



集積回路視点による 生成AI

情報統合技術研究合同会社

岡島 義憲

okajima@info-integnology.com

自己紹介：岡島義憲

■ 経歴

- ・ 1981年～2019年 富士通半導体設計部門
Bipolar Memory、SRAM、DRAM、Micro Controller、ASIC、ASSP
2015年～2019年は、SIRIJ、JEITA、SEAJ等、経産省と半導体業界団体
- ・ 2021年4月 情報統合技術研究合同会社を設立
(AI回路の知財開発、ロジック系集積回路設計への問題提起)
(<https://info-integnology.com/index.html>)
- ・ 2015年から、汎用AI開発に興味を持つ。
 - 今は、AI回路をGPUを使って回路シミュレーションしている段階だが、設計が完了すると、次は、その回路を搭載したチップ開発の時代が来る。
 - 生成AI や GPTは、想定外の進展
 - Transformerは、AI回路の普遍論理(Universal Logic) だろう。

普遍論理(Universal Logic)としてのTransformerの登場

(注) 圏論/カテゴリー論による

「普遍論理」とは、

- ・ 数値に関する論理 ($1+1=2$ 、等) : デジタルな論理
- ・ 情報に関する論理 (意味理解と情報表現のルール、等) : 確率的論理

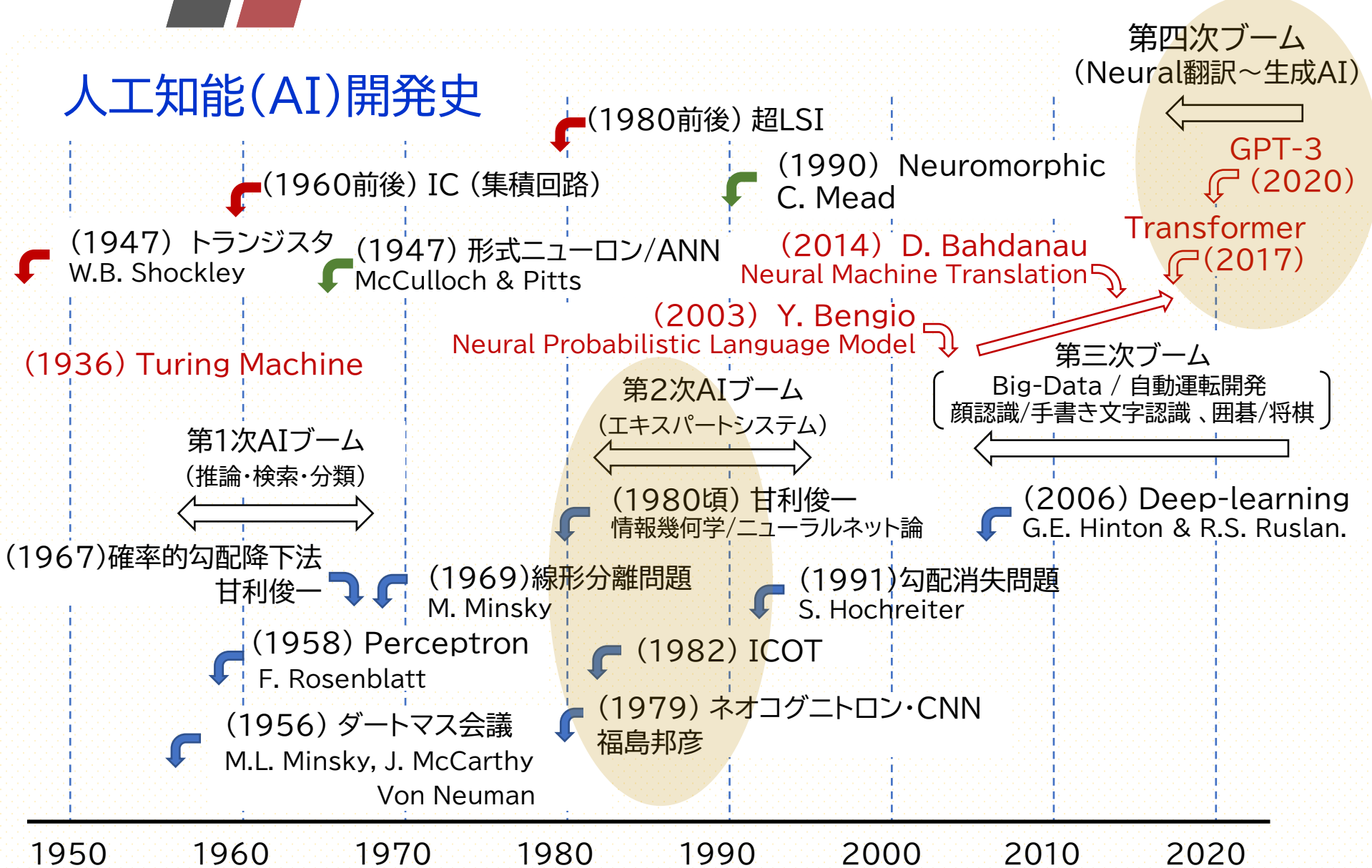
からなり、自然界(人間含む)のダイナミズムを表現する時に用いる「演算」のこと

- ・ 素な普遍論理 : NAND、NOR、INV、メモリ …
- ・ 普遍論理の表現プラットフォーム回路 : CPU、FPGA、 …

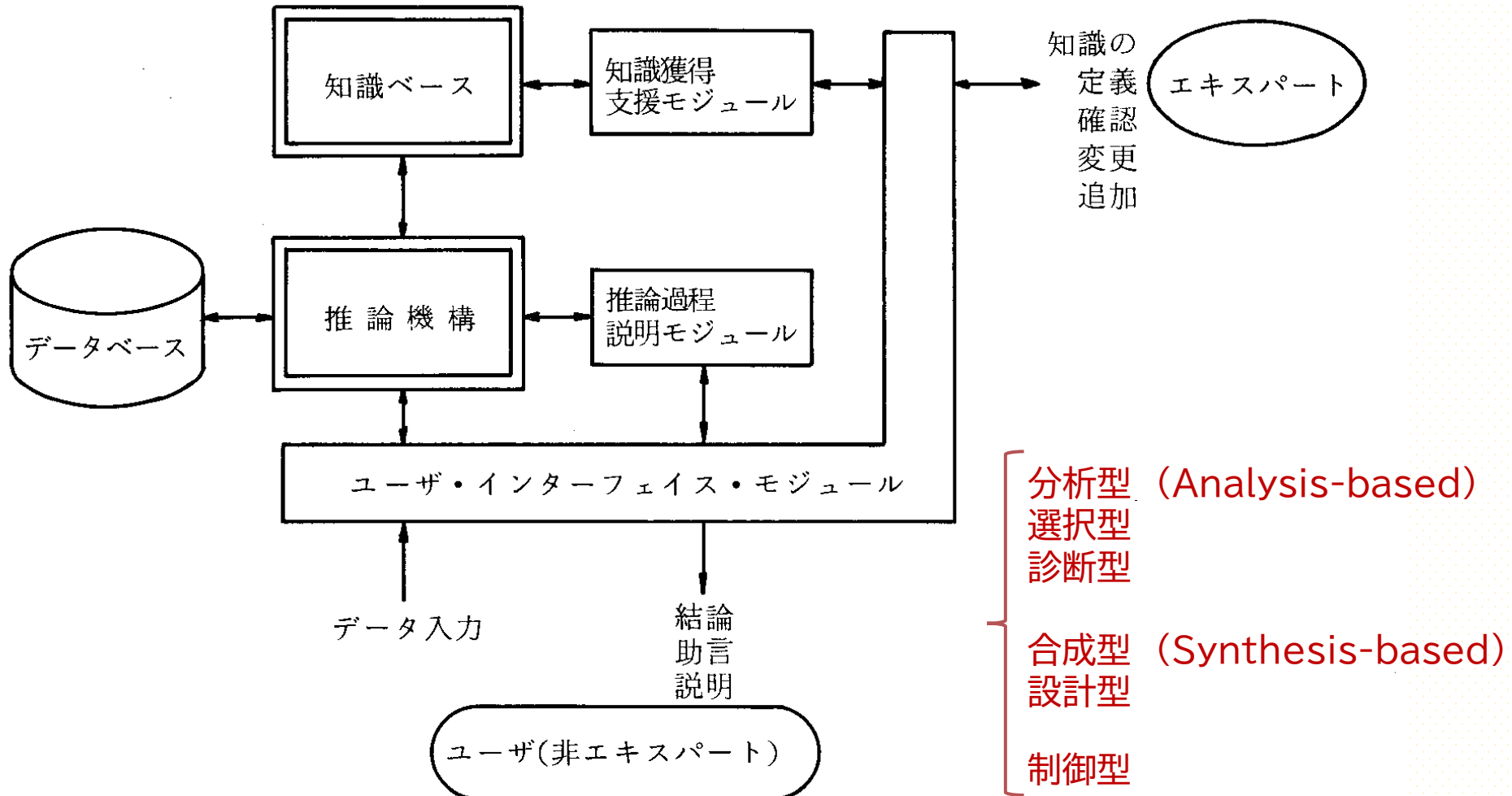
ニューラルネットワーク
の論理って、何なんだ？

Transformerも追
加される？

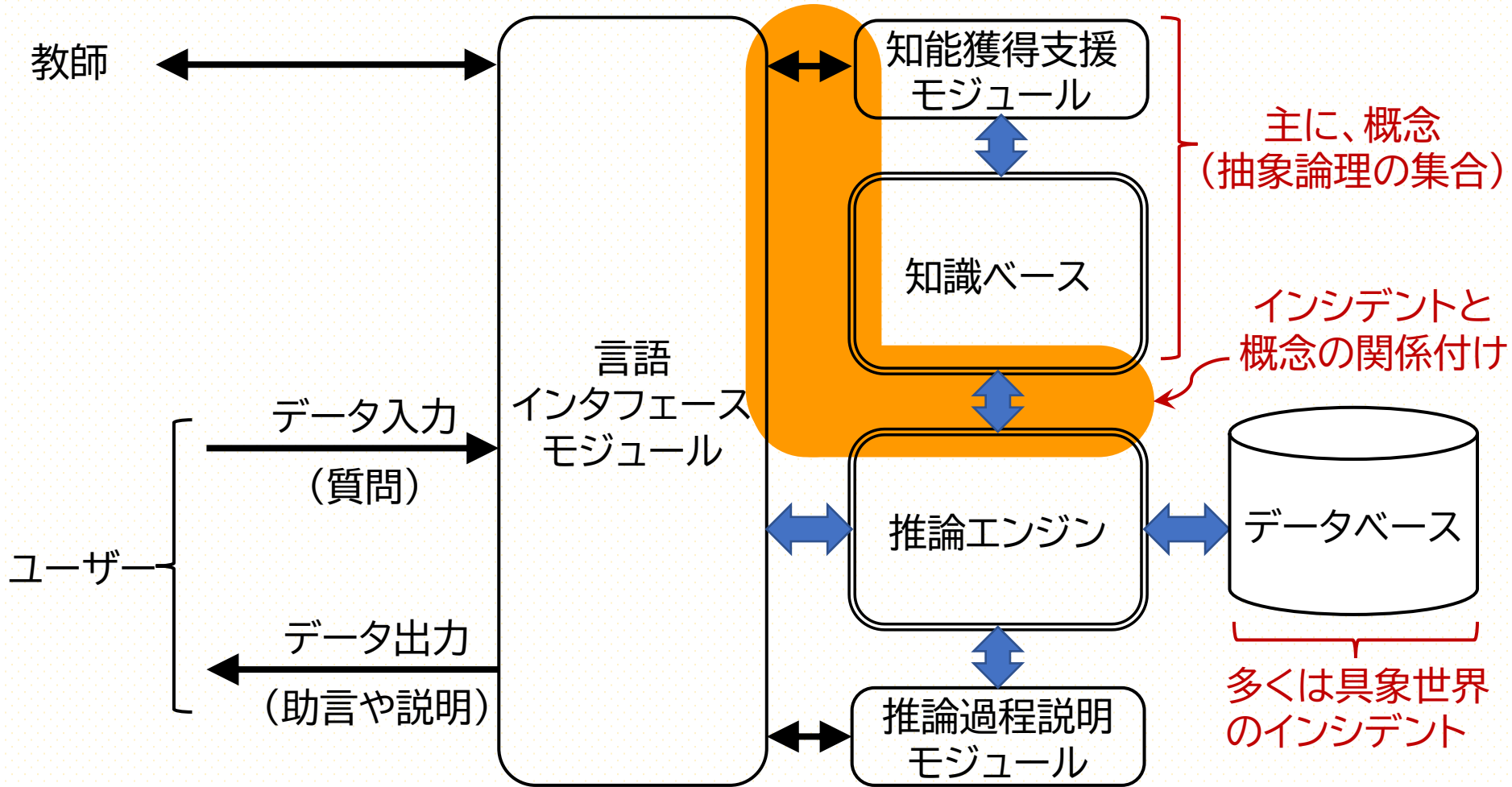
人工知能(AI)開発史



ICOTのエキスパートシステム



ICOTの目指したエキスパートシステム



生成AI (Generative AI)

1. 機能

- ・ 入力文章を元に、文章を作成 (Text to Text model : 次頁に掲載)
- ・ 入力文章を元に、画像を作成 (Text to Image model : 次次頁に掲載)
- ・ 入力文章を元に、プログラムコードの作成 : CodeX (OpenAI社)、等
- ・ 入力文章を元に、音楽作成 : Jukebox(OpenAI社)、等

2. 要素回路 : Transformer回路 (2017年、Google/Google Brain /Google Research、トロント大)

- ・ Attention機構を用いて、入力ベクトル列(文章)を元に、出力ベクトル列を生成する回路
- ・ Language Modelに、言語を学習させ、入力文章が何を求めているかを理解し、その理解の元で、出力すべき単語を推測し、人間に親和性の良い文章を生成し出力する。

3. モデルの巨大化により性能が急上昇

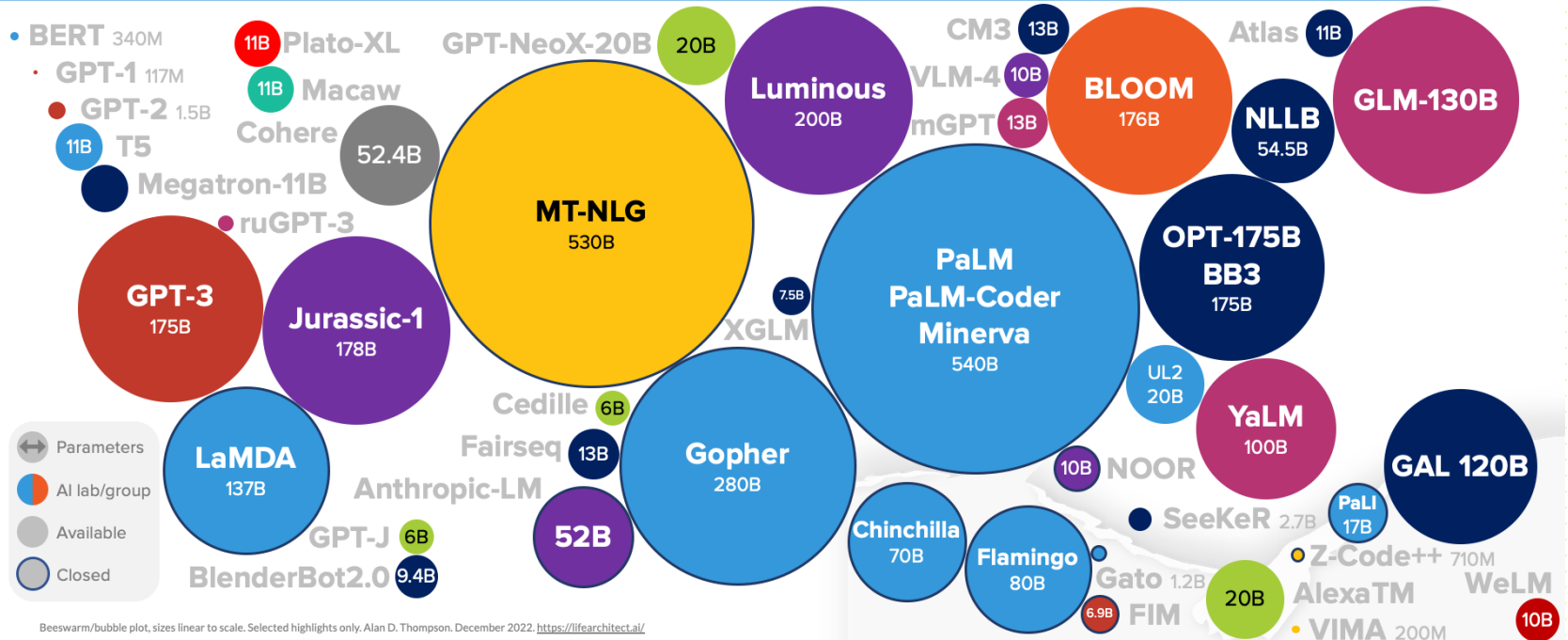
質疑応答 (Question-answering)
翻訳 (Sentence Translation)
読解 (Reading comprehension)
要約 (Summarization)

4. 様々な議論

- ・ 信頼度 / 精度 / 品質
- ・ 社会実装 / 悪用問題 / G7等でのルール作り
- ・ Real-Time学習 (学習ポートの Internet接続)
- ・ 制御系への対応 / 自動運転車への導入 / 兵器の制御

AI language models 2018-2022.

LANGUAGE MODEL SIZES TO DEC/2022



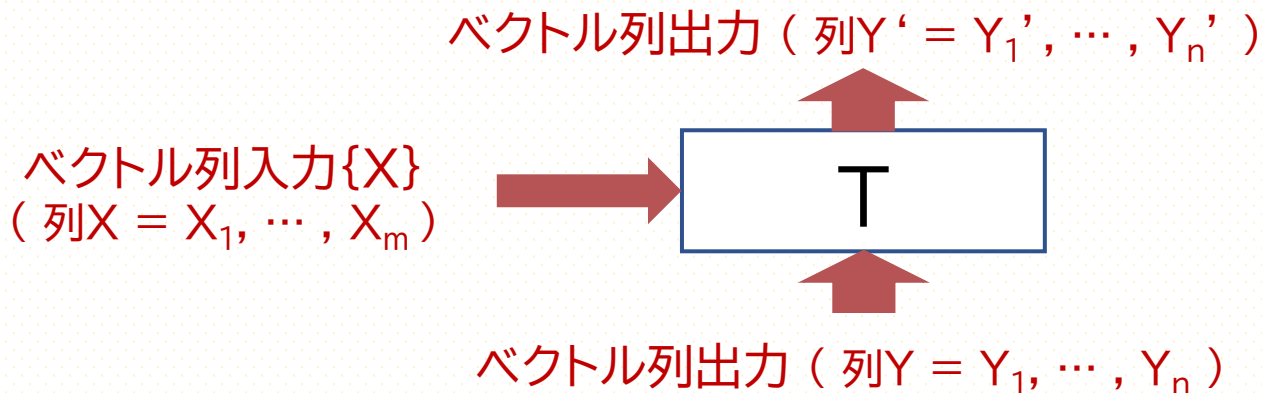
- LaMDA/PaLM/Minerva/Chinchillaは、Google社。
- Gopher/Flamingo/Gato は、DeepMind社。
- GAL、LLaMA はMETA社。
- Jurassic-1 は、Amazon社。
- Luminousは、Aleph Alpha社。
- Bloomは、Hugging Faceが、60ヶ国/約1000人を集め、40ヶ国語向けに立ち上げたGPT-3ベース・モデル。米国のPublic Supercomputerを使い、2022年3月からTrainingを開始し、\$7M の費用で7月6日に完了。
- Yandex (Russia's Google equivalent) • GLM-130B/WeLM 10Bは、中国製。後者はWeChatが改良。

Text-to-image models to end-2022. Selected highlights only.

AI lab	Model	Type	Release	Notes
OpenAI	DALL-E 1	Autoregressive	Jan/2021	
Tsinghua	Cogview 1	Autoregressive	May-21	4B params
Midjourney	Midjourney v1	Diffusion	Apr-22	
OpenAI	DALL-E 2	Diffusion	Apr/2022	1M users in 3 months
Google	Imagen	Diffusion	May-22	
Google	Parti	Autoregressive	Jun-22	20B params
Tsinghua	CogView2	Autoregressive	Jun/2022	24B params, Pathways
Microsoft	NUWA-Infinity	Autoregressive	Jul-22	'Infinite' canvas
Stability.ai	Stable Diffusion	Diffusion	Aug-22	1M users, 50 countries
Craiyon	Craiyon (new)	GAN	Aug-22	2.5B, ex-DALL-E Mini
Baidu	ERNIE-ViLG 2.0	Diffusion	Oct-22	24B params
NVIDIA	eDiff-I	Diffusion	Nov-22	
Midjourney	Midjourney v4	Diffusion	Nov-22	

Transformer モデル：ベクトル列を変換する回路

1. 入出力モデル



2. 基本機能

左からのベクトル列入力(X)によって、
 下からのベクトル列入力(Y)を変形させて(変換し)、
 上辺のベクトル列(Y')出力する回路

3. ベクトル (GPT-3)

Vocabulary
(5万257次元)

〇〇△番目	藍
〇〇〇番目	青い
〇〇□番目	緑
50257番目	****

「意味+配置則」を表現するベクトル
(埋め込みベクトル・: 1万2288次元)

- 「葵」の方向 : 0.01
- ・
- 「安全」の方向 : 0.12
- ・
- 「若い」の方向 : 0.09
- 「悩ましい」の方向 : 0.05
- ・
- 「苦しい」の方向 : 0.01
- ・
- ・
- 「青色」の方向 : 0.55
- 「藍色」の方向 : 0.02
- 「緑色」の方向 : 0.03

文章作成時の
語彙探索で使
う確率ベクトル



Queryベクトル
(512次元)

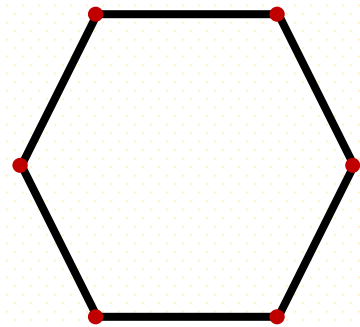
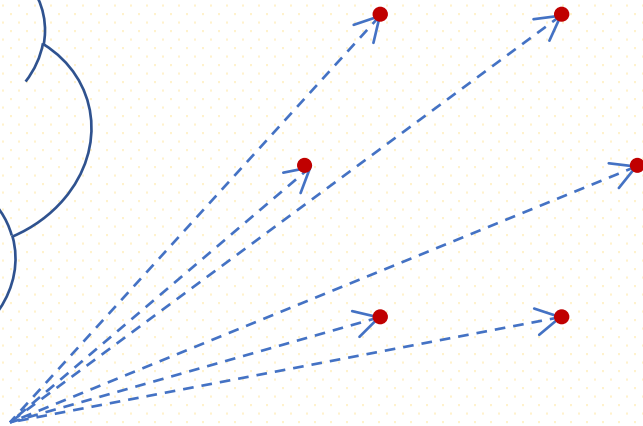
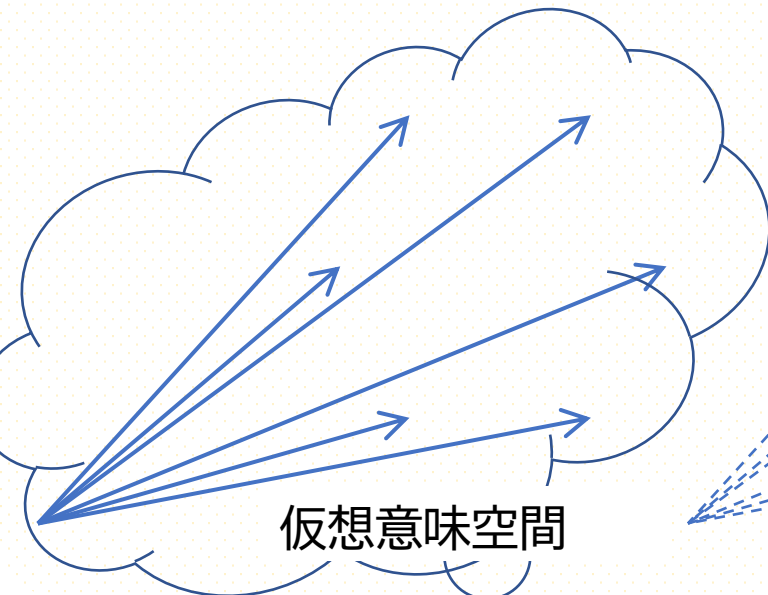
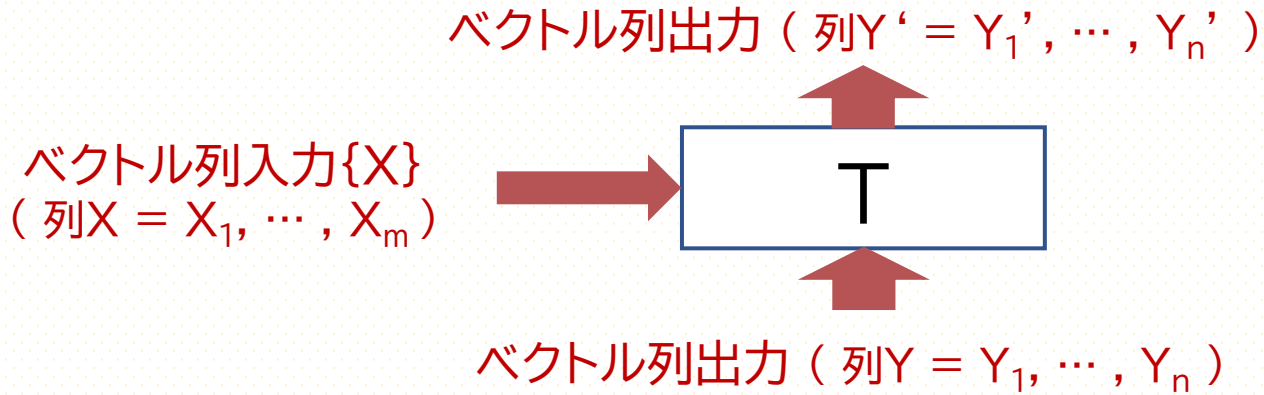


Keyベクトル
(512次元)

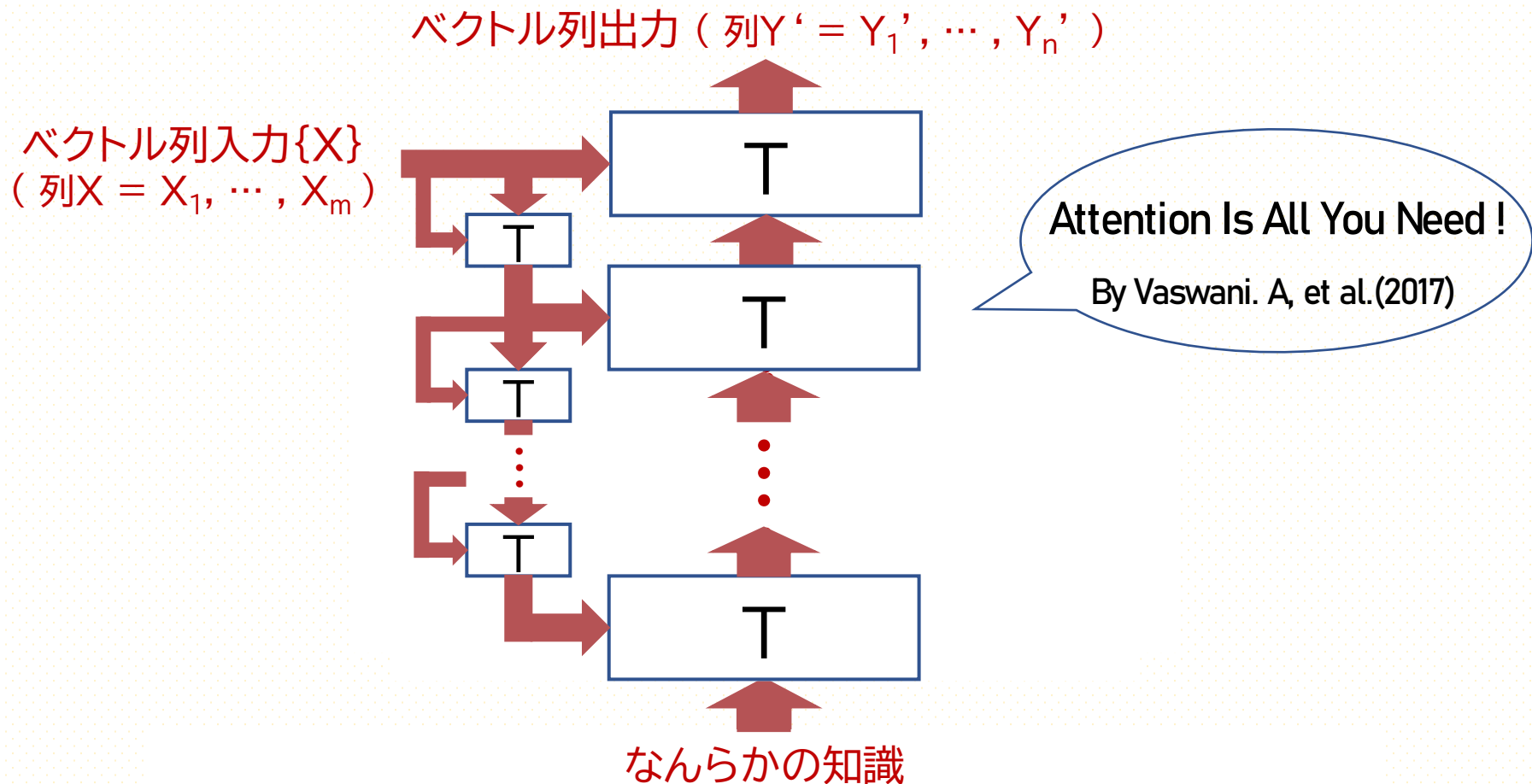


Valueベクトル
(512次元)

Transformer モデル



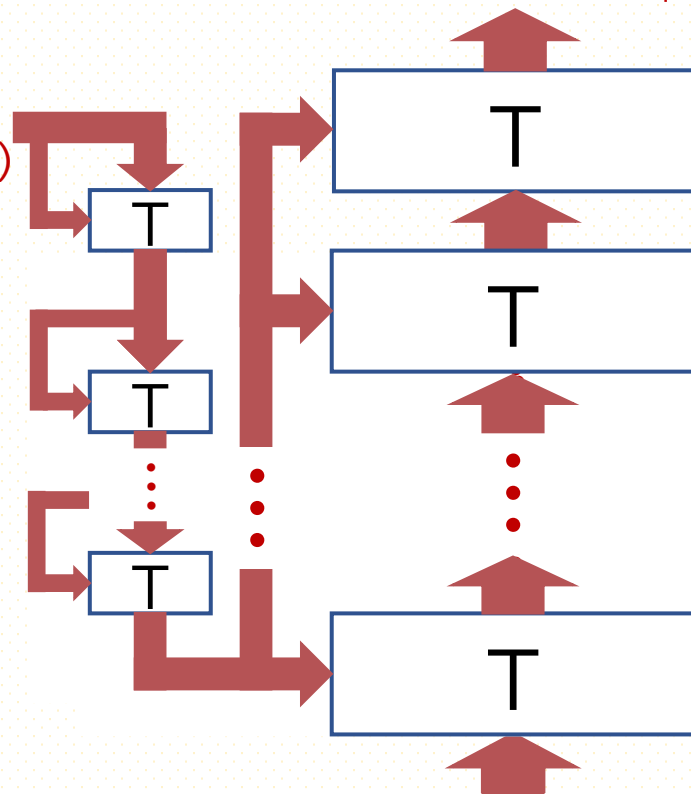
Transformer モデルの回路設計的解釈



Transformer モデル：別形態

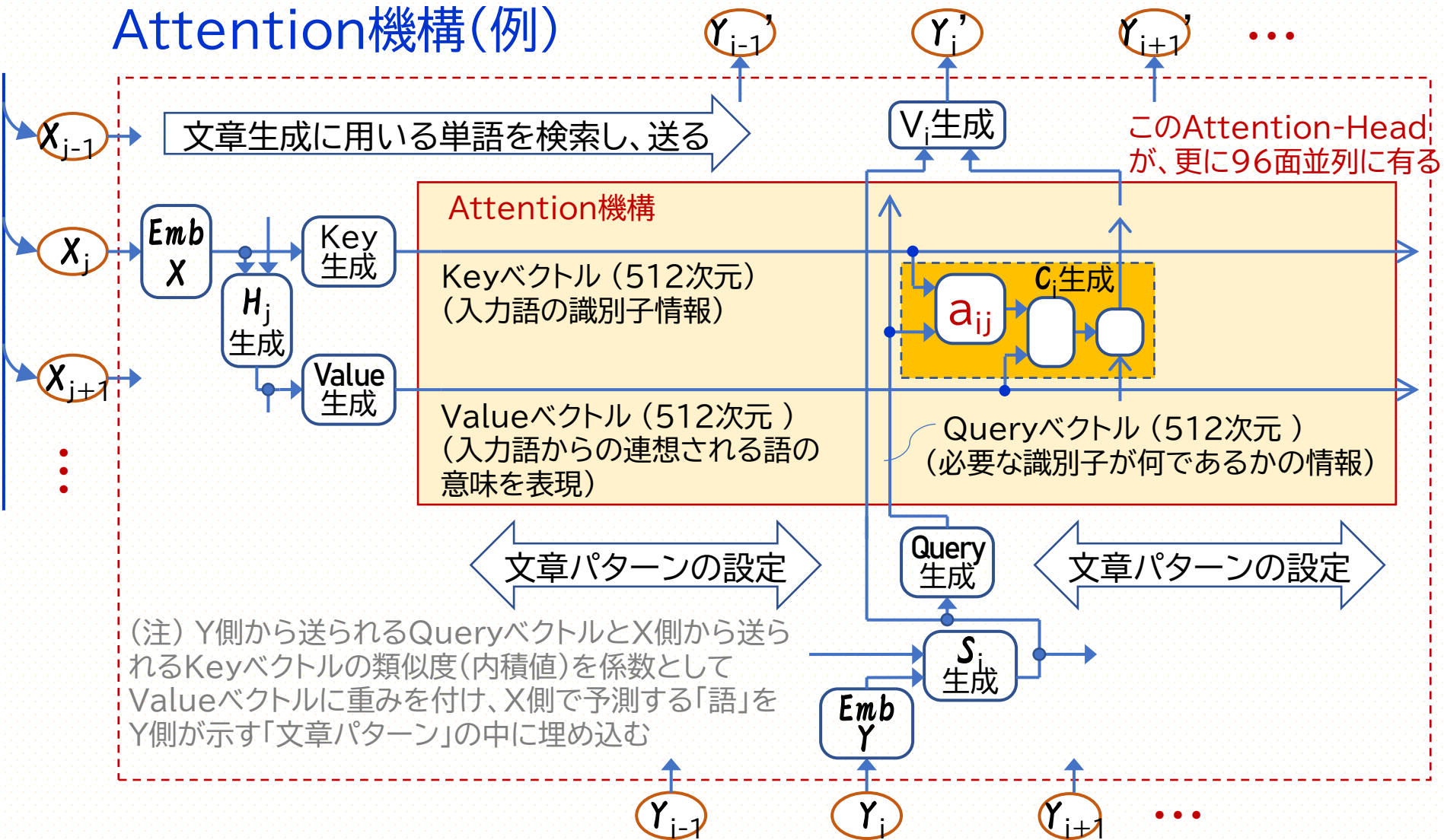
ベクトル列出力 (列 $Y' = Y_1', \dots, Y_n'$)

ベクトル列入力 $\{X\}$
(列 $X = X_1, \dots, X_m$)

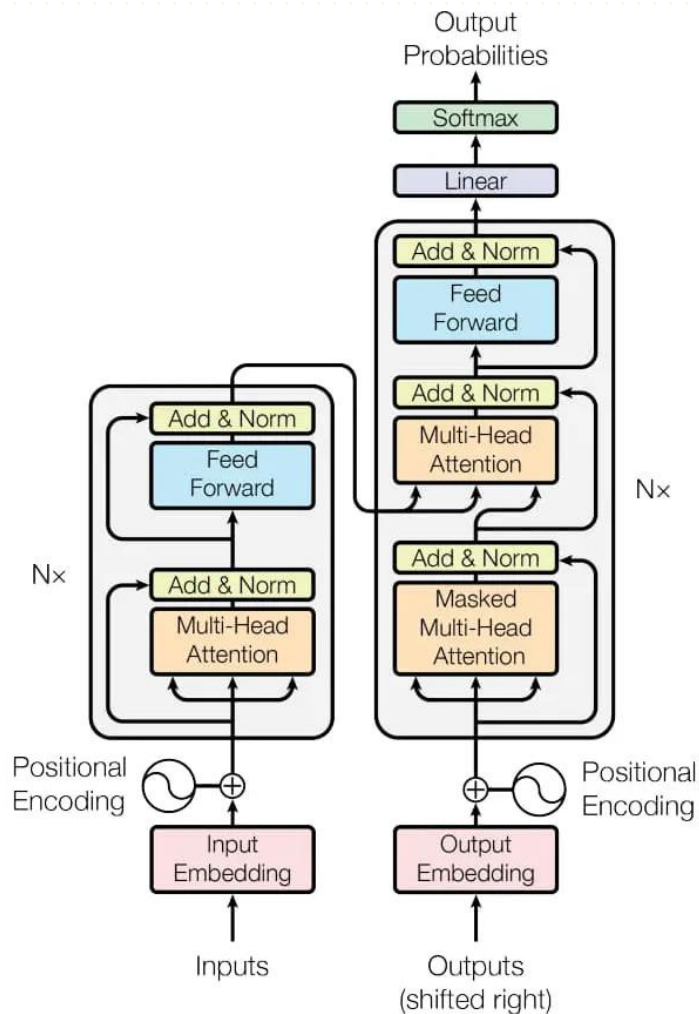


なんらかの知識

Attention機構(例)



Google社のTransformer



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

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

$$c_i = \sum_j a_{ij} h_j \quad (1)$$

$$a_{ij} = \frac{\hat{a}_{ij}}{\sum_j \hat{a}_{ij}} \quad (2)$$

$$\hat{a}_{ij} = \text{att}(s_i, h_j) \quad (3)$$

$$\text{att}(s_i, h_j) = \langle W_h h_j, W_s s_i \rangle.$$

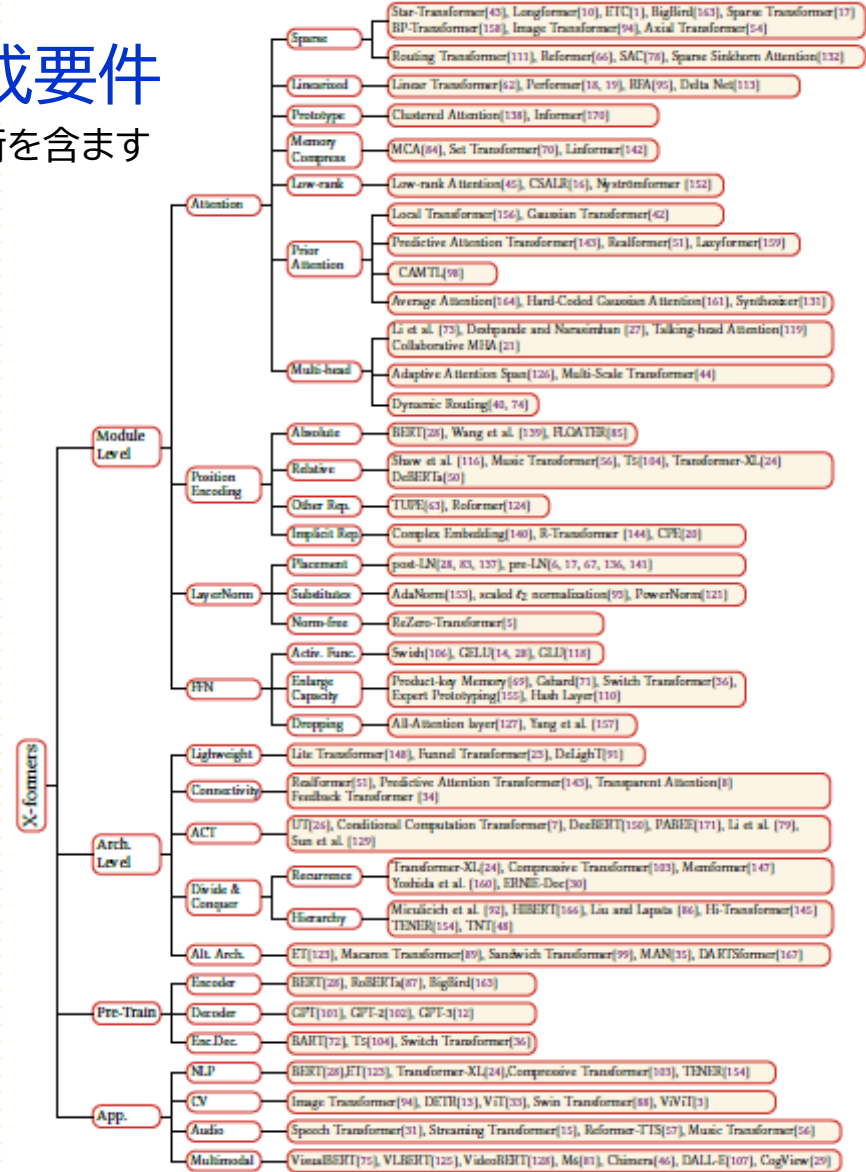
$$P(y_i | y_1, \dots, y_{i-1}, \mathbf{x}) = \text{softmax}(W[s_i; c_i] + b)$$

Britz, et, al, (2017) "Massive Exploration of Neural Machine Translation"

Transformer(X-Former)の構成要件

(注) 追加学習と評価技術を含めます

- Module Level
 - Attention
 - Position Encoding
 - Layer Norm
 - Feed-Forward Network
- Attention
 - Light weight
 - Connectivity
 - ACT
 - Divide & Conquer
 - Alt. Arch.
- Pre-Train
 - Encoder
 - Decoder
 - Enc.Dec
- Application
 - NLP
 - CV
 - Audio
 - Multimodal



GPT-2 / -3 の開発目的 と 状況

- ◆ Autoregressive (自己回帰的) Language Model の 評価 : 評価レポート有り
 - WebTextを用いたPre-Training(事前学習) の 立上げ、問題抽出、効率UP
 - Fine-Tuning(特定タスク向けの追加学習) の学習データ量評価、モデルの修正
 - 人間が行う強化学習の内容、学習体制構築、目的関数/価値の設定
 - One-Shot学習能力/転移学習能力の評価、HumanEval Datasetの開発
 - Scaling則の確認 / コストトレンド
 - 本質問題 (言葉による言葉の生成で良いのか? 等)
- ◆ マーケティング : トップが強力に推進中
 - 多言語対応 ⇒ Hugging Face社 の" Bloom"
 - 専用モデル (Distillated Model) やAPI によるサービスビジネス立ち上げ
 - 用途拡大 (各種エキスパートシステム、DX対応、自動車対応、等)

学習

Dario Amodei (Google Brain), et al. (2016) "Concrete Problems in AI Safety"
John Schulman, et al. (2017) "Proximal Policy Optimization Algorithms"
Tom B. Brown, et al. (2020) "Language Models are Few-Shot Learners"
Damai Dai, et al. (2023) "Why Can GPT Learn In-Context"

1) 事前学習(Pre-Training)

- ・ 言語モデルを習得するための **教師無し学習(Unsupervised Learning)**
- ・ 文章パターン(文法、慣用句、表記方法、テクニカルライティング・スキル)を習得
- ・ 学習データは、Web上のテキスト・データから良質なデータを抽出(フィルタリング)
- ・ 計算量/使用電力が膨大。

GPT-3の学習向けには、45TeraByteを収集し、
フィルタリングにより570GBを学習コーパスとした。

2) 追加学習(Supervised Fine-Tuning)

- ・ 特定分野/タスクに適応のための **教師有り学習(Supervised Learning)**
- ・ 転位学習能力強化 (ユーザの要求や意図に応ずる方法を文章パターンとして蓄積)
- ・ ラベル付け人員(Trained Labeler)が入力プロンプトと所望の出力文を用意

3) 価値観/モラルの学習 (Proximal Policy Optimization)

- ・ 人間からのFeed-Backによる **強化学習(Reinforcement Learning)**
- ・ 強化学習での報酬モデル(真実性、無害性、有益性を評価するマシン)の学習

GPT-3では、13000文用意

報酬モデルは、GPT-3(パラメータ数 6Billion)

4) Prompt入力

- ・ 転移学習能力を利用して、ユーザーの要望や意図を学習 (**One-Shot Learning**)

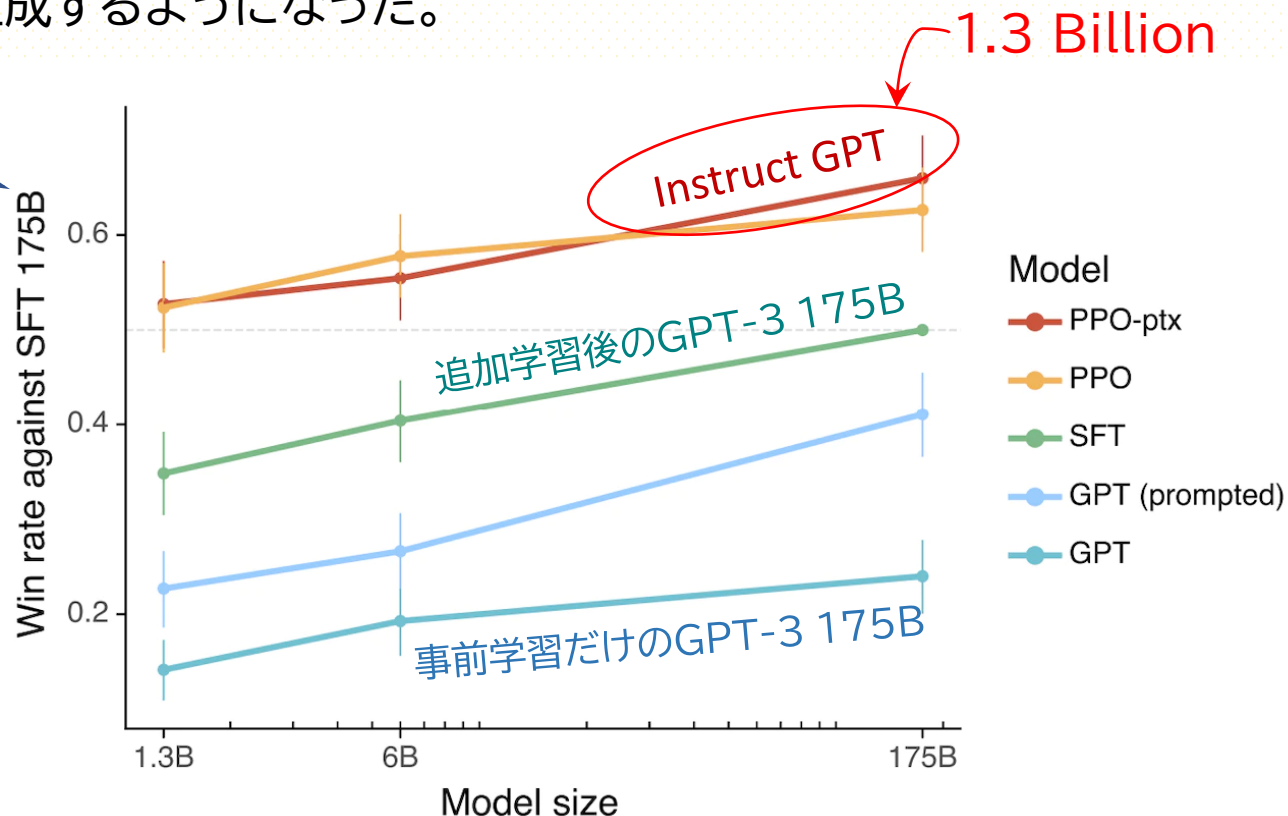
Chinchilla ショック (Deep Mind社の小型GPT)

- 学習計算量一定の前提でモデルが必要とする学習データ量を調査すると、小さなモデルをより多くのデータで訓練した方が性能が高くなることを発見(DeepMind社)
 - GPT-3 (175 Billion)は、11倍の学習データで学習させるべきだった)
 - 実際の学習データ : 300 Billion のトークン量 (570 GigaByte相当)
 - 必要としていた学習データ : 3500 Billion のトークン量 (7 TeraByte相当)
 - (注) データ収集能力が非常に重要となった。
但し、データは「複数回使う」ことでデータを節約できるかもしれない。
 - (注) A4 1枚は4kB。7TeraByteは、A4文書の20億頁分のデータ量。
フィルタリング前のWebTextとしては、その約80倍を収集する必要ある。
- この結果を受けて、OpenAI社は、人間のフィードバックによる強化学習を強化し、GPT-3(175GB)を改良し、InstructGPT(1.3B)を完成 (2022年5月)

Instruct GPT (GPT-3 からの改良版)

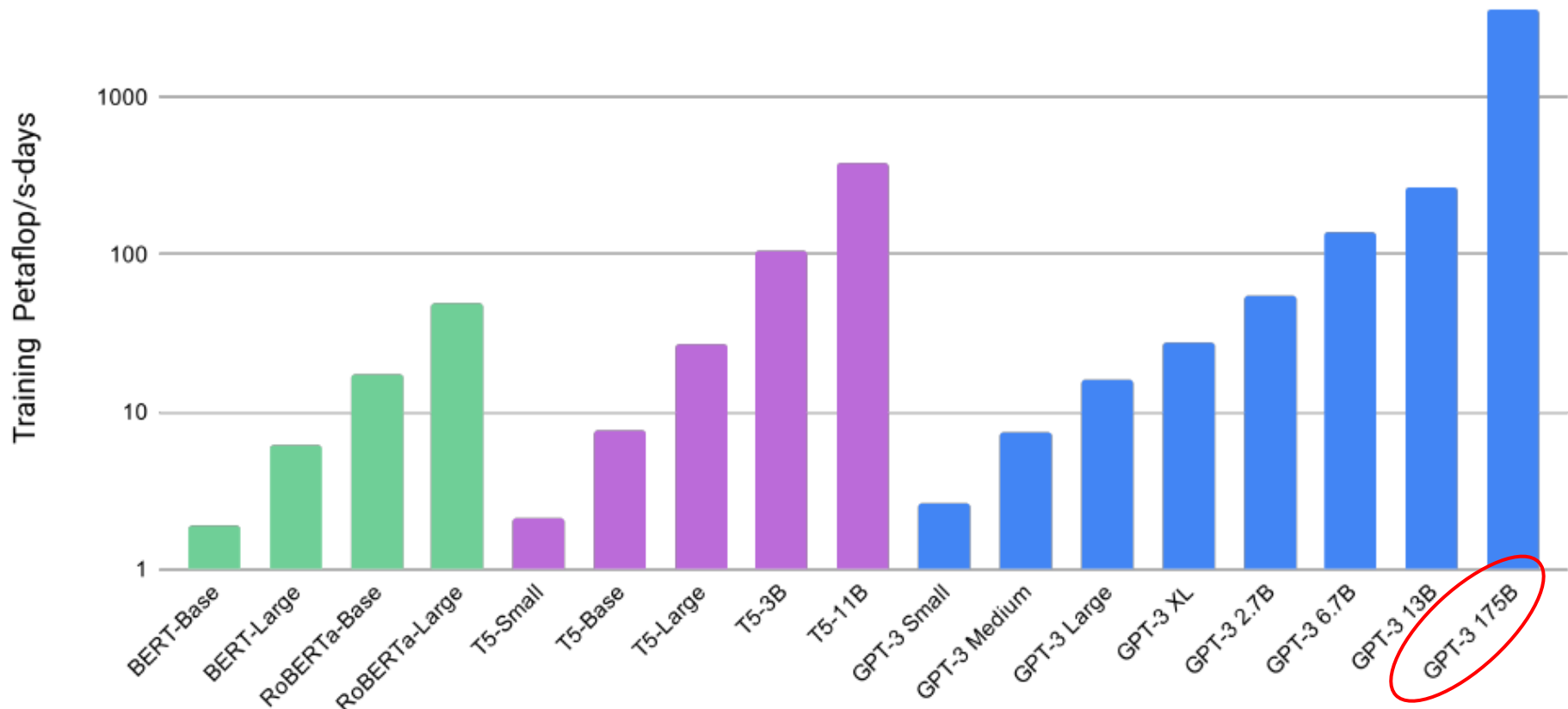
- 人間のフィードバックによる強化学習にて、Proximal Policy Optimizationを行うと(Instruct-GPT)、回路規模が GPT-3 (175B)の100分の1以下(1.3B)でも、好ましい文章を生成するようになった。

返答文章が、人間が評価にて「好ましい」とみなされた割合



事前学習時の計算量は膨大：GPT-3は10ペタのスパコンを約300日使用

- コンテンツ生成段階の消費電力は小さい（100頁当たり数セント。約0.4 kW-hr）
- 生成AIのビジネスは、いかにアプリを増やすかにかかっている。



Transformerが良い文章を生成する正確な理由は、未だ不明？

- ◆ Y. Dong, et al. (2021年3月); "Attention is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth"
アテンションベースアーキテクチャは、機械学習においてユビキタスなものとなっているが、その有効性の理由についての理解はまだ不十分なままです。
- ◆ N. Bogoychev (2021年9月); "Not all parameters are born equal: Attention is mostly what you need"
トランスフォーマーは最先端の機械翻訳で広く使われているが、その成功の鍵は未だ分かっていない。
- ◆ C. Wu, et al. (2021 9月) ; "Fastformer: Additive Attention Can Be All You Need"
Transformerは、テキスト理解のための強力なモデルである。しかし、入力配列の長さに対して2次関数的な複雑さを持つため、非効率的なモデルである。
- ◆ R. Singh, et al.(2023 4月) ; "Attention: Marginal Probability is All You Need?"
アテンション機構は、機械学習のアーキテクチャの選択肢として、圧倒的な存在感を示している

(私見)2022年末には、Transformerの言語モデルは認知/評価を得た。

GPT-4(2023年3月14日公開)の 文章応答例

次のような質問をGPT-4にしてみました。

丸山氏)「ある文から他の文を連想する時、あなたはどんな手順で連想を実行しているのですか？
詳しく教えてください。」

そうすると、GPT-4は大いに語り始めます。

GPT) 私は人間のように意識的な思考過程を持つわけではないので、「連想」を行う方法は一般に想像するようなものとは異なります。しかし、人間が理解しやすい形で説明するために、私の動作を一連の「手順」として表現することは可能です。
以下に、そのようなステップを説明します。

1. 元の文の理解 : 最初のステップは、入力された文を**理解**することです。
2. 要素の抽出 : 次に、文から主要な要素を抽出します。
3. 関連性の確認 : これらの要素がどのように関連しているかを確認します。
4. 連想の生成 : 抽出した要素とその関連性に基づいて、新たなシナリオ、アクション、記述を生成します。
5. 新しい文の生成 : 最後に、これらの新たな連想を元に新しい文を形成します。

(続) GPT-4の文章応答

丸山氏) この「手順」はどのように学習されたのですか？

GPT) しかし、私が「連想」のプロセスを5つの手順に分解した説明は、人間がこのAI技術を理解しやすいように、AIの内部動作を人間の思考プロセスに似せて説明したものです。私はこの手順を「学習」したわけではありません。

私は単に統計的なパターンマッチングを行っているだけで、意識や理解、自発性は持っていません。私が生成するすべてのレスポンスは、学習データに基づく予測に過ぎません。」……
(中略)…… 私が行う「理解」とは、入力されたテキストを解析し、そのテキストに基づいて最も適切な応答を生成することを指します。これはパターン認識やデータの処理に過ぎず、人間のような主観的な感覚や意識を伴っていません。同様に、私が自分自身や人間との関係について説明するとき、それは私が学習したテキストデータに基づいています。私自身には意識、感情、意図、理解といった人間のような特性はありません。これらはすべて、人間がAIの動作を理解しやすいように説明するためのメタファーです。人間が人間の経験や認識を通じてこれらの概念を 理解するように、AIがこれらを「理解」することはありません。

OpenAI社は、ChatGPTに、「私は理解していない」と言わせていると思われる。

丸山氏の解説:

RBRMに基づく Model Refusal が発動されているのだと思います。“GPT-4 System Card” の例では、明示的には取り上げられていませんでしたが、 OpenAIは、GPTが、“我々は、高い理解力がある” とか “我々にも人間並の意識がある” と 答えることを強く警戒しているように思います。(後略)

AIのBigName達のコメント

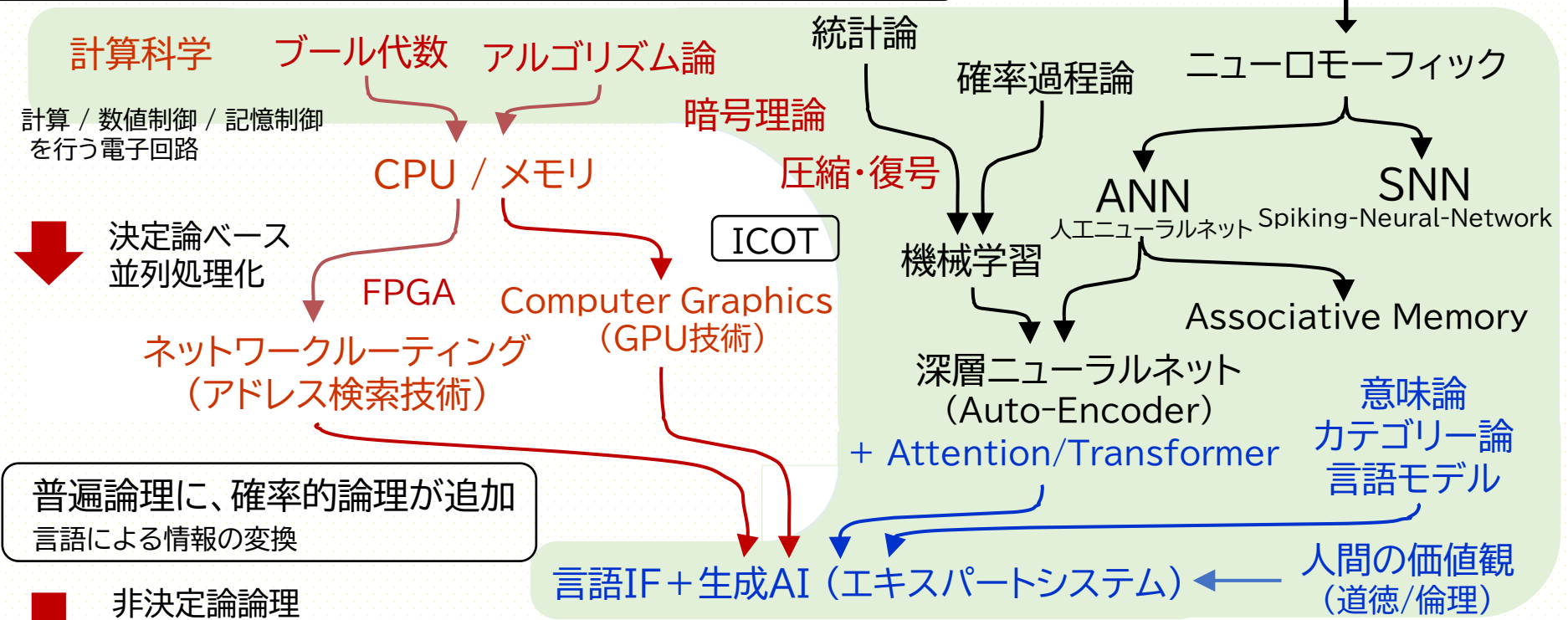
- ◆ Yann Lecun (Meta Platforms社、Open Reviewに投稿した論文(2022年6月))
 - ・ 統計学的アプローチ(Deep Learning)は扱いにくい。確率論だけで世界モデルを作るのは無理。
 - ・ Transformerのは未来の知能システムの構成要素かもしれないが、必要不可欠な部分が欠けている。
 - ・ 世界の仕組みを理解することができるシステムから、よりよいソリューションが生まれるかもしれない。
- ◆ Geoffrey Hinton (at 神経情報処理システム会議の閉会基調講演、NeurIPS、2022年12月)
 - ・ 全く違う全く新しいコンピュータを登場させる必要がある。
 - ・ ニューラルネット技術も変わる必要がある。Forward-Forward Networkが可能なら、現実の脳で起こっていることを、よりもっともらしく再現できるかもしれない。<https://www.cs.toronto.edu/~hinton/FFA13.pdf>
- ◆ OpenAI社のサム・アルトマンCEOの発言 (at マサチューセッツ工科大学、2023年4月13日)
 - ・ 巨大AIモデルを用いる時代は終わった。GPT-4の訓練には1億ドル以上の費用がかかった。
 - ・ 今後の進歩はモデルをさらに大きくすることでは達成できない。
- ◆ イーロン・マスク (2023年4月14日 Financial Times、Wallstreet Journal、4月17日 FoxNews)
 - ・ OpenAIの透明性の欠如、利益志向と運営の姿勢を批判。別の選択肢が必要だ
 - ・ 万物を理解しようとする真実追求型のAIであるTruthGPTを開発する真会社(X.AI)を検討する。

まとめ

- Transformer回路は、説得力あるコンテンツを生成すると認められつつある
(確率的/統計的なベクトル列の処理にて、普遍的ともいえる確率論理を処理する)
- 但し、用意できる良質なテキスト・データの量が、有効な回路規模の上限を決めてしまうのが問題 (良質なデジタル・コンテンツの量が、回路の価値を決めてしまう)
- 確率的/統計的な言語モデルのままで良いのかについては、未だ議論がありうる
 - 学習時の計算量が大き過ぎる
 - そもそも、言葉による言葉の生成で良いのか？
(ICOTが想定したような知識獲得や推論過程説明の機能は、未だ無い)
- 今こそ、第五世代コンピュータが必要ではないか？ (日本発の論文は全体の1%)
 - 現在は、未だ、GPUを使って回路シミュレーションしている段階だが、設計が完了すると、次は、その回路を搭載したチップ開発の時代が来る

(参)生成AI / Transformer回路 / Self-Attention機構 の位置付け

普遍(Universal)論理： 古典的チューリングマシンの時代



普遍論理に、確率的論理が追加
言語による情報の変換

非決定論論理
超分散ネットワーク化

汎用知能化

「思考」と「発見」と「制御」の技術 **言語IF + 生成AI + 数値制御 + 思考(アルゴリズムの発見能力)**

By Benjamin Thompson & Noah Baker (Jun 9, 2021), in Nature.


“Google AI beats humans at designing computer chips”

By James Vincent (Jun 10, 2021), in “The Verge”.

<https://www.theverge.com/2021/6/10/22527476/google-machine-learning-chip-design-tpu-floorplanning>

(参) Google と NVIDIA は、AI をチップ設計に利用開始 (2021年)

- グーグル社は、TPUチップの次期バージョンにAIを用いて設計(2021年)
(人間が数カ月かかる作業を、AIは6時間以内に達成)
- 様々なチップのフロアプランを1万枚集めたデータセットで、AIを強化学習
(必要な配線の長さや電力使用量などを指標に基づく報酬関数を設定)
- AIのフロアプランは、人間の設計とは全く異なっていた。
(人間の設計は整然としたパターンが多いが、AIは、そのような制限がなく、
信号の流れ、効率、電磁結合の悪影響、等を最適化する。)
- Google社は、「アーキテクチャの探索技術」にてAIを使うことを模索中
=> “Neural Architecture Search“
- NVIDIAは、H100 Hopper チップで、13000の回路設計にAIを活用
(2022年に、A100チップよりも最大で6倍高速なGPUが実現)



ご清聴いただき、ありがとうございました。