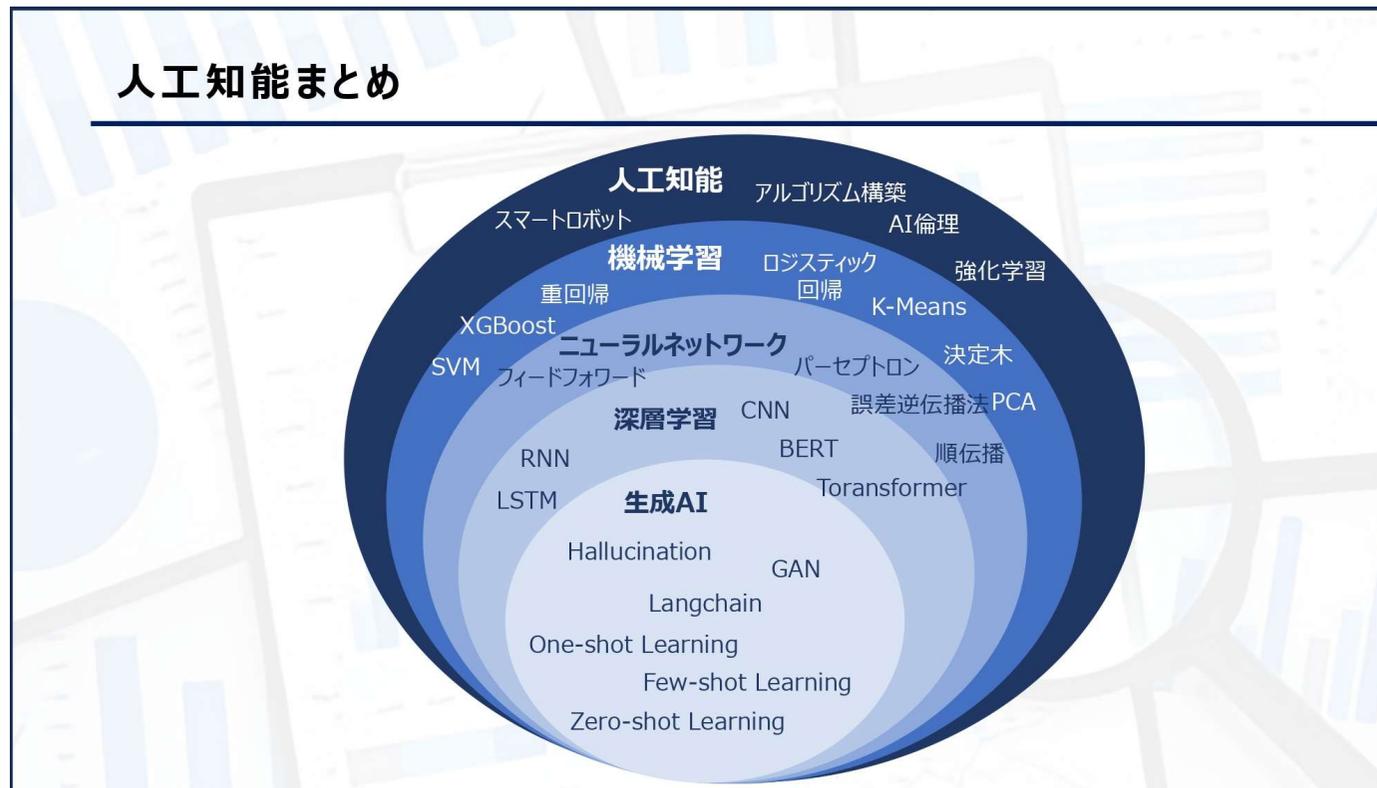


# AI とはなにか？



## Copilot

 Copilot Microsoft Bing Image Creator

「夜の東京タワーとスカイツリーを作成してください。」



## ChatGPT



「50代のイケてる、ヤングのハートをがっちりキャッチする  
ジャージの着こなし方を教えてください。」

こちらが50代男性のイケてるジャージの着こなしイメージです。  
ジャージをスタイリッシュに着こなすポイントが詰まっています。  
自信を持って、ファッションを楽しんでください！

# AIツールについて

## 用途別 本気で使える AI ツール 12 選

No. 1	 <b>Claude</b> 文章作成・推論	最新の Claude3.5 Sonnet は複数の性能で ChatGPT を超え No.1。自然な日本語生成が得意。
No. 2	 <b>Canva</b> デザイン	豊富なテンプレートで、初心者でも美しいグラフィックデザインを簡単に作成可能。
No. 3	 <b>Gamma</b> プレゼン資料	単語を入れるだけで AI が自動でスライドを作成。これまでの資料を作っていた時間から解放。
No. 4	 <b>Perplexity</b> 市場調査	複数の言語モデルを使用可能な高度な検索エンジン。最新情報を元に包括的な市場分析を提供。
No. 5	 <b>tl; dv</b> 議事録	会議録音から自動で議事録を生成。重要ポイントを抽出し、会議の振り返りの時間を削減。
No. 6	 <b>Coze</b> AIチャットボット	お手軽にカスタマイズ可能なチャットボットを作成。顧客サポートや情報提供などに対応。
No. 7	 <b>Dify</b> AIアプリ開発	ノーコードで AI エージェントを開発。専門知識不要で独自の AI エージェントを素早く構築可能。
No. 8	 <b>Midjourney</b> 画像生成	テキスト入力から高品質な画像を生成。アート作品やビジュアル素材の制作に活用可能。
No. 9	 <b>Mapify</b> マインドマップ	AI サポートでマインドマップを作成。アイデアを網羅的に整理し、効率的な思考整理を支援。
No. 10	 <b>Notion</b> タスク管理	柔軟なワークスペースでタスク管理と情報整理。AI 支援の文書作成機能で生産性向上をサポート。
No. 11	 <b>Webサイト Create.xyz</b>	AI で Web サイトを簡単作成。デザインからコンテンツ作成まで、美しいサイトを素早く構築。
No. 12	 <b>Chat GPT</b> URL 解析	WebPilot 機能で指定 URL の内容を分析し、要約や重要ポイントを抽出。

## ChatGPT 以外にも活用して！ 2024 年 タスク別効率上昇 AI ツール 8 選

POINT 1	 <b>Claude / 文書作成</b>	Claude は高度な自然言語処理技術を活用した文書作成・推論 AI ツールです。ユーザーの入力に基づき、文章の生成、要約、分析、質問応答などを行います。複雑な概念の説明や創造的な文章作成にも対応し、様々な分野での文書作成作業を効率化します。また、論理的推論能力を活かして、問題解決や意思決定のサポートも可能です。
POINT 2	 <b>Gemini / 分析</b>	Gemini は、Google AI が開発した高性能な分析ツールです。複雑なデータセットを処理し、深い洞察を導き出す能力に優れています。自然言語理解と多様なデータタイプの統合分析が可能で、ビジネス戦略立案やトレンド予測、科学研究などの分野で活用できます。高度な機械学習アルゴリズムにより、精度の高い分析結果を提供が可能です。
POINT 3	 <b>Perplexity / AI 検索エンジン</b>	Perplexity は、AI を活用した次世代の検索エンジンです。従来の検索エンジンとは異なり、自然言語理解と機械学習を駆使して、ユーザーの真意を深く理解し、関連性の高い情報を効率的に提供します。複雑な質問にも適切に対応し、信頼性の高い情報源から最新のデータを集約して回答を生成します。さらに、対話形式で情報を掘り下げることができ、より詳細で正確な検索結果を得られます。
POINT 4	 <b>ChatGPT-4o / テンプレート作成</b>	ChatGPT-4o は、高度な言語モデルを活用し、ビジネス文書、プレゼンテーション資料、メールなど、様々な用途に応じたテンプレートを効率的に生成することが可能です。ユーザーの要望や業界特有の要件を理解し、カスタマイズ可能な柔軟なテンプレートを提案できます。また、一貫性のある文体や構成を維持しながら、創造的な要素も取り入れることができ、プロフェッショナルな文書作成をサポートします。
POINT 5	 <b>Notion / タスク管理</b>	Notion は、柔軟性と多機能性に優れたタスク管理ツールです。プロジェクト計画、To-Do リスト、スケジュール管理などを一元化し、直感的なインターフェースで効率的な作業を管理できます。カスタマイズ可能なデータベースやカンバンボード、ガントチャートなどの機能を活用し、個人やチームのタスクを視覚的に管理。AI を活用した自動化機能により、タスクの優先順位付けや進捗管理を最適化し、生産性を向上させます。
POINT 6	 <b>Canva / デザイン</b>	Canva は、直感的で使いやすいオンラインデザインツールです。プロ級のグラフィックデザインを、デザインスキルがない人でも簡単に作成できます。豊富なテンプレート、画像、フォント、グラフィック要素が簡単に編集可能で、実に様々な用途に対応できます。AI 機能を活用して、デザインの提案や画像生成、レイアウトの最適化を行い、クリエイティブな高品質なデジタルコンテンツを効率的に制作することも可能です。
POINT 7	 <b>Dify / アプリ開発</b>	Dify は、AI を活用した効率的なアプリケーション開発プラットフォームです。コーディングの専門知識が少ない人でも、直感的なインターフェースを通じて AI アプリケーションを迅速に構築できます。自然言語処理や機械学習モデルの統合が容易で、チャットボット、データ分析ツール、予測モデルなど、様々な AI アプリケーションの開発に対応しています。
POINT 8	 <b>Create.xyz / web サイト構築</b>	Create.xyz は、AI を活用した革新的なウェブサイト構築ツールです。ユーザーの要望や業界に合わせた、自動的にカスタマイズされたウェブデザインを生成します。コーディング不要で、直感的なドラッグ&ドロップインターフェースを使用して簡単に編集可能。レスポンシブデザイン、SEO 最適化、高速読み込みなどの現代的な要件に自動対応し、プロフェッショナルな外観と機能性を両立。さらに、AI がコンテンツの提案や最適化を行い、魅力的で効果的なウェブサイトを短時間で構築できます。

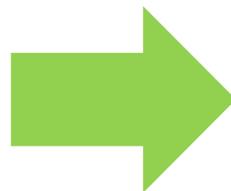
## 技術革新のイメージ（産業革命と現代）

19世紀～



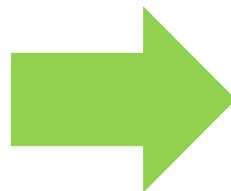
## 技術革新のイメージ（産業革命と現代）

19世紀～

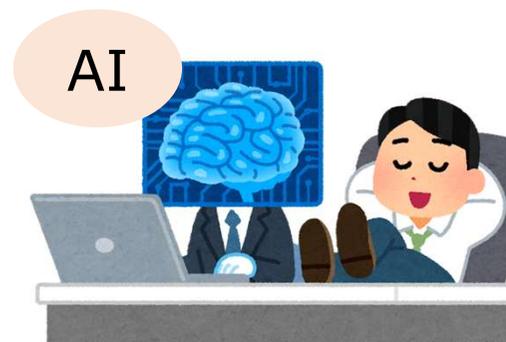
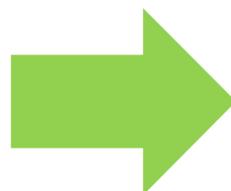


## 技術革新のイメージ（産業革命と現代）

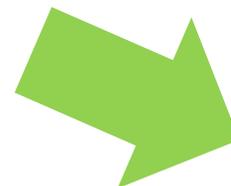
19世紀～



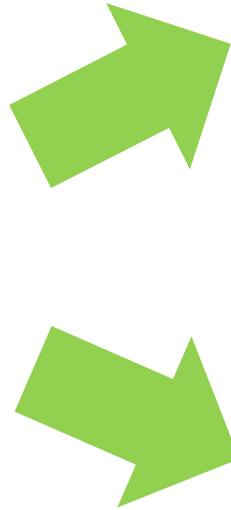
現代～



## 自動車を運転して・・・



## 自動車を運転して…



# 自動車を運転して...

ツールを中心に使いたい...



**Attention Is All You Need**

Ashish Vaswani* Google Brain avaswani@google.com	Noam Shazeer* Google Brain noam@google.com	Niki Parmar* Google Research niki@google.com	Jakob Uszkoreit* Google Research uszkoreit@google.com
Llion Jones* Google Research llion@google.com	Aidan N. Gomez* <sup>1</sup> University of Toronto aidan@cs.toronto.edu	Lukasz Kaiser* Google Brain lukaszkaiser@google.com	
Bia Babcock* <sup>1</sup> 11114.pil@khi18gsm11.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 nine 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 sequence-to-sequence models, 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 consistency parsing both with large and limited training data.

<sup>\*</sup>Equal contribution. Llion Jones is student. Jakob proposed parallel RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Biao, developed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representations and became the other person involved in nearly every detail. Niki developed, implemented, tested and evaluated cross-encoder models separate to our original codebase and auto-Dataset. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualization. Lukasz and Aidan spent countless long days debugging various parts of and improving seq2seq/Decoder, replacing our earlier codebase, greatly improving results and massively accelerating our research.

<sup>1</sup>Work performed while at Google Brain.

<sup>2</sup>Work performed while at Google Research.

仕組みを知っておくべきか...

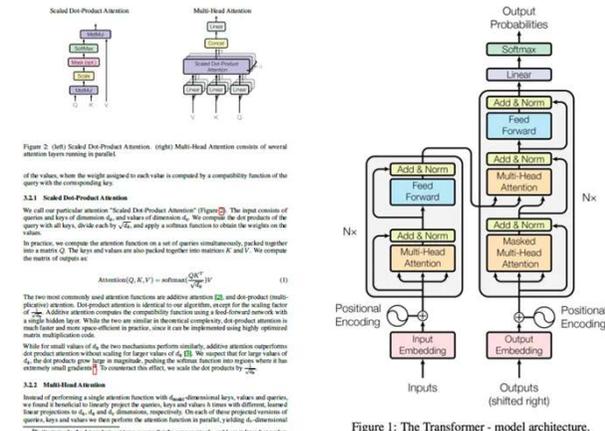


Figure 1: The Transformer - model architecture.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

## 生成AIで『便利になった』ということ

最近、知人の経営者から聞いた。

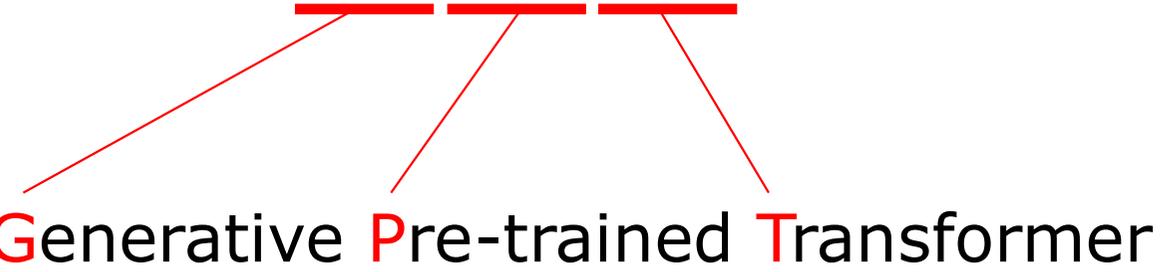
「生成AIで『便利になった』というのは、電子レンジができて『便利になった』と同じこと」

という言葉は、とても的を得ていると思った。(竹内謙礼)

