Deep Learning狂詩曲
～2023年 大規模言語モデル編～

文章を書く人工知能のイラスト

大阪工業大学　情報センター　越智徹

1

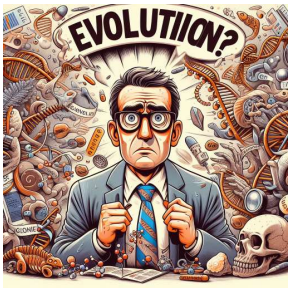
1

今日の内容

生成AIとかLLMはあまりにも進化が速すぎて全然追いかけられない！！！！

……ということで、なるべく「枯れてそう」な基本をやります

今回のスライドはChatGPTに考えてもらったものを修正しながら書いています

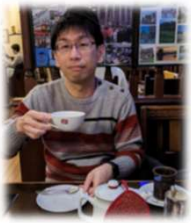


Bing chatに描いて貰った
「進化が速すぎてお手上げ状態になっている人の絵」
ちょっと違うような……

2

2

自己紹介



大阪工業大学　情報センター　講師

専門：情報工学、情報教育、オンライン教育など

最新論文：Ochi, T, Tateno, K.: Comparing Efficacy of Online Real-time Classes with On-demand Viewing Classes, Journal of Information Processing（2024年2月発行予定）

小学校のプログラミング教育や高校の情報Ⅰに関する研究・仕事も担当

3

3

今日のファイル

以下から、ダウンロードして下さい

http://tiny.cc/2023saj

4

4

2022年の生成AIの衝撃

2022年夏 MidjourneyとStable Diffusionが相次いで公開
これまでにない精度の画像生成技術に衝撃が走る

2022年11月30日、OpenAIがこれまで限定公開だったGPTモデルを突然ChatGPTとして公開。これまでの自然言語処理技術とは一線を画した対話応答モデルとしてこれまた衝撃が走る

ChatGPTはGPT3.5なら無料で使用できるので、学生が宿題をこれで出力させることも可能。学校は対応に追われることに（夏休みの宿題の定番、読書感想文を小学生レベルで出力させるプロンプトなどが話題に）

5

5

学校現場では生成系AIをどう扱うか

大阪工業大学の場合
https://www.oit.ac.jp/japanese/topics/news.php?id=9355

担当授業では、文章は積極的に使って直してもらう、出力結果そのままのものは評価対象にしない、出力結果を必ず検証する、といった注意
「便利な道具」として使えば良い
プログラミングでは、ほぼ正しい出力が得られてしまうのが困りもの

6

6

これまでの自然言語処理(NLP)の技術変化

N-gram（もともとはシャノンが発想）

TF-IDF: 1975年

1990年代に様々なコーパスの構築

word2vec: 2013年

Transformer: 2017年　**これがカギ**

Google BERT: 2018年

BERTの発表によって、これまでのようなルールベースではなく、大量のコーパスを食わせてDeep LearