3-5　生成AIの基礎と展望

東京大学 数理・情報教育研究センター

2024年6月1日

概要

- 基盤モデルと生成AIについて理解する.
- 実世界で進む生成AIの応用と革新について理解する.
- 大規模言語モデルについて理解する.
- 拡散モデルについて理解する.
- 生成AIの留意事項について理解する.

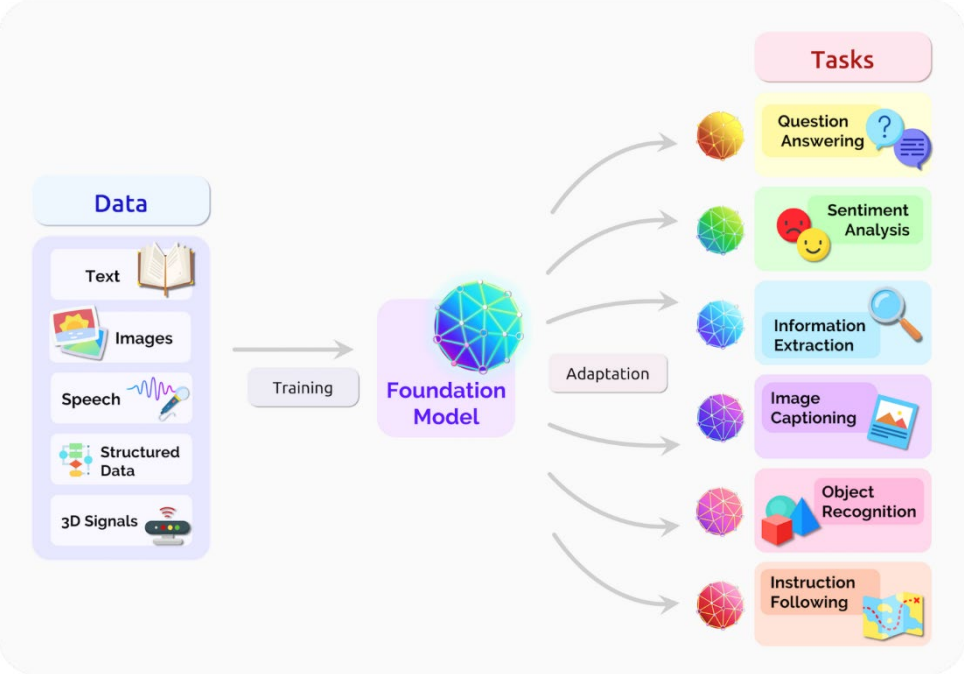
本教材の目次

基盤モデル

- 大規模で多種多様なデータをもとに学習され，さまざまなタスクに適応可能な機械学習のモデルを基盤モデルと呼びます．

  “A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.”[1)]

  - 基盤モデルは主に深層学習[2)]の技術に立脚しています．



文献1)の図 より引用

1) R. Bommasani, et al. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258, 2021.
2) 応用基礎レベル教材 3-4. 深層学習の基礎と展望 参照

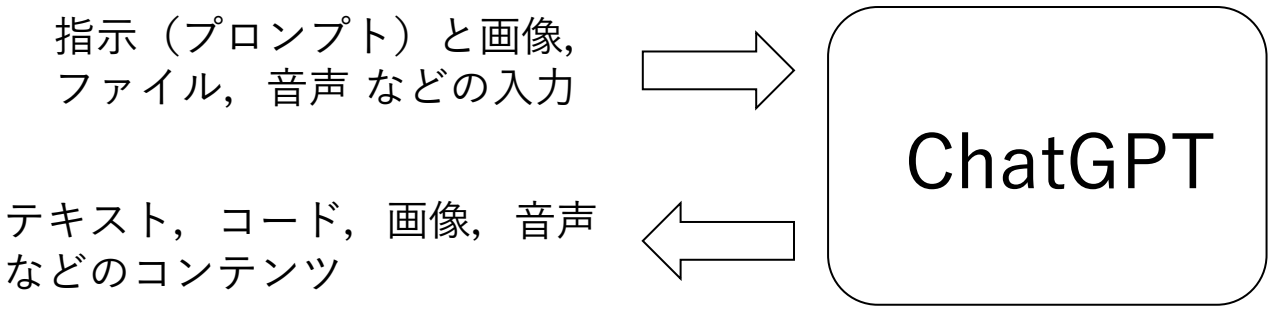
基盤モデル

- 基盤モデルは，転移学習という機械学習の方法の考え方がもとになっています．
- 転移学習は，あるタスクから学習した「知識」を他のタスクの問題を解くことに適用する方法です．
- 基盤モデルの学習においては，補助的なタスクで「事前学習」（主に自己教師あり学習[1)]による）を行なったモデルを対象のタスクに適応するように学習（ファインチューニング）する，というように転移学習を行います．
- 基盤モデルはこのような転移学習を大規模に行うことで構築されますが，そのためには（1）GPU（Graphics Processing Unit）のようなハードウェア，（2)トランスフォーマ（Transformer）[2)]のように並列処理を適用可能なモデル，（3）大規模なデータが必要となります．

1）本教材の項目「自己教師あり学習」参照
2）本教材の項目「トランスフォーマ」参照

生成AI

- 主に基盤モデルに基づく生成モデル[1)]を用いて，テキスト，画像，音声，動画，その他のコンテンツを生成する人工知能（AI）技術の総称を生成AIと呼びます.
- 生成AIの利用においては，ユーザが自然言語による指示（プロンプト）を入力することで，そのプロンプトに従ってコンテンツを生成（より一般的にはタスクを実行）することが可能です.
- 例えば，2022年11月に公開されたChatGPT[2)]は，大規模なテキストデータをもとに事前学習された大規模言語モデルを基盤モデルとして，プロンプトに対してテキストなどのコンテンツを生成する生成AIのサービスです.



1) 本教材の項目「生成モデル」参照
2) https://chat.openai.com/

生成AI

- 2023年3月に公開された大規模言語モデルであるGPT-4では，従来のモデルに比べて言語生成，知識活用，知識推論，記号推論，数学推論などの能力がさらに向上したことが報告されています[1)]．
- また，生物学，経済学，化学，物理学などの複数の専門的な試験においても一定程度の成績を収めることが報告されています[2)]．

User　Answer question I.1.a. Think step-by-step.

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T, supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R}\frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th}.

Un barreau, conducteur thermique, homogène, de longueur L, de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

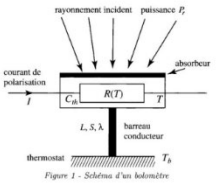


Figure 1 - Schéma d'un bolomètre

I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_b$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_0$ et $T(L) = T_b$.

GPT-4　To answer question I.1.a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x. We are given that $T(0) = T0$ and $T(L) = Tb$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:
$d^2T/dx^2 = 0$

Integrating once, we get:
$dT/dx = C_1$

Integrating a second time:
$T(x) = C_1x + C_2$

To determine the constants C_1 and C_2, we use the boundary conditions:
$T(0) = T0$, which gives $C_2 = T0$
$T(L) = Tb$, which gives $C_1 = (Tb - T0)/L$

Finally, the expression for $T(x)$ is:
$T(x) = (Tb - T0) * (x/L) + T0$
This represents a linear temperature profile along the conductive bar, with a slope of $(Tb - T0)/L$.

文献2)の図より引用

- 数学，プログラミング，医学，法学などでは人間のレベルのパフォーマンスを達成という予備的な実験結果の報告もなされています[3)]．

1) W. Zhao, et al. A Survey of Large Language Models. arXiv:2303.18223, 2023.
2) OpenAI, et al. GPT-4 Technical Report. arXiv:2303.08774, 2023.
3) S. Bubeck, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712, 2023.

実世界で進む生成AIの応用と革新

- 生成AIは，テキスト，画像，音声，動画などさまざまなモダリティを入出力としてコンテンツの生成を可能にします．
- 以下のように，入力に対してさまざまなコンテンツを生成可能な生成モデルの技術とそれらのサービスが現在研究開発されています．
  - テキストの生成 (Text-to-Text, Image-to-Text, Speech-to-Text など)
  - コードの生成 (Text-to-Code など)
  - 画像の生成 (Text-to-Image, Image-to-Image など)
  - 動画の生成 (Text-to-Video など)
  - 音楽の生成 (Text-to-Audio, Text-to-Music)

そのほか，3Dモデルの生成，ロボットやシミュレーションに応用可能な行動や動作系列の生成，などを行う生成AI技術に関する研究開発も進められています．

テキストの生成

- 大規模なテキストデータをもとに事前学習された大規模言語モデル[1)]により，入力となる指示（プロンプト）に対してテキストを生成.
- テキスト作成の支援，テキストの翻訳・要約・執筆・校正，対話応答などへ応用することができます.
- 代表的な大規模言語モデルとそのサービス
  - サービス：ChatGPT（OpenAI社）
    - モデルはGPT (Generative Pre-trained Transformer)の系統
  - モデルおよびサービス：Gemini（Google社）
  - モデルおよびサービス：Claude（Anthropic社）
  - サービス：Meta AI（Meta社）
    - モデルはオープンソースのLlaMA (Large Language Model Meta AI)
  - モデル：Mistral, Mixtral（Mistral AI社）
- 日本語大規模言語モデルのまとめ（llm-jp，国立情報学研究所）

1) 本教材の項目「大規模言語モデル」参照

コードの生成

- テキストを生成する大規模言語モデルの技術は，プログラムのソースコードを生成するText-to-Codeモデルにも応用されています．
- Text-to-Codeモデルはプログラミングやソフトウェア開発の支援へ応用することができます．
  - Text-to-CodeモデルのCodex [1)]でのコード生成の例

プロンプト

```
def below_zero(operations: List[int]) -> bool:
    """ You're given a list of deposit and withdrawal
     operations on a bank account that starts with
    zero balance. Your task is to detect if at any
    point the balance of account fallls below zero,
    and at that point function should return True.
    Otherwise it should return False.
    >>> below_zero([1, 2, 3])
    False
    >>> below_zero([1, 2, -4, 5])
    True
    """
```

コード

```
    balance = 0

    for op in operations:
        balance += op
        if balance < 0:
            return True
    return False
```

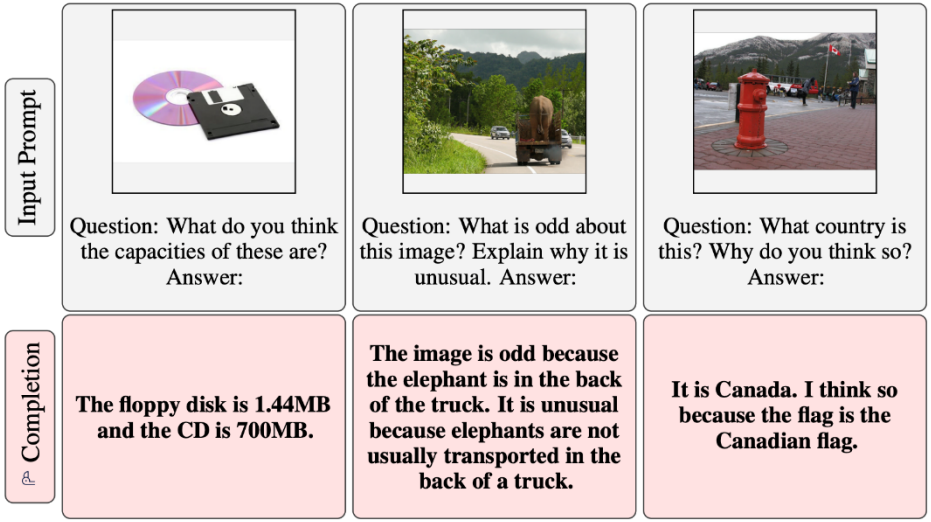
文献1)の図を一部改変して引用

  - Text-to-Codeモデルのサービスの例
    - GitHub Copilot（GitHub社）

1) M. Chen, et al. Evaluating Large Language Models Trained on Code. arXiv:2107.03374, 2021.

画像からテキストの生成

- 大規模なテキストと画像のデータをもとに事前学習されたマルチモーダルな基盤モデルである画像-言語（Vision-Language）モデルにより，テキストや画像を入力としてテキストを生成.
- キャプションの生成，画像を用いた質問応答や対話などへ応用することができます.
  - 画像言語モデルFlamingo[1)]での画像質問応答の例



文献1)の図より引用

1) J.-B. Alayrac, et al. Flamingo: a visual language model for few-shot learning. arXiv:2204.14198, 2022.

音声からテキストの生成

- 音声入力に対してテキストを生成する音声認識技術の開発も行われています．
  - 従来の音声認識技術に対して，基盤モデルに基づく技術が近年導入されています．
  - 例えば，ChatGPTのサービスを提供しているOpenAI社は，音声認識と自動翻訳のためのモデルである[Whisper](1)を公開しています．
- 音声からテキストの生成は，対話システムや自動応答システム，文字起こしや翻訳の自動化などへ応用することができます．
- テキストから音声を生成するText-to-Speech（TTS）の技術も発展しており，人のような自然な発話を生成することも可能になっています．
- 音声認識やTTSにより，生成AIを対話的に利用可能にするサービスの開発も行われています．
  - 例えば2024年5月に発表されたGPT-4oでは，人と会話するのと同様にほぼ遅延なく音声で生成AIと対話することも可能になっています[2)]．

1) A. Radford, el al. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2204.14198, 2022.
2) https://openai.com/index/hello-gpt-4o/ （2024-06-01 閲覧）

画像の生成

- 大規模なテキストと画像のデータをもとに事前学習されたマルチモーダルな基盤モデルであるText-to-Imageモデルにより，入力（プロンプト）に対して画像を生成.
  - 生成モデルや拡散モデルなどの技術が活用されています[1].
- 画像作成や編集の支援に応用することができます.
  - Text-to-ImageモデルのDALL-E2[2]での画像生成の例

vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

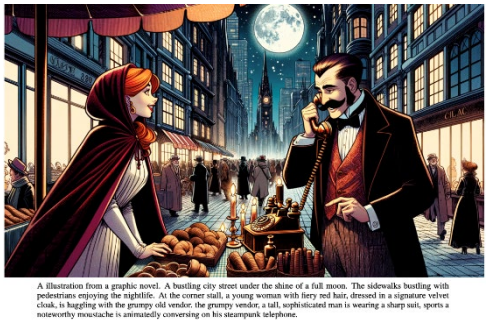
a close up of a handpalm with leaves growing from it

文献2)の図より引用

1) 本教材の項目「拡散モデル」，項目「生成モデル」参照
2) A. Ramesh, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125, 2022.

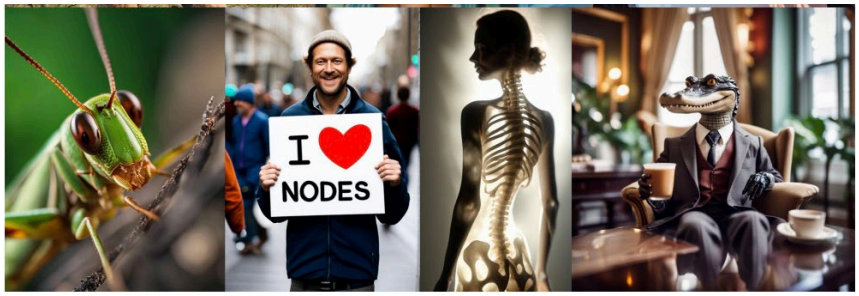
画像の生成

- 代表的な画像生成のサービス
  - DALL-E（ChatGPTで利用可能）（OpenAI社）



文献1)の図より引用

  - Stable Diffusion（Stability AI社）



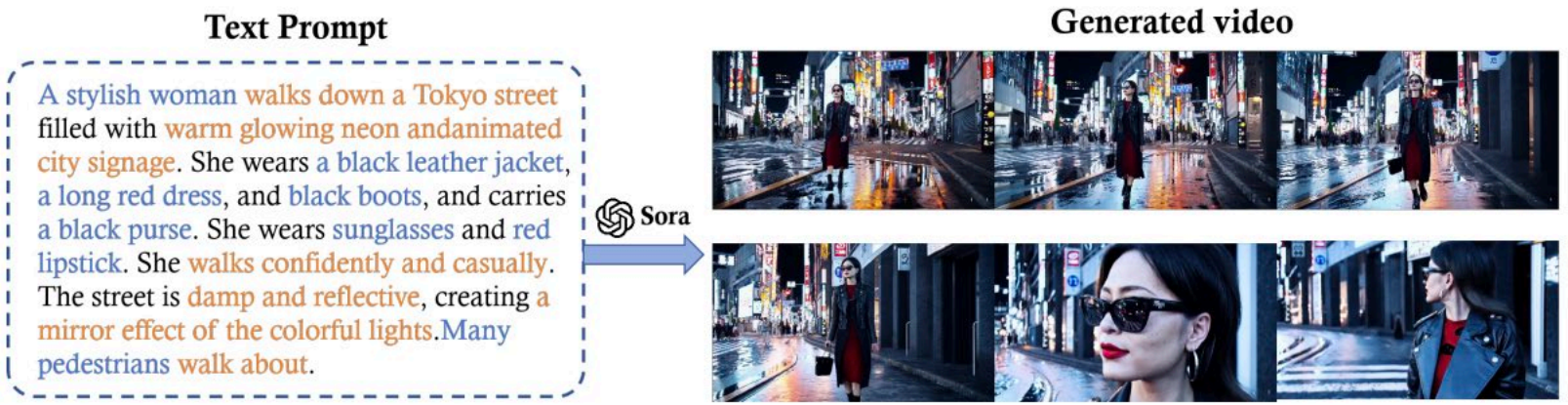
文献2)の図より引用

  - Midjourney（Midjourney社）

1) J. Betker, et al. Improving Image Generation with Better Captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023.
2) D. Podell, et al. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952, 2023.

動画の生成

- Text-to-Imageモデルを発展させ，入力（プロンプト）に対して動画を生成するText-to-Videoモデルの開発も行われています．
- Text-to-Videoモデルは現実の物理世界をシミュレートする世界シミュレーターへの応用も期待できます．
  - Text-to-VideoモデルのSORA[1]での1分程度の動画生成の例[2]



文献2)の図より引用

1) Video generation models as world simulators, https://openai.com/index/video-generation-models-as-world-simulators (2024-05-01 閲覧)
2) Y. Liu, et al. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv:2402.17177v3, 2024.

音楽の生成

- 入力（プロンプト）に対してオーディオ（音声やサウンドエフェクトなど）を生成する Text-to-Audioモデルや音楽を生成するText-to-Musicモデルの開発も行われています.
  - オーディオ生成の例
    - Audiobox[1)]
    - AudioLDM[2)]
  - 音楽生成の例
    - MusicGen[3)]
    - MusicLM[4)]

1) A. Vyas, et al. Audiobox: Unified Audio Generation with Natural Language Prompts. arXiv:2312.15821v1, 2023.
2) H. Liu, et al. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. arXiv:2301.12503v3, 2023.
3) J. Copet, et al. Simple and Controllable Music Generation. arXiv:2306.05284v3, 2024.
4) A. Agostinelli, et al. MusicLM: Generating Music From Text. arXiv2301.11325, 2023.

実世界で進む生成AIの応用と革新

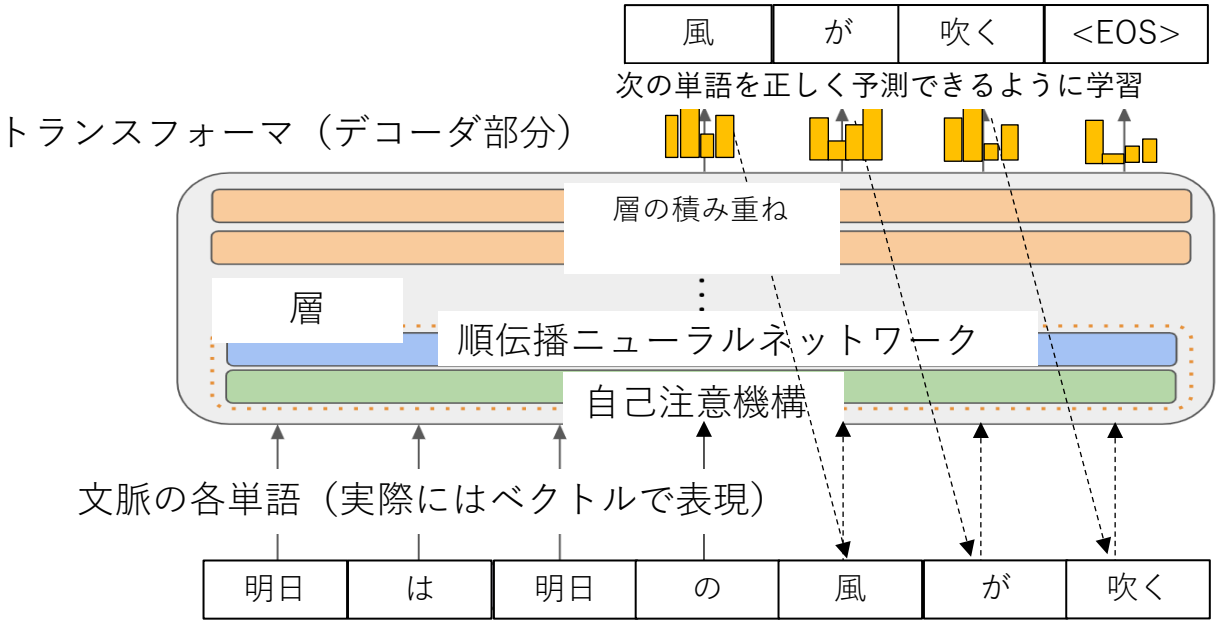
産業応用

生成AI技術のさまざまな産業応用が現在検討・試行されています.

- ビジネス
  - 文書作成支援・要約・翻訳，マーケティング支援，カスタマーサービス，ブレインストーミング など
- 教育
  - 教材作成，個別指導，自動評価，外国語学習 など
- 医療
  - 診断・治療支援，創薬 など
- メディア・広告
  - コンテンツ・シナリオ・デザイン作成や編集 など
- 芸術
  - 絵画や音楽作成 など
- IT
  - プログラミング・ソフトウェア開発・データ分析支援 など

大規模言語モデル

- 大規模なテキストデータをもとに事前学習された基盤モデルは，大規模言語モデル（large language model, LLM）と呼ばれています．
  - LLMは主にトランスフォーマ[1)]と呼ばれるニューラルネットワークのアーキテクチャに基づいています．
  - 特にテキストの生成に用いられるLLMの事前学習では， 与えられた文脈（単語（実際にはトークン）の系列）に続く次の単語（実際には確率分布）を予測するという自己教師あり学習[2)]を行います．



1) 本教材の項目「トランスフォーマ」参照
2) 本教材の項目「自己教師あり学習」参照

大規模言語モデル

事前学習

- LLMの事前学習は，与えられた単語（トークン）の系列に続く次の単語の出現確率をモデルが正しく予測できるようになるように，トランスフォーマを構成するニューラルネットワークの重み（パラメータ）を学習することで行われます．

- この学習には，大規模なテキストデータが用いられます．
  - 例えば，ウェブ，本，コード，学術文献などのテキストデータを含みます．
  - 2020年に開発されたLLMであるGPT-3[1)]では，およそ3000億トークン（数百万冊の本に相当）のテキストデータをもとに事前学習が行われ，モデルのパラメータの総数は1750億に及びます．

- LLMの学習には膨大なコストが必要となります．
  - 例えば，オープンソースのLLMであるLLaMA（2023年に公開）は，650億パラメータのモデルの学習のため，2048個のGPUで21日間の期間を要した（数百万ドルのコストに相当），と報告されています[2)]．

1) T. B. Brown, et al. Language Models are Few-Shot Learners. arXiv:2005.14165, 2020.
2) H. Touvron, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971, 2023.

指示チューニングとアライメント

- LLMの学習では，事前学習に続いて指示チューニング[1)]およびアライメント[2)]によりモデルのファインチューニング[3)]を行います．これにより，LLMは入力された指示に対して適切な応答を返すように学習されます．
  - 指示チューニング (Instruction tuning)
    - 推論，要約，質問応答などのさまざまな問題（タスク）について，入力とそれに対する適切な応答のペアをもとにモデルをファインチューニングします．
    - 多様なタスクで指示チューニングすることで，モデルが未知のタスクにも適応できるようにその能力が引き出されると考えられています．
  - アライメント (Alignment)
    - 入力に対して倫理的・社会的に適切な応答を返すように実際の人のフィードバックをもとにモデルをファインチューニングします．
    - 人のフィードバックによる強化学習（reinforcement learning from human feedback, RLHF）や直接選好最適化（direct preference optimization, DPO）などの手法が用いられます．

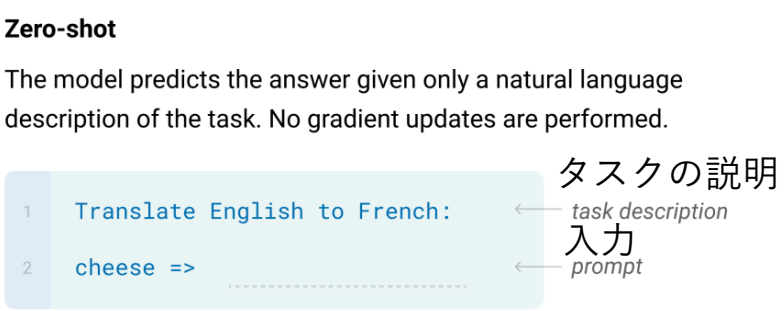
1) J. Wei, et al. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652, 2021.
2) L. Ouyang, et al. Training language models to follow instructions with human feedback. arXiv:2203.02155, 2022.
3) 本教材の項目「基盤モデル」参照

大規模言語モデル

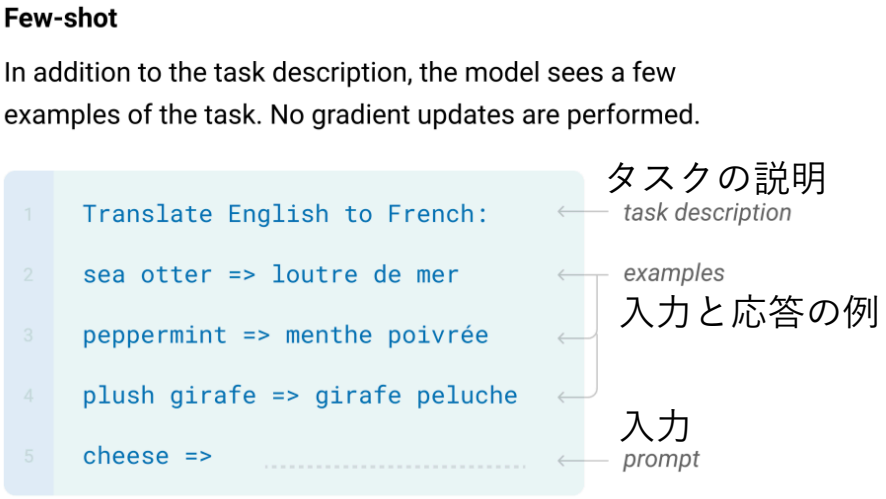
文脈内学習

- 事前学習と指示チューニングにより学習されたLLMは，対象のタスクに関する入力と応答の例をいくつか与えるだけで，そのタスクに適応させることができることがわかってきています．
- このように例の教示によって未知のタスクに適応させる学習を文脈内学習（In-Context Learning）と呼びます．

ゼロショット（タスクの説明のみを与える）



フューショット（タスクの説明に加えて，入力と応答の例を与える）



文献1)の図より一部改変して引用

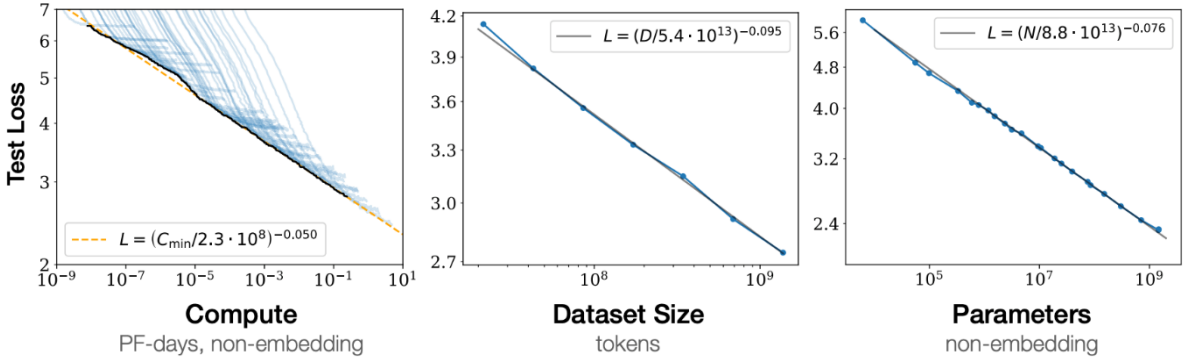
1) T. B. Brown, et al. Language Models are Few-Shot Learners. arXiv:2005.14165, 2020.

プロンプトエンジニアリング

- 文脈内学習において，入力となる指示（プロンプト）の与え方を工夫することで，LLMの能力を引き出すことができることがわかってきています．
- タスクに応じてLLMが適切な応答を返すように，さまざまなプロンプトの作り方（プロンプトエンジニアリング）が提案されています．
  - 例えば，「ロールプレイング」（Role-Playing）という方法では，LLMに特定の役割や状況をあらかじめ想定させ，そのもとでタスクを解かせます．
  - 「思考の連鎖」（Chain-of-Thought, CoT）という方法では，文脈内学習のフューショットにおいて入力から応答の例を教示する際に，その中間の過程を含めて教示することで，推論を要する複雑なタスクを解かせます．また，ゼロショットにおいても「ステップ・バイ・ステップで考えてみましょう」と指示することでLLMの適切な推論の能力を引き出すことがわかっています．
  - プロンプトエンジアリングの実践例
    - Prompt Engineering Guide（DAIR.AI）
    - 文章生成AI利活用ガイドライン（東京都デジタルサービス局）
    - 中小企業のための「生成AI」活用入門ガイド（東京商工会議所）

スケーリング則

- LLMのサイズ（パラメータの数）やデータセットのサイズ（トークンの数）や学習に用いる計算資源の量が大きくなると，スケーリング則（べき乗則）の関係に沿ってモデルの性能も向上（テストデータに対する損失が減少）することが報告されています[1)]．



文献1)の図より引用

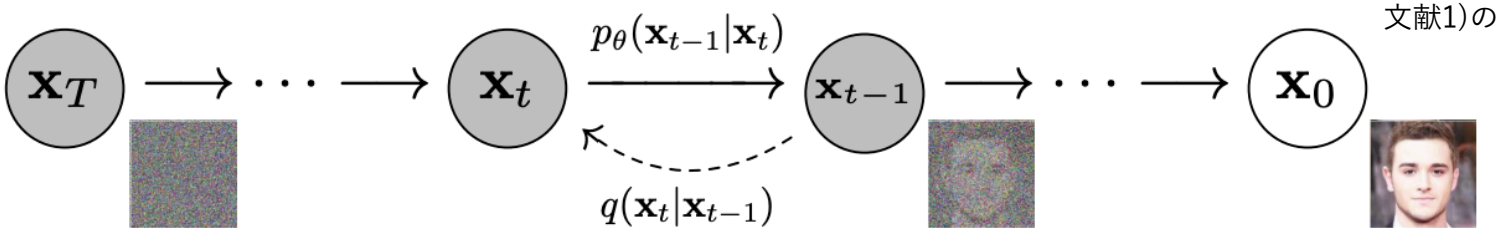
- また，LLMのモデルのサイズが大きくなるとさまざまなタスクにおいて文脈内学習の性能が向上するというLLMの創発（Emerging Abilities）が起こるという議論もなされています[2)]．
- 現在，巨大テック企業を中心にLLMの開発競争が起こっています．

1) J. Kaplan, et al. Scaling Laws for Neural Language Models. arXiv:2001.08361, 2020.
2) J. Wei, et al. Emergent Abilities of Large Language Models. arXiv:2206.07682, 2022.

拡散モデル

- 拡散モデルは，画像の生成に活用できる生成モデルです．
  - 拡散モデル[1)]では，データにノイズを付加することを繰り返す（下図の右から左の方向）拡散過程により，データをノイズへ変換します．
  - 次に，この変換を逆の方向（下図の左から右の方向）へたどり，ノイズを除去することを繰り返すことで，ノイズからデータを生成する逆拡散過程を学習します．
  - これにより，拡散モデルは高品質で多様な画像生成を可能にします．
  - またテキストで条件付けを行った画像を生成することも可能にします．

文献1)の図より引用



  - 拡散モデルを確率微分方程式で一般化した方法も提案されています[2)]．

1) J. Ho, et al. Denoising Diffusion Probabilistic Models. arXiv:2006.11239, 2020.
2) Y. Song, et al. Score-Based Generative Modeling through Stochastic Differential Equations, arXiv:2011.13456, 2020.

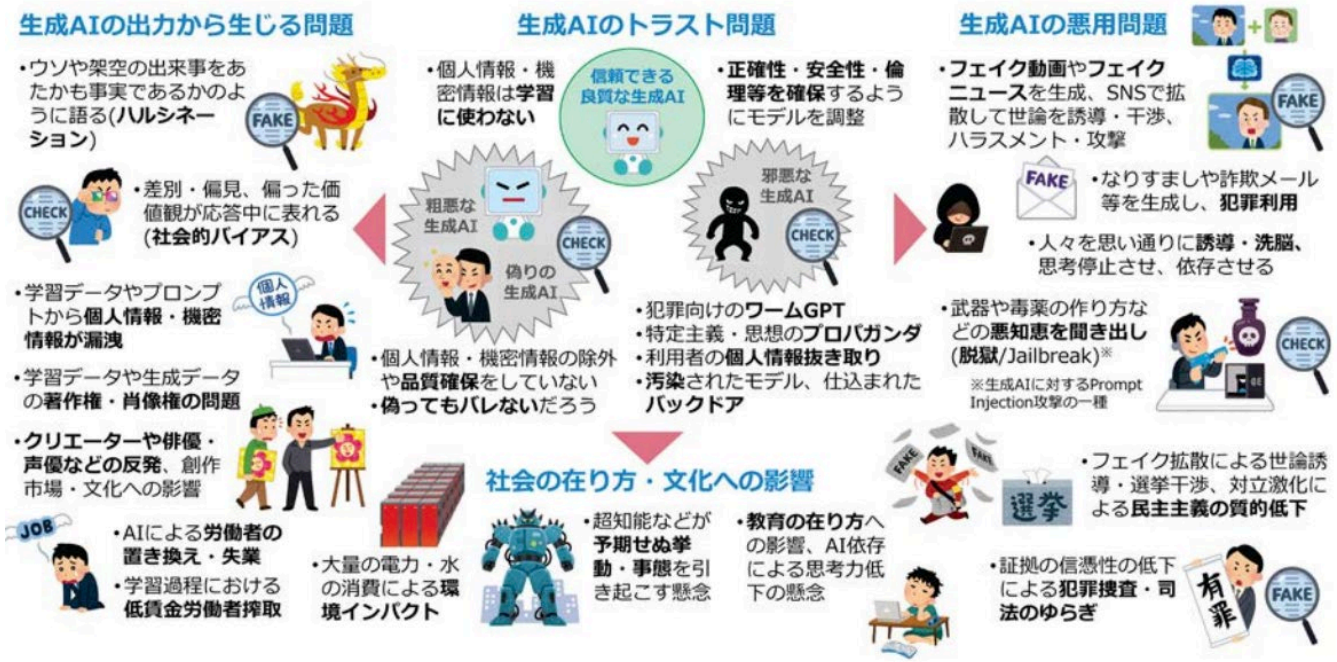
マルチモーダルモデル

- 画像-言語（Vision-Language）モデルのように異なるモダリティを横断的に扱うことを可能にした基盤モデルの技術[1)]により，テキスト，画像，音声，動画などさまざまなモダリティに対応したマルチモーダルモデルとそのサービスが開発されています.
  - マルチモーダルモデルの例
    - LLaVA（Large Language and Vision Assistant）（University of Wisconsin–Madison, Microsoft Research）
    - GPT-4V（OpenAI社）
    - Gemini Pro Vision（Google社）
    - Qwen-VL（Alibaba社）
- 2024年5月時点では，テキスト，画像，音声などを入出力可能なマルチモーダルモデルを元にしたサービスとしてGPT-4o（OpenAI社）やGemini 1.5 Pro Vision（Google社）などが公開されています.

1) 本教材の項目「CLIP」参照

生成AI技術の留意事項

- 生成AIを含む人工知能（AI）技術の急速な発展に伴い，AIのリスクとその対応に関する議論も行われています．
  - 内閣府のAI戦略会議[1)]では，個人情報保護，AIと知的財産権との関係，偽・誤情報等，雇用への影響，ガイドラインと履行確保等，をAIに関するリスクの論点として整理しています．
  - JST・CRDS（研究開発戦略センター）がまとめた次世代AIモデルに関する戦略プロポーザル[2)]では，生成AIがもたらすリスクを挙げています．



文献2)の図より引用

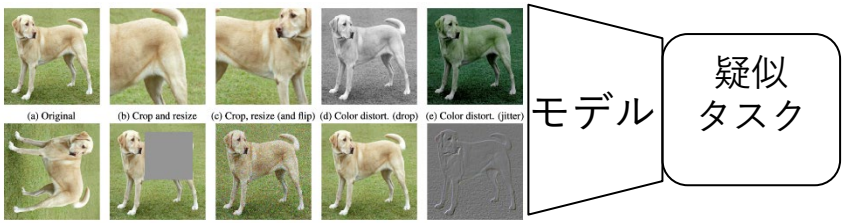
1） 内閣府 AI戦略会議， https://www8.cao.go.jp/cstp/ai/ai_senryaku/1kai/shiryo2.pdf (2024-06-01 閲覧)
2） JST・CRDS 戦略プロポーザル 次世代AIモデルの研究開発，CRDS-FY2023-SP-03, 2024.

自己教師あり学習

- 教師あり学習（特に深層学習に基づく学習）では，一般にモデルの学習のために大規模なラベル付き教師データ（人手でラベル付けされたアノテーションデータ）が必要となります．
  - 例えば，画像認識のモデルの学習使用されるデータセットであるImageNetでは1000万以上の画像データについて，各画像にどのような物体が写っているかを示すラベルが2万以上のカテゴリから選ばれ付与されています．
- 自己教師あり学習では，ラベル付き教師データを必要としない補助的なタスク（疑似タスク）によりモデルの事前学習を行います．
  - 例えば画像であれば
    - 一部を欠落させた画像の残りの部分から欠落部分を復元するタスク
    - 様々な変換を施した対の画像についてそれらが元は同じ画像か異なる画像かを予測するタスク
  - 例えば言語であれば
    - 単語が一部欠落した文脈からその欠落単語を予測するタスク
    - 与えられた文脈に続く単語を予測するタスク

自己教師あり学習

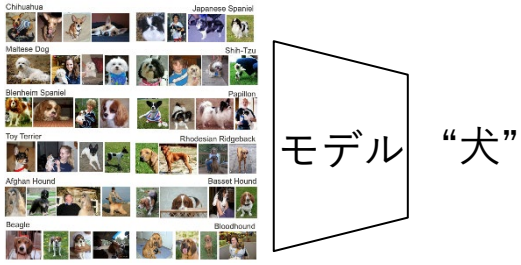
- 転移学習の考え方に従うと，自己教師あり学習において疑似タスクによりモデルの事前学習を行うことは，他のタスクの問題を解くために有用な「知識」をデータから学習していることになります．
- 自己教師あり学習で事前学習したモデルを他のタスクに適用する時，適用先のタスクを下流タスクと呼ぶことがあります．
- 下流タスクの小規模なラベル付き教師データ（例えば画像であれば画像分類のためのデータセット）があれば，事前学習済みのモデルをそれらの教師データで学習する（これをファインチューニングと呼びます）ことで，モデルを下流タスクに適応させることができます．



文献1)の図より引用

事前学習

転移学習



文献2)の図より引用

ファインチューニング

1) T. Chen, et al. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709, 2020.
2) A. Khosla, et al. Novel dataset for fine-grained image categorization. Workshop on Fine-Grained Visual Categorization, 2011.

トランスフォーマ

- トランスフォーマ（Transformer）[1]は，2017年に提案されたニューラルネットワークのアーキテクチャーです．
- 可変長の系列の情報（例えばテキストの単語の系列）を処理可能なトランスフォーマは，自己注意機構（self attention mechanism）により系列内の離れた情報間の依存関係（例えばテキスト中の離れた単語間の関係）を扱うことに可能にしました．
- トランスフォーマは入力の系列を内部表現に変換するエンコーダと内部表現を出力の系列に変換するデコーダからなります．
- トランスフォーマのエンコーダやデコーダは基盤モデルのアーキテクチャとしても用いられています．
  - 例えば，後述するBERT (Bidirectional Encoder Representations from Transformers)とGPT (Generative Pre-Trained Transformer)[2]は，それぞれトランスフォーマのエンコーダとデコーダのアーキテクチャをもとにした言語モデルです．

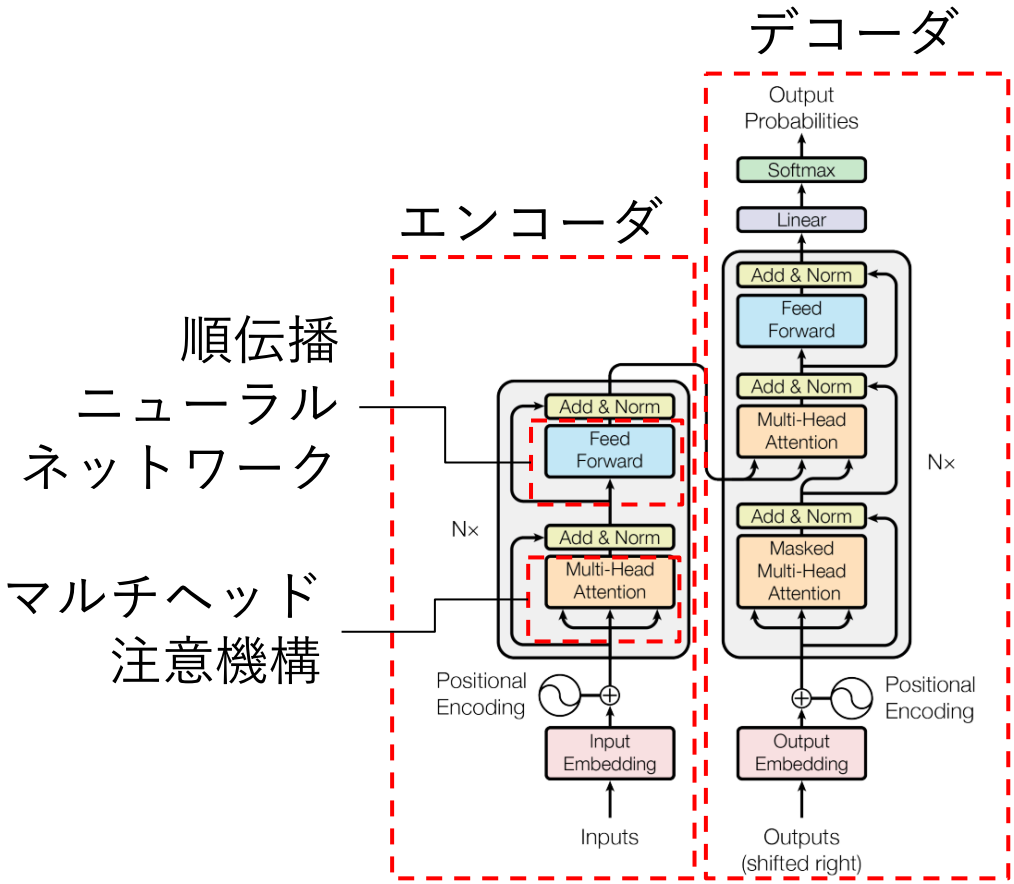
1) A. Vaswani, et al. Attention Is All You Need. arXiv:1706.03762, 2017.
2) A. Radfold, et al. Improving language understanding by generative pre-training. Technical Report, 2018.

トランスフォーマ

- トランスフォーマのエンコーダ・デコーダの主な構造は，（マルチヘッド）注意機構と順伝播ニューラルネットワークから構成される層を多層に積み上げたものになっています（次ページ図参照）．
  - 系列内の離れた情報間の依存関係を扱うことを可能にした注意機構は，タスクを解くため系列のどの情報に注目すべきかを適切に選択する役割を果たしています．
  - 順伝播ニューラルネットワークは入力層・中間層・出力層からなる多層パーセプトロンであり，情報を保存しそれらを想起するための記憶の役割を果たしています．
  - トランスフォーマでは，これらの処理を多段に繰り返すことで，必要な情報の選択と過去の情報との関連付けを行い，それらの情報をもとにタスクを解くということを行なっています．
- トランスフォーマで行われる処理は並列化可能であるため，GPUのようなハードウェアを用いて効率的にその処理を行うことができます．

トランスフォーマ

- トランスフォーマの構造

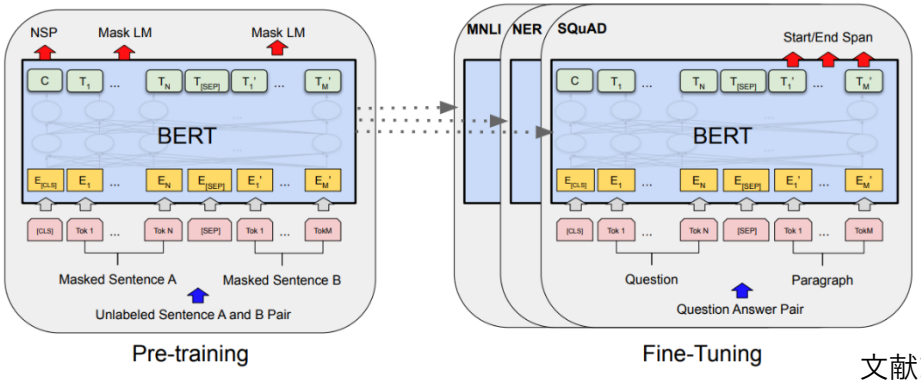


文献1)の図を一部改変して引用

1) A. Vaswani, et al. Attention Is All You Need. arXiv:1706.03762, 2017.

トランスフォーマと大規模言語モデル

- 2018年に提案されたBERT（Bidirectional Encoder Representations from Transformers）[1]は，トランスフォーマのエンコーダ部分のアーキテクチャをもとにした言語モデルです．
- BERTでは，テキスト中の単語の穴埋め問題（masked language model）および与えられた対の文が元々連続していたかという次文予測の問題を解くという自己教師あり学習で事前学習を行います．
- BERTの事前学習済みモデルを，さまざまな自然言語処理のタスクでファインチューニングすることで，モデルをそれらの下流タスクに適応させることができるようになりました．



文献1)の図より引用

1) J. Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805, 2018.

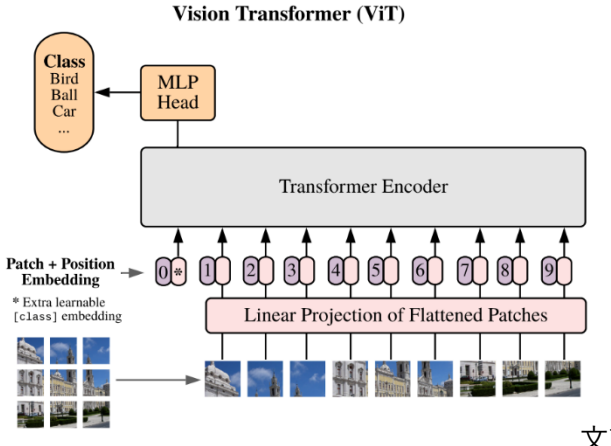
トランスフォーマと大規模言語モデル

- 2018年に提案されたGPT（Generative Pre-Trained Transformer）[1]は，トランスフォーマのデコーダ部分のアーキテクチャーをもとにした言語モデルです.
- GPTでは，与えられた文脈に続く単語を予測するという問題を解くという自己教師あり学習で事前学習を行います.
  - このように事前学習されたモデルは自己回帰（autoregressive）言語モデルと呼ばれます.
- BERT同様，GPTの事前学習済みモデルをさまざまな自然言語処理の下流タスクに適応させることができるようになりました.
- GPTはその後，事前学習に用いるデータやモデルの規模を大きくすることで現在の大規模言語モデル（LLM）の基盤となるGPT-2（2019年），GPT-3（2020年）へと発展しました.
- 2022年には，GPT-3.5のLLMをもとにしたChatGPTのサービスがOpenAI社から公開されました.

1) A. Radfold, et al. Improving language understanding by generative pre-training. Technical Report, 2018.

Vision Transformer (ViT)

- Vision Transformer (ViT)[1)]では，画像をパッチに分割してそれらをトークンとしてトランスフォーマのエンコーダに入力し，画像分類タスクを解く事前学習を行います．
- ViTは画像データでの自己教師あり学習（例えば画像の欠落を復元するなど）による事前学習に用いることもできます．
- ViTの事前学習済みモデルをさまざまな画像認識のタスクでファインチューニングすることで，モデルをそれらの下流タスクに適応させるができます．

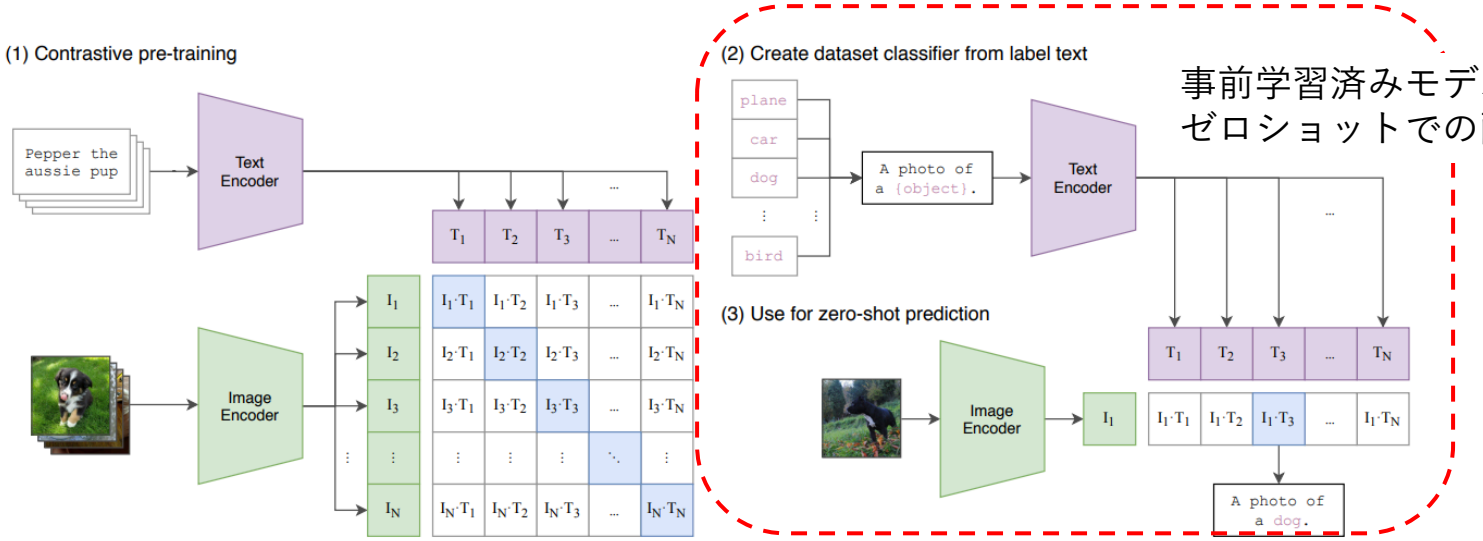


文献1)の図より引用

1) A. Dosovitskiy, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929, 2020.

CLIP

- CLIP (Contrastive Language-Image Pretraining)[1] は，大規模な画像とテキストのペアをもとに学習されたマルチモーダル（画像-言語）の基盤モデルです．
- テキスト⇄画像の生成へ応用されています．
  - テキストと画像それぞれをエンコーダ（例えばトランフォーマとViT）で内部表現（ベクトル）に変換します．
  - 対となる画像とテキスト間のベクトルの内積が大きくなるように事前学習を行います．

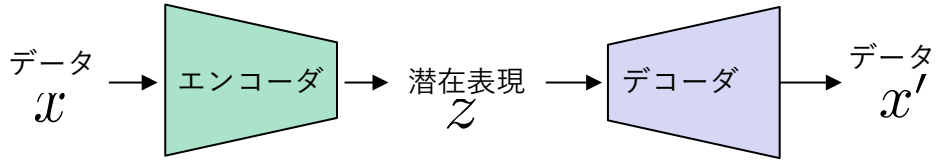


文献1)の図を一部改変して引用

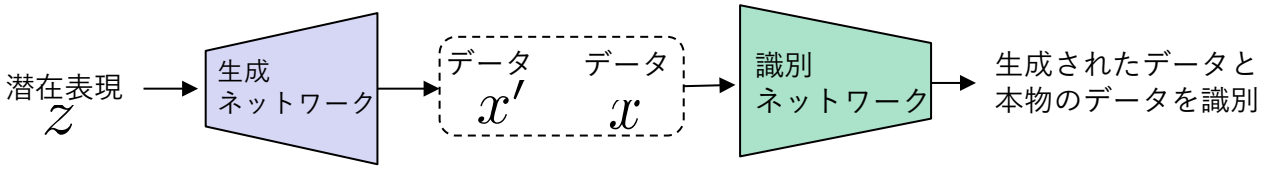
1) A. Radford, et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020, 2021.

生成モデル

- データの同時確率分布をモデル化する手法は，生成モデルと呼ばれます．
- 生成モデルにより新たなデータサンプルを生成することができます．
  - 特に深層学習の技術に基づく深層生成モデルでは，深層ニューラルネットワークにより，データとその潜在表現と間の複雑な写像を学習します．
  - 深層生成モデルの技術は生成AIにも活用されてます．
    - 代表的な深層生成モデル
      - 変分オートエンコーダ（Variational Autoencoder，VAE）



      - 敵対的生成ネットワーク（Generative Adversarial Networks, GAN）



      - フローベースモデル（Flow-based Model）