GPT

ChatGPT

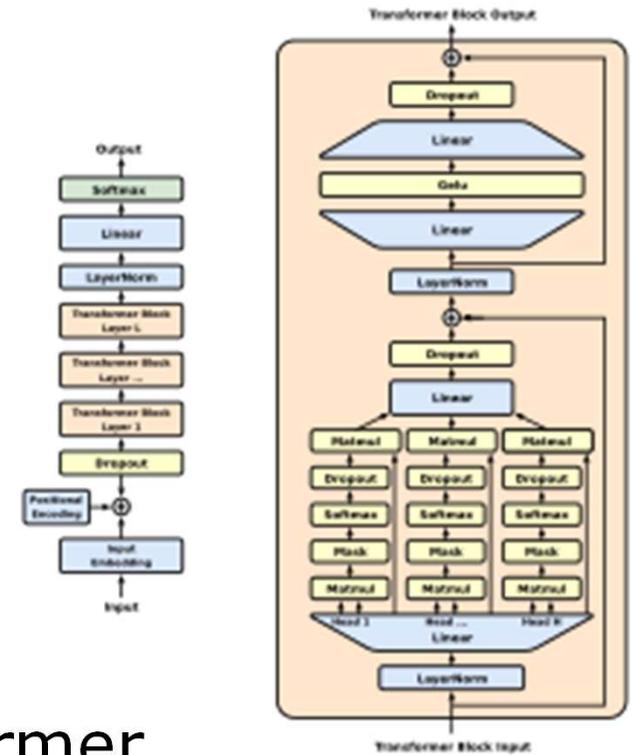
# ChatGPT



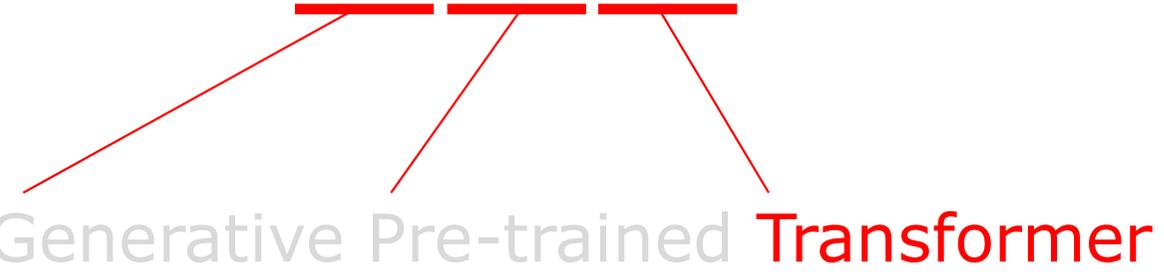
Generative Pre-trained Transformer

# ChatGPT

Generative Pre-trained Transformer



# ChatGPT



Generative Pre-trained Transformer

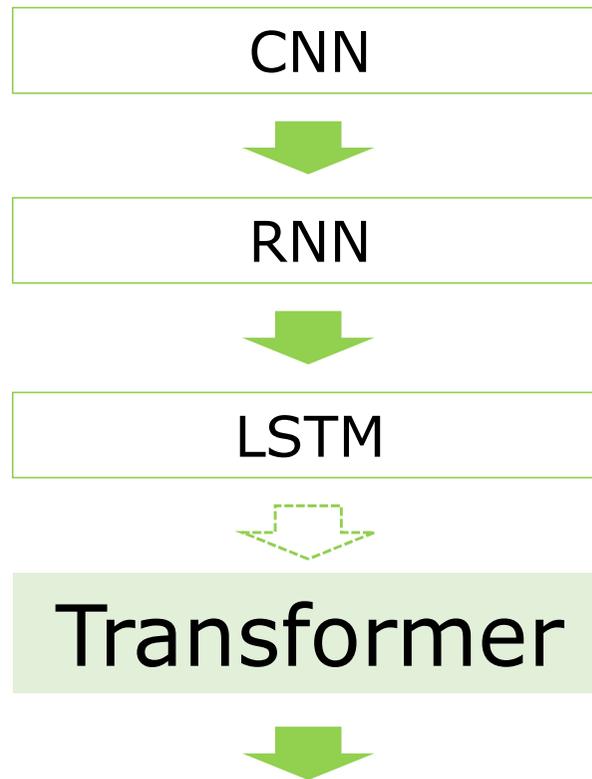
Mapify × Gamma

# Transformerの概要

Transformerは、自然言語処理や機械翻訳に使用される深層学習モデルです。自己注意機構を利用して、入力データの異なる部分間の関係を学習します。

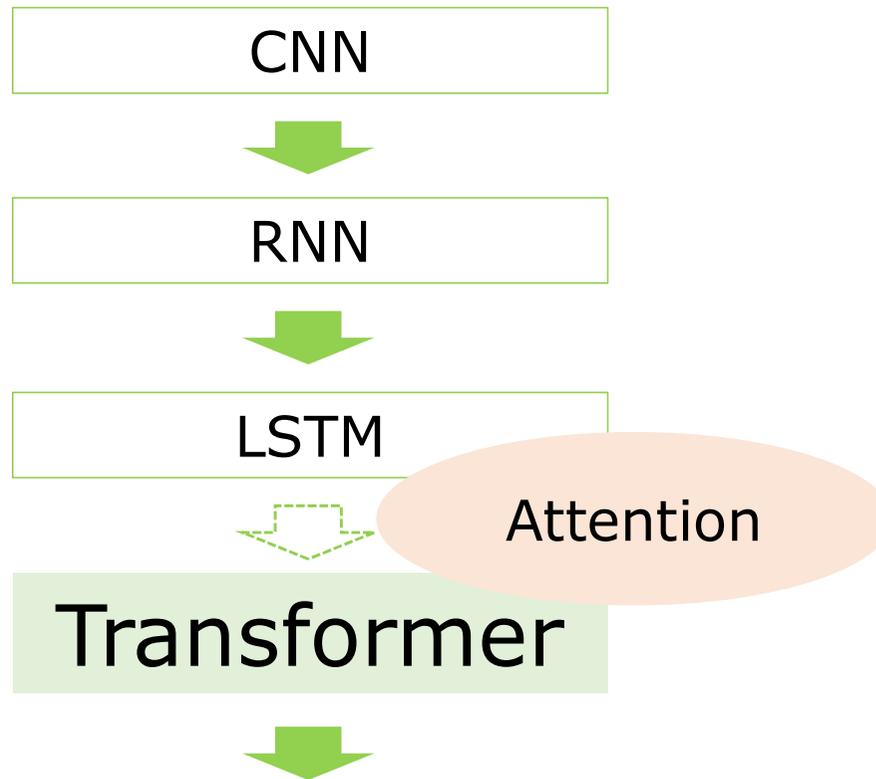


## Transformerへのアーキテクチャー



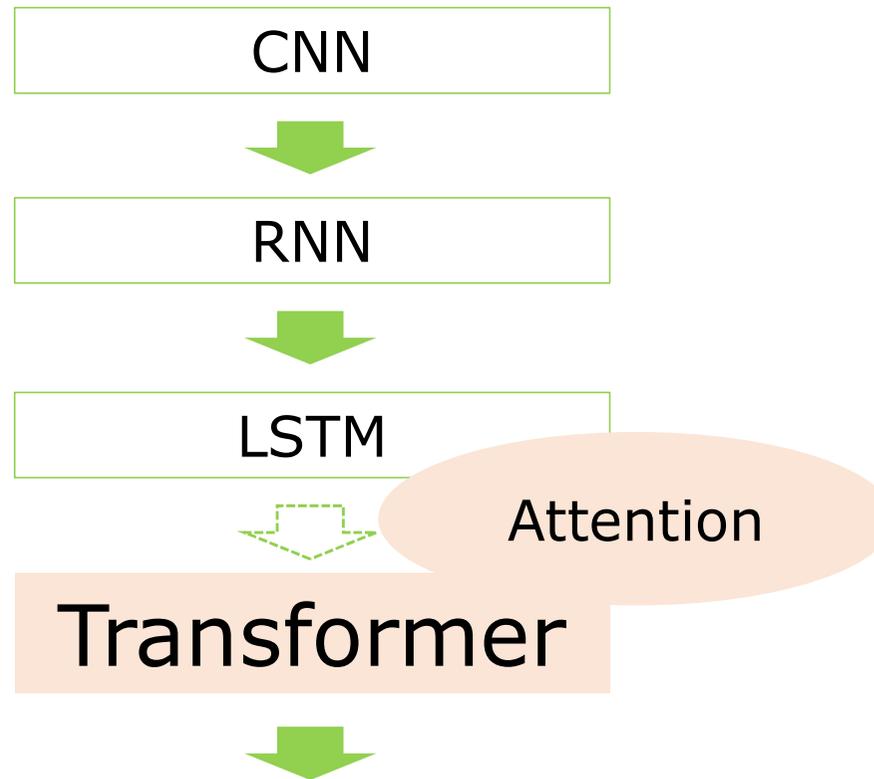
GPT : Generative Pre-trained Transformer

## Transformerへのアーキテクチャー



GPT : Generative Pre-trained Transformer

## Transformerへのアーキテクチャー



GPT : Generative Pre-trained Transformer

# Transformer



## "Attention Is All You Need" (Vaswani et al, 2017)

<https://arxiv.org/pdf/1706.03762>

**Attention Is All You Need**

---

<p><b>Ashish Vaswani*</b> Google Brain avaswani@google.com</p>	<p><b>Noam Shazeer*</b> Google Brain noam@google.com</p>	<p><b>Niki Parmar*</b> Google Research nikip@google.com</p>	<p><b>Jakob Uszkoreit*</b> Google Research usz@google.com</p>
<p><b>Llion Jones*</b> Google Research llion@google.com</p>	<p><b>Aidan N. Gomez†</b> University of Toronto aidan@cs.toronto.edu</p>	<p><b>Lukasz Kaiser*</b> Google Brain lukasz.kaiser@google.com</p>	
<p><b>Iliia Polosukhin*†</b> iliia.polosukhin@gmail.com</p>			

**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.

\*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Iliia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.  
‡Work performed while at Google Research.

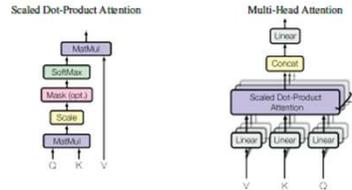


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

**3.2.1 Scaled Dot-Product Attention**

We call our particular attention "Scaled Dot-Product Attention" (Figure 2). The input consists of queries and keys of dimension  $d_k$ , and values of dimension  $d_v$ . We compute the dot products of the query with all keys, divide each by  $\sqrt{d_k}$ , and apply a softmax function to obtain the weights on the values.

In practice, we compute the attention function on a set of queries simultaneously, packed together into a matrix  $Q$ . The keys and values are also packed together into matrices  $K$  and  $V$ . We compute the matrix of outputs as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The two most commonly used attention functions are additive attention [2], and dot-product (multiplicative) attention. Dot-product attention is identical to our algorithm, except for the scaling factor of  $\frac{1}{\sqrt{d_k}}$ . Additive attention computes the compatibility function using a feed-forward network with a single hidden layer. While the two are similar in theoretical complexity, dot-product attention is much faster and more space-efficient in practice, since it can be implemented using highly optimized matrix multiplication code.

While for small values of  $d_k$  the two mechanisms perform similarly, additive attention outperforms dot product attention without scaling for larger values of  $d_k$  [3]. We suspect that for large values of  $d_k$ , the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients [4]. To counteract this effect, we scale the dot products by  $\frac{1}{\sqrt{d_k}}$ .

**3.2.2 Multi-Head Attention**

Instead of performing a single attention function with  $d_{model}$ -dimensional keys, values and queries, we found it beneficial to linearly project the queries, keys and values  $h$  times with different, learned linear projections to  $d_k$ ,  $d_v$  and  $d_q$  dimensions, respectively. On each of these projected versions of queries, keys and values we then perform the attention function in parallel, yielding  $d_h$ -dimensional

<sup>4</sup>To illustrate why the dot products get large, assume that the components of  $q$  and  $k$  are independent random variables with mean 0 and variance 1. Then their dot product,  $q \cdot k = \sum_{i=1}^{d_k} q_i k_i$ , has mean 0 and variance  $d_k$ .

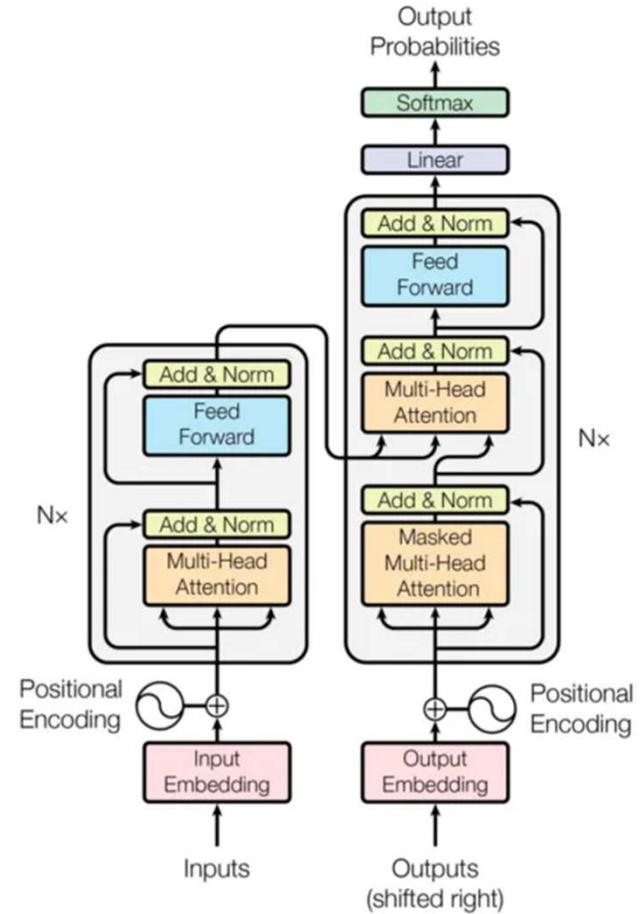


Figure 1: The Transformer - model architecture.

## LSTM (Long Short-Term Memory) の拡張版

### Transformerを性能で凌駕、AIの新たな可能性を拓く5月の注目論文

野々村 泰吾 AI・データラボ、浅川 森輝 クロスメディア編集部/AI・データラボ

2024.06.07  
有料会員限定



全1991文字

生成AI（人工知能）を含む最新のAI研究動向を知るため、世界中の研究者やエンジニアが参照しているのが、論文運搬サイト「arXiv（アーカイブ）」である。米OpenAI（オープンAI）や米Google（グーグル）などAI開発を主導するIT企業の多くが、研究成果をarXivに競って投稿している。

そんなarXivの投稿論文から、2024年5月（1日～31日）にSNSのX（旧Twitter）で多く言及されたAI分野の注目論文を紹介する。調査には米Meltwater（メルトウォーター）のSNS分析ツールを利用した。対象はXの全世界のオリジナル投稿、コメント、再投稿、引用投稿である。調査は、日経BPが2024年1月に新設したAI・データラボの活動の一環として実施した。

#### Transformer並みの拡張性をLSTMで実現

5月に最も多く言及された論文は、オーストリアの研究チームが発表した「xLSTM: Extended Long Short-Term Memory」である。深層学習ベースの言語モデルのアーキテクチャーであるLSTM（Long Short-Term Memory）を改良し、数十億パラメーターの言語モデルにおいてTransformer並みかそれ以上の拡張性（スケーラビリティ）を持たせたという。

かつてLSTMは高性能な言語モデルを実現する技術として一世を風靡したが、複数のGPUによる並列処理が難しく、性能を拡張しづらい弱点があった。その後、並列処理が可能なTransformerが台頭し、ChatGPTをはじめとする現在の生成AIの発展につながった。今回、拡張性を高めたLSTMが登場したことで、従来とは異なる特性を備えた大規模言語モデル（LLM）が生まれる可能性がある。

関連論文

<https://arxiv.org/abs/2405.04517>

日経 **XTECH**

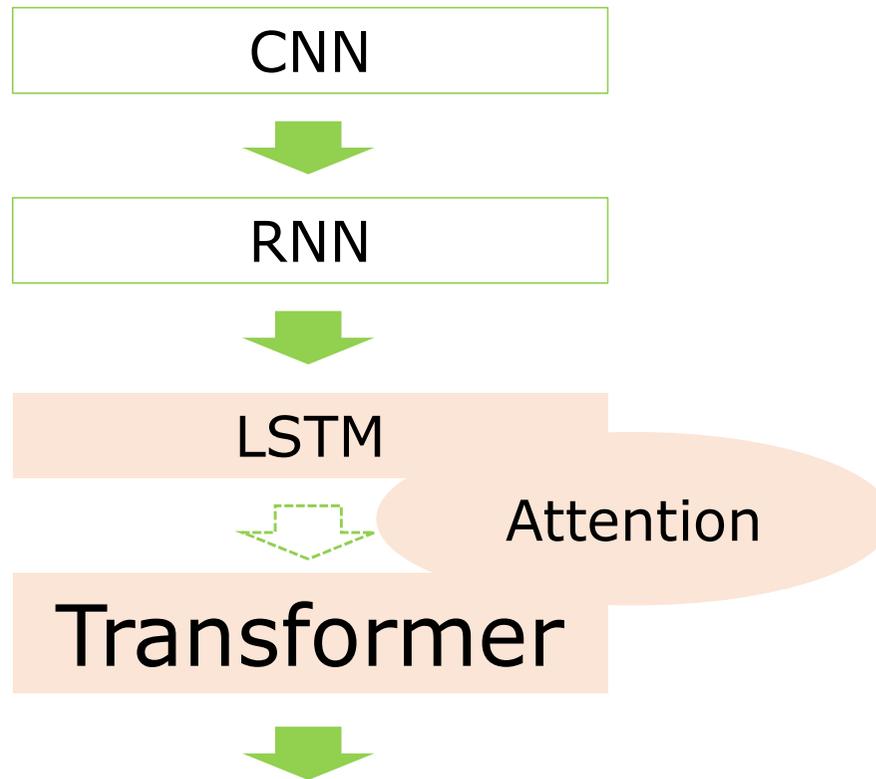
<https://xtech.nikkei.com/atcl/nxt/column/18/02801/060400005/>



<https://arxiv.org/abs/2405.04517>

- 深層学習ベースの言語モデルの**アーキテクチャー**である**LSTM（Long Short-Term Memory）**を改良。（2024年6月現在）
- 数十億パラメーターの言語モデルにおいて**Transformer**並みかそれ以上の拡張性（スケーラビリティ）を持たせた。
- LSTMは高性能な言語モデルを実現する技術として一世を風靡したが、複数のGPUによる**並列処理**が難しく、性能を拡張しづらい弱点があった。
- その後、並列処理が可能な**Transformer**が台頭、**ChatGPT**をはじめとする現在の**生成AI**の発展につながった。
- 今回、拡張性を高めたLSTMが登場したことで、従来とは異なる特性を備えた**大規模言語モデル（LLM）**が生まれる可能性がある。

## Transformerへのアーキテクチャー



GPT : Generative Pre-trained Transformer

## Diffusion model (拡散モデル)

### Diffusionモデルとは？

#### 概要

Diffusionモデルは、画像生成AIサービスで利用される生成モデルの一つ。

スコアベースや拡散確率モデルに分類され、画像生成タスクにおいて高精度を実現。

2015年に提案され、2020年に改良版が発表された。

#### 主なサービス

Stable Diffusion、DALL・E2、GLIDE、GoogleのImagenなどがこのモデルを採用。

テキストから画像を生成する能力を持つ。

### Diffusionモデルの仕組み

#### Forward Process

元の画像にノイズを加え、最終的にノイズだけに変換するプロセス。

ガウスノイズを少しずつ加え、ガウス分布を得る。

ステップごとのパラメータ学習が不要で、アーキテクチャがシンプル。

#### Reverse Process

Forward Processの逆で、ガウス分布からノイズを除去して画像を生成。

条件付き確率を推定するモデルを学習する必要がある。

実装は比較的シンプルなコードで可能。

## Diffusion model（拡散モデル）

### DiffusionモデルとVAEの違い

#### VAEの概要

VAEはオートエンコーダの一種で、入力とは異なる出力を生成。

潜在変数モデルにおけるモデルエビデンスの推論に使用される。

#### Diffusionモデルとの比較

Diffusionモデルはエンコード側の学習パラメータが存在しない。

生成プロセスはVAEと似ているが、より高品質な画像生成が可能。

### DiffusionモデルとGANの違い

#### GANの概要

GANは本物のデータと間違われるデータを生成する手法。

2種類のモデルを対峙させることで学習を進める。

#### Diffusionモデルとの比較

GANは学習の不安定さや多様性の欠如が課題。

Diffusionモデルはより安定した生成が可能。

## Diffusion model（拡散モデル）

### DiffusionモデルとFlow-based modelsの違い

#### Flow-based modelsの概要

潜在変数からデータを生成する深層生成モデルの一種。

可逆的な関数でなければならず、モデルの表現力が制限される。

#### Diffusionモデルとの比較

Diffusionモデルは確率微分方程式でモデル化される。

ノイズを少しずつ加え、実データに近づける生成過程を持つ。

### Diffusionモデルが使用されているAIサービス

#### Stable Diffusion

ユーザーがテキストを入力することで高品質な画像を生成。

オープンソースであり、誰でも利用可能。

#### DALL・E2

OpenAIが開発した画像生成AIツール。

クリエイティビティに優れ、テキストを基に画像を作成。

## Diffusion model（拡散モデル）

### Diffusionモデルの実装方法

PyTorchによる実装

PyTorchはPython向けのオープンソース機械学習ライブラリ。

複雑なネットワークの実装が容易で、効率的な開発が可能。

開発の外部委託

自社での開発は時間とコストがかかるため、専門企業に依頼することが推奨される。

## LSTM (Long Short-Term Memory)