「統計的機械学習」の中核としての
統計数理シンポジウム 2023/05/25

生成モデルは世界を
どのように理解しているのか

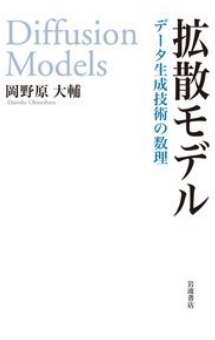
株式会社 Preferred Networks
岡野原 大輔
@hillbig

アジェンダ

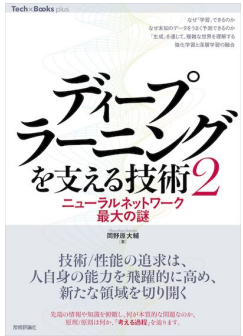
- 現在の代表的な生成モデル
  　大規模言語モデル/ 拡散モデル
- 自己教師あり学習 / メタ学習
- 未解決問題

関連書籍

岩波書店 2023
一般向け



岩波書店 2023
専門家向け



技術評論社 2021 2022
ディープラーニングの基礎知識

日経BP 2022
個別の深い話題

生成モデル

$x \sim p(X \mid C)$

X: 生成対象　C: 条件

- 生成モデル：対象ドメインのデータを生成できるようなモデル
  - テキスト、画像、動画、化合物、行動列 等
  - 条件を通じて、制約、指示、対象ドメインなどを指定する（条件付き生成モデルの方が学習の面でも使いやすさの面でも有利であり一般的に使われる）
- 大量の学習データ、手法改良により高忠実かつ多様なデータを狙った形で作れるようになってきた

生成モデルは事前学習時に学んだこと以外も実現できる汎用性を持っている

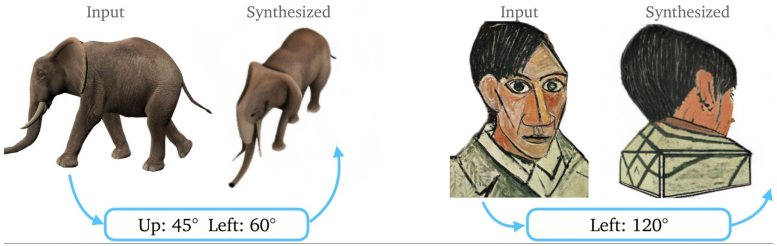
- 生成モデルを使って、指示し新たなタスクをゼロショット、フューショットで実現できる　（左図）
- また、事前学習時には知らない新たな制御方法も後で追加できる
  - 例：2D画像のみで学習後、3次元の視点変化を学習（右図）

```
Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese =>
```

タスクをその場で学習する能力
in-context learning
（本文中学習）

GPT-3によるゼロショット能力の発見[Brown+ 2020]

生成能力を失わず後付けで別の条件付も学べる能力



[Liu+ 2023]

現在の生成モデルの代表例
大規模言語モデル
拡散モデル

大規模言語モデル (1 / 4)
言語モデル

単語列* $w_{1:n} = w_1 w_2 w_3 ... w_n$の自己回帰モデルによる生成モデル

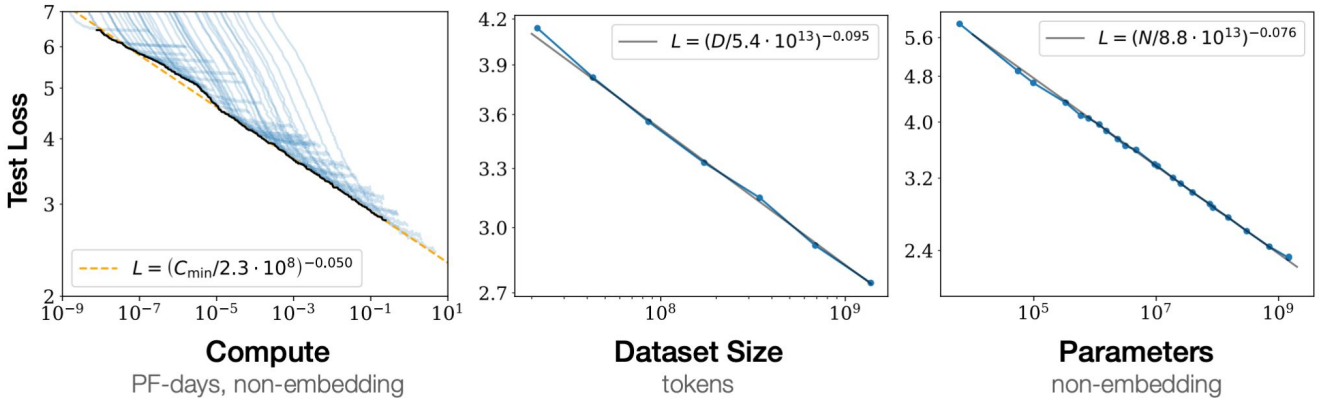
$p(w_{1:n}) = \Pi_i \, p(w_i | w_{1:i-1})$

Transformerとよばれるモデルを使って条件付確率をモデル化

利用用途：
　単語列の尤もらしさを評価できる → 機械翻訳, 音声認識
　後続単語を生成できる　→　現在のAI対話システム（ChatGPTなど）

* 現在はBPEなどで自動決定されたバイト列の塊のトークンを単語として言語モデルを作

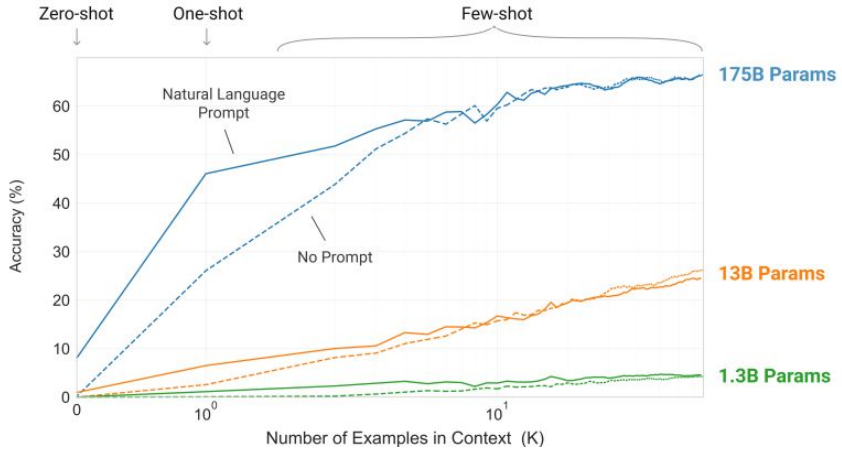
大規模言語モデル (2 / 4)
言語モデルのべき乗則 [Kaplan+ 2020]



言語モデルでTransformerをモデルとして使った場合に「学習時投入計算量」「学習データ量」「モデルサイズ」と、「検証データのクロスエントロピー損失」との間にべき乗則が成り立つ

注：以前は最適なモデルサイズが他の因子の伸びより大きいと考えられていた、最近ではいずれも比例が最適と考えられる [Hoffmann+ 2022] [Google 2023]

大規模言語モデル (3 /4)
ゼロショット学習、フューショット学習

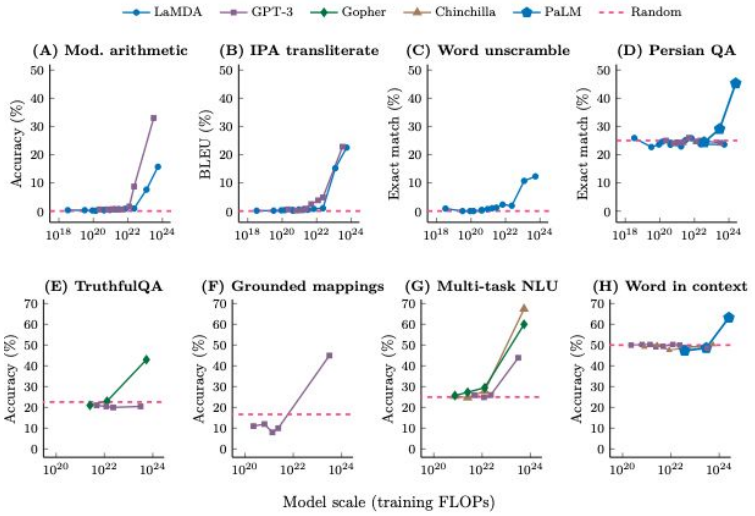


[Brown+ 2020]

大規模言語モデルはゼロショット/フューショット（追加学習データ無し/少数学習データ）で新しいタスクを解ける

高い分布外汎化能力を持ち、多くのタスクで大量の教師ありデータで学習をした場合に匹敵する性能を達成できる

大規模言語モデル (4/4)
大規模化による創発



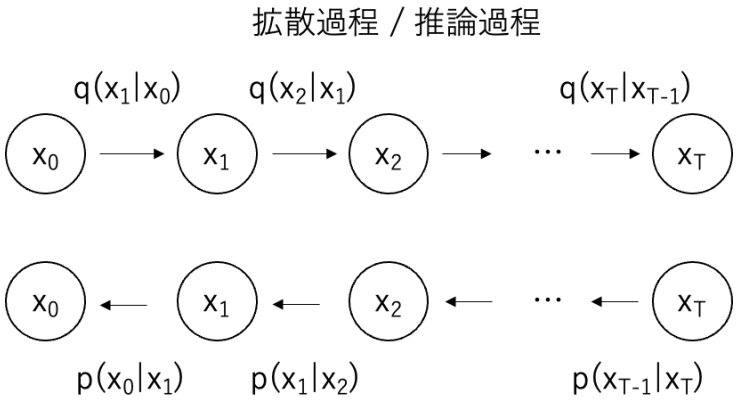
大規模化していくだけで、
様々な能力が創発する
[Wei+ 2022]
創発：後続タスクの性能が急激に改善

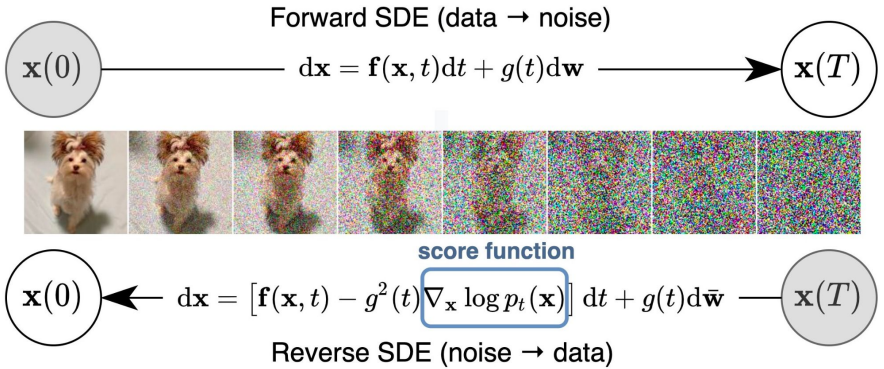
- 論理思考/質問応答/抽象的思考
本文中学習による事前知識の上書き能力等

突然創発しているのか滑らかに創発するのかは議論がある
[Schaeffer+ 2023]

拡散モデル (1/4) [Sohl-Dickstein+ 2015] [Song+ 2019] [Ho+ 2020]

- 非平衡熱力学を源流に持つ、深層生成モデルの一種
- データにノイズを徐々に加えていく拡散過程を逆向きに辿る逆拡散過程（生成過程）によって生成モデルを定義する
- データを破壊することで生成方法を学習する





https://yang-song.net/blog/2021/score/

midjourney v5の出力例 Yuki Homma @y__homm https://twitter.com/y__homm/status/1636186478899494912

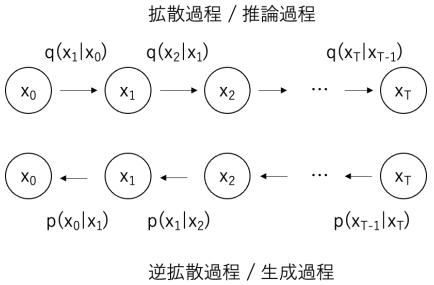
拡散モデル (2/4)

拡散モデルは複数の確率層からなるVAE

- 拡散過程（固定の推論）
- 逆拡散過程（生成）
- データの対数尤度の変分下限（ELBO）最大化で学習

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$



$$\log p_\theta(\mathbf{x}_0) = \log \int p(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$$

$$\geq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log p(\mathbf{x}_T) + \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] := L(\theta)$$

データを拡散して破壊した時に、元に復元できる経路を求める
物理の言葉でいえば、事前分布から目標分布へ変換する経路の中で
発生する散逸（自由エネルギー減少）が最小の経路を求める

拡散モデル (3/4)
拡散モデルを使った学習と推論（生成）

学習

データに様々な強度のノイズをのせ、デノイジングできるよう学習する

推論（生成）

完全なノイズからデータをサンプリングし、それをデノイジング強度を下げながらデノイジングするのを繰り返す

局所解に陥らないよう各強度の撹乱後分布でランジュバンモンテカルロを使ってサンプリングするスコアベースドモデル [Song+ 2019]と拡散モデルは目的関数の係数などを除いて一致する [Ho+ 2020]

拡散モデル (4/4)
拡散モデルは多くのタスクに使える

- 編集として補完、超解像、Zero-shot編集が可能
- その他アプリケーションとして密度推定、非可逆圧縮、敵対的摂動頑健性向上、最適化などでも最高精度を達成
- 学習時に使わなかった別情報での条件付生成を少量のデータで適応できる
  深度、3D、スタイル

ControlNet [Zhang+ 2023]

本講演でとりあげる話題

話題１：
　なぜ生成を学習するだけで生成とは関係ない情報も理解しているのか

話題２：
　なぜ生成モデルの分布外汎化能力が高いのか

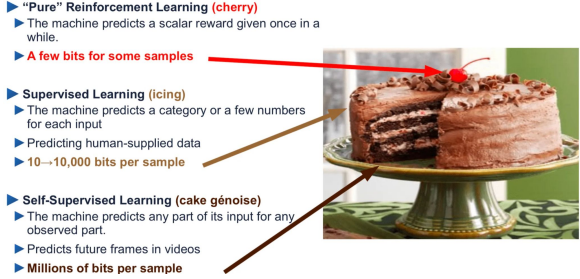
本講演でとりあげる話題

話題１：
　なぜ生成を学習するだけで生成とは関係ない情報も理解しているのか

→　生成というタスクによる自己教師あり学習を通じて、データを理解する必要に駆られているから

話題２：
　なぜ生成モデルの分布外汎化能力が高いのか

自己教師あり学習



[Lecun 2019]

正解がタダでいくらでも手に入る教師あり学習問題

- 予測：過去から未来を予測する
- 欠損補間：一部を欠損させ残りから欠損を予測
- 対比：意味が同じものと違うものを対比させる

そのタスクの達成自体が目標ではなく、そのタスクを達成するための副作用として別の能力を獲得する

教師あり学習と違って、膨大かつ多様なデータを利用でき、特定タスク向けでないデータ理解ができるようになる

自己教師有り学習
大規模言語モデルの場合

「こうしたことから、私は父と一緒に＊＊＊へ行き相談した」

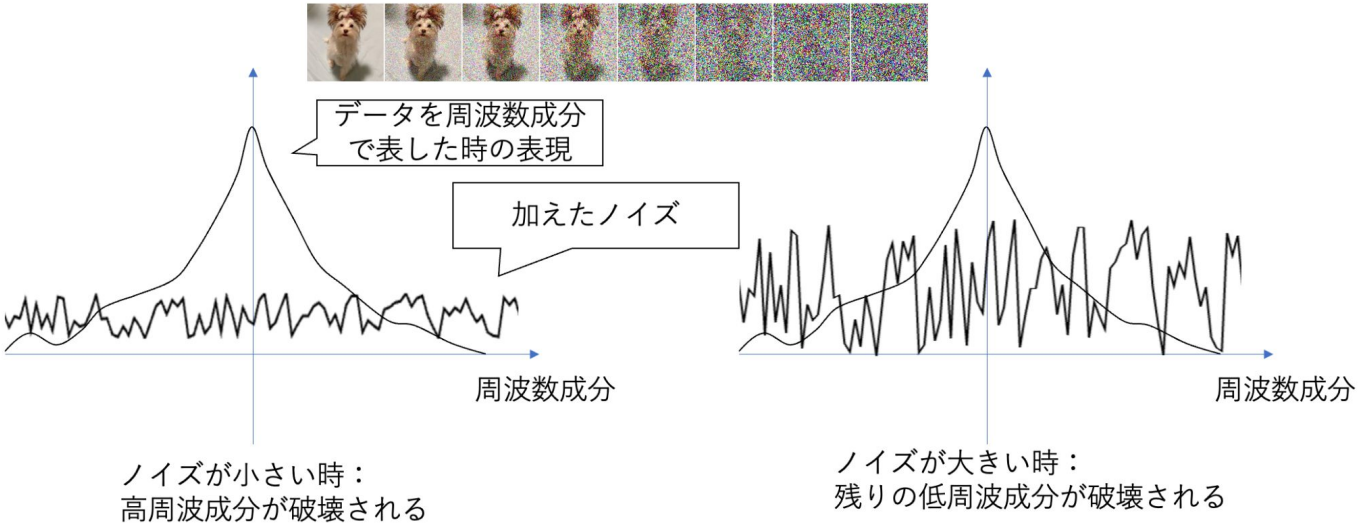
***に入る単語を予測するには？
（私や父はどういう人か、こうしたこととは何か、行って相談する場所はなにか　など）

- 次の単語をうまく予測するために文やその背後の情報を理解する必要に駆られる
  - 学習過程で最初は単語の意味、それができれば句や文の意味、その文の背後にある情報を理解していく
  - うまく予測できることと、理解することは完全に一致しないが、多く一致する

拡散モデルの場合

様々な強度でのデノイジングは画像理解を必要とする

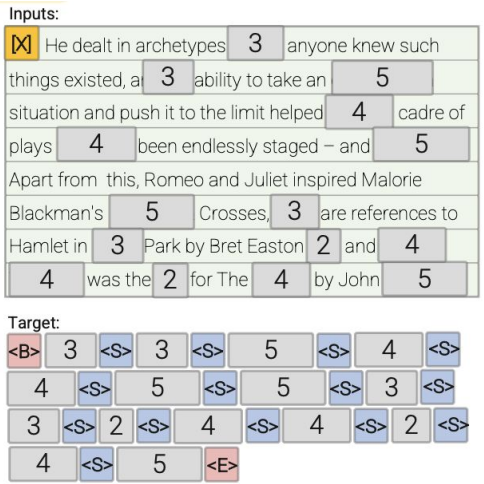
- 初期（ノイズ小）は高周波成分、詳細の理解
- 後期（ノイズ大）は低周波成分、画像全体の理解



生成以外の自己教師あり学習も有効

生成は有効な自己教師あり学習だが、
必ずしも生成ではないタスクも有効

- 例: 様々なマスク補完 UL2[Tay+ 2022]（右図）



しかし、生成タスクは対象の理解を測る上でとても優れており、自己教師あり学習の中心的なタスクになり続ける

”What I cannot create, I do not understand”, [Feynman]

本講演でとりあげる話題

話題１：
　なぜ生成を学習するだけで生成とは関係ない情報も理解しているのか

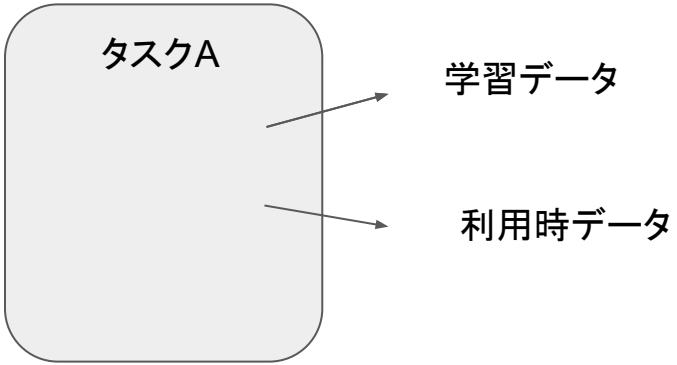
→　生成というタスクによる自己教師あり学習を通じて、データを理解する必要に駆られているから

話題２：
　なぜこれらのモデルは分布外汎化能力が高いのか
→　Transformerと逐次的な生成が（意図せず）メタ学習を実現し、問題にあわせてモデルを急速に適応する能力を獲得しているから

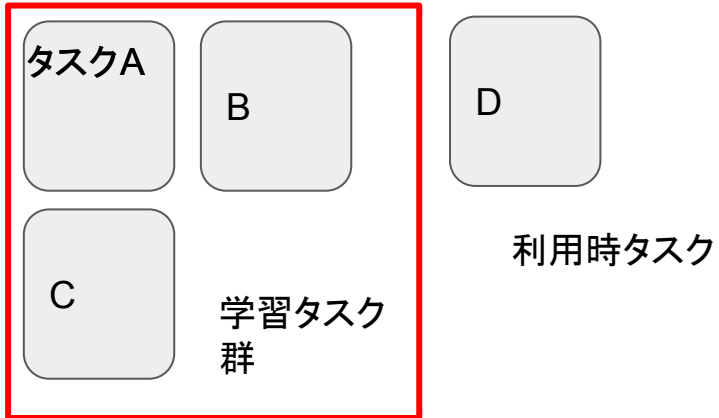
メタ学習

「学習の仕方」を学習する手法。現在のタスクに応じてモデルパラメータを急速に適応する手法が一般的。MAML[Finn 2017]等
　一つの固定したモデルで汎化することを諦め、その場で適応できる能力を持って分布外汎化を目指すアプローチといえる

従来の問題設定
単一タスクで、学習データと利用時データが
同じ確率分布から得られている場合の汎化を目指す



メタ学習の問題設定
たくさんのタスクで、一部のタスク群で学習の仕方を学習し、新しいタスクの一部が与えられた時、そのタスクに急速に対応する



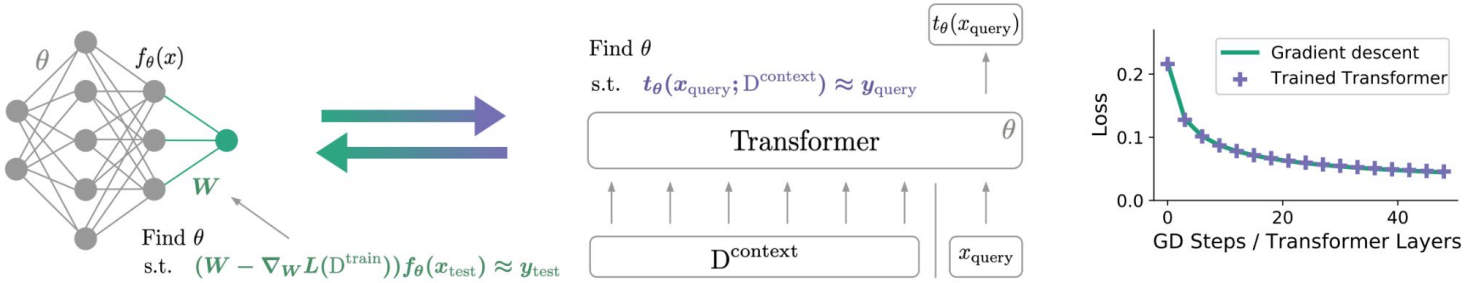
自己回帰モデル+Transformerはメタ学習をシミュレーションしている

本文中学習はメタ学習によって実現している

- 言語モデルは問題毎にモデルを急速に適応している
  - 前から順に予測する毎に結果に応じ適応する
- 拡散モデルも条件部がTransformerで適応している

- 本来であればモデル（パラメータ）は学習時のみ更新し、利用時には更新できないが、注意機構がパラメータを変更させるのと同じ役割を果たしている
  （次ページ）

本文中学習は勾配降下法を少なくともシミュレーションするだけでなく、それを超える能力を獲得



Transformerの各層が、勾配降下法と同じ役割を果たし、これまでの生成した内容に応じて適応している [Oswald 2022]
→ プロンプトで役割を指定したり例示するとモデルが変わる

事前知識を本文中学習で上書きする能力、事前知識にない抽象的な入出力関係を学習する能力を獲得する [Wei+2023]

今後の課題

世界の理解の限界

- 自己教師あり学習由来で世界を理解しているため、それで得られない情報の理解はできない
  - ハルシネーション（嘘回答）の根本的な解決は現在の枠組みでは難しく、別の（自己）教師信号や推論が必要になるのでは
  - 一方、工学的には対話システムは外部サービスとつながることで解決できる部分も多い
- 拡散モデルは実際の世界の生成過程を模倣していないので、究極的な因果関係の理解はできない
  - 条件付をすることで生成を分解し、その上で操作できるだけ
- 言語モデルはそもそも真の生成過程がよくわかっていない

今後の課題

・ハルシネーション対策
    ・ 汎化の問題と究極的には関わる可能性
・条件付学習の成功や限界の理解
・本文中学習/メタ学習の理解
    ・ i.i.dではない問題設定でどのように問題を定式化し、理論的な保証を与えることができるのか
・真の生成過程は獲得できるのか
・生成モデル自体ではなく、その内部モデルが持っている知識をどのように制御、活用できるのか

文献

[Liu+ 2023] “Zero-1-to-3: Zero-shot One Image to 3D Object”, arXiv:2303.11328

[Sohl-Dickstein+ 2015] “Deep Unsupervised learning using nonequilibrium thermodynamics”, ICML 2015

[Song+ 2019] “Generative Modeling by Estimating Gradients of the Data Distribution”, NeurIPS 2019

[Ho+ 2020] “Denosing Diffusion Probabilistic Mdoels”, NeurIPS 2020

[Tay+ 2022] UL2: Unifying Language Learning Paradigms, arXiv:2205.05131

[Finn+ 2017] “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”, ICML 2017

[Ozwald 2023] “Transformers learn in-context by gradient descent”, arXiv:2212.07677

[Schaeffer+ 2023] Are Emergent Abilities of Large Language Models a Mirage?, arXiv:2304.15004

[Brown+ 2020] Language Models are Few-Shot Learners, arXiv:2005.14165

[Kaplan+ 2020] Scaling Laws for Natural Language Models, arXiv:2001.08361

[Hoffmann+ 2022] Training Compute-Optimal Large Language Models, arXiv:2203.15556

[Google+ 2023] PaLM 2 Technical Report, https://ai.google/static/documents/palm2techreport.pdf

[Wei+ 2022] Emergency Abilities of Large Language Models, TMLR 2022