generating Wikipedia articles on "The Transformer", and parsing. These convinced the team that the Transformer is a general purpose language model, and not just good for translation.[9]

As of 2024, the paper has been cited more than 100,000 times.[10]

For their 100M-parameter Transformer model, they suggested learning rate should be linearly scaled up from 0 to maximal value for the first part of the training (i.e. 2% of the total number of training steps), and to use dropout, to stabilize training.

Authors

[edit]

The authors of the paper are: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. All eight authors were "equal contributors" to the paper; the listed order was randomized. The *Wired* article highlights the group's diversity:[8]

Six of the eight authors were born outside the United States; the other two are children of two green-card-carrying Germans who were temporarily in California and a first-generation American whose family had fled persecution, respectively.

By 2023, all eight authors had left Google and founded their own AI start-ups (except Łukasz Kaiser, who joined OpenAI).[8][10]

Historical context

[edit]

Main articles: Transformer (deep learning architecture) § History, and Seq2seq § History

See also: Timeline of machine learning

Predecessors

[edit]

For many years, sequence modelling and generation was done by using plain recurrent neural networks (RNNs). A well-cited early example was the Elman network (1990). In theory, the information from one token can propagate arbitrarily far down the sequence, but in practice the vanishing-gradient problem leaves the model's state at the end of a long sentence without precise, extractable information about preceding tokens.

A key breakthrough was LSTM (1995),[note 1] a RNN which used various innovations to overcome the vanishing gradient problem, allowing efficient learning of long-sequence modelling. One key innovation was the use of an attention mechanism which used neurons that multiply the outputs of other neurons, so-called *multiplicative units*.[11] Neural networks using multiplicative units were later called *sigma-pi networks*[12] or *higher-order networks*.[13] LSTM became the standard architecture for long sequence modelling until the 2017 publication of Transformers. However, LSTM still used sequential processing, like most other RNNs.[note 2] Specifically, RNNs operate one token at a time from first to last; they cannot operate in parallel over all tokens in a sequence.

Modern Transformers overcome this problem, but unlike RNNs, they require computation time that is quadratic in the size of the context window. The linearly scaling fast weight controller (1992) learns to compute a weight matrix for further processing depending on the input.[14] One of its two networks has "fast weights" or "dynamic links" (1981).[15][16][17] A slow neural network learns by gradient descent to generate keys and values for computing the weight changes of the fast neural network which computes answers to queries.[14] This was later shown to be equivalent to the unnormalized linear Transformer.[18][19]

Attention with seq2seq

[edit]

Main article: Seq2seq § History

The idea of encoder-decoder sequence transduction had been developed in the early 2010s (see previous papers[20][21]). The papers most commonly cited as the originators that produced seq2seq are two concurrently published papers from 2014.[20][21]

A 380M-parameter model for machine translation uses two long short-term memories (LSTM).[21] Its architecture consists of two parts. The *encoder* is an LSTM that takes in a sequence of tokens and

turns it into a vector. The *decoder* is another LSTM that converts the vector into a sequence of tokens. Similarly, another 130M-parameter model used gated recurrent units (GRU) instead of LSTM.[20] Later research showed that GRUs are neither better nor worse than LSTMs for seq2seq.[22][23]

These early seq2seq models had no attention mechanism, and the state vector is accessible only after the *last* word of the source text was processed. Although in theory such a vector retains the information about the whole original sentence, in practice the information is poorly preserved. This is because the input is processed sequentially by one recurrent network into a *fixed*-size output vector, which is then processed by another recurrent network into an output. If the input is long, then the output vector would not be able to contain all relevant information, degrading the output. As evidence, reversing the input sentence improved seq2seq translation.[24]

The *RNNsearch* model introduced an attention mechanism to seq2seq for machine translation to solve the bottleneck problem (of the *fixed-size* output vector), allowing the model to process long-distance dependencies more easily. The name is because it "emulates searching through a source sentence during decoding a translation".[4]

The relative performances were compared between global (that of *RNNsearch*) and local (sliding window) attention model architectures for machine translation, finding that mixed attention had higher quality than global attention, while local attention reduced translation time.[25]

In 2016, Google Translate was revamped to Google Neural Machine Translation, which replaced the previous model based on statistical machine translation. The new model was a seq2seq model where the encoder and the decoder were both 8 layers of bidirectional LSTM.[26] It took nine months to develop, and it outperformed the statistical approach, which took ten years to develop.[27]

Parallelizing attention

[edit]

Main article: Attention (machine learning) § History

Seq2seq models with attention (including self-attention) still suffered from the same issue with recurrent networks, which is that they are hard to parallelize, which prevented them to be accelerated on GPUs. In 2016, *decomposable attention* applied a self-attention mechanism to feedforward networks, which are easy to parallelize, and achieved SOTA result in textual entailment with an order of magnitude less parameters than LSTMs.[28] One of its authors, Jakob Uszkoreit, suspected that attention *without* recurrence is sufficient for language translation, thus the title "attention is *all* you need".[29] That hypothesis was against conventional wisdom of the time, and even his father, a well-known computational linguist, was skeptical.[29] In the same year, self-attention (called *intra-attention or intra-sentence attention*) was proposed for LSTMs.[30]

In 2017, the original (100M-sized) encoder-decoder transformer model was proposed in the "Attention is all you need" paper. At the time, the focus of the research was on improving seq2seq for machine translation, by removing its recurrence to process all tokens in parallel, but preserving its dot-product attention mechanism to keep its text processing performance.[1] Its parallelizability was an important factor to its widespread use in large neural networks.[31]

AI boom era

[edit]

Already in spring 2017, even before the "Attention is all you need" preprint was published, one of the co-authors applied the "decoder-only" variation of the architecture to generate fictitious Wikipedia articles.[32] Transformer architecture is now used in many generative models that contribute to the ongoing AI boom.

In language modelling, ELMo (2018) was a bi-directional LSTM that produces contextualized word embeddings, improving upon the line of research from bag of words and word2vec. It was followed by BERT (2018), an encoder-only Transformer model.[33] In 2019 October, Google started using BERT to process search queries.[34] In 2020, Google Translate replaced the previous RNN-encoder–RNN-decoder model by a Transformer-encoder–RNN-decoder model.[35]

Starting in 2018, the OpenAI GPT series of decoder-only Transformers became state of the art in natural language generation. In 2022, a chatbot based on GPT-3, ChatGPT, became unexpectedly popular,[36] triggering a boom around large language models.[37][38]

Since 2020, Transformers have been applied in modalities beyond text, including the vision transformer,[39] speech recognition,[40] robotics,[41] and multimodal.[42] The vision transformer, in turn, stimulated new developments in convolutional neural networks.[43] Image and video generators like DALL-E (2021), Stable Diffusion 3 (2024),[44] and Sora (2024), are based on the Transformer architecture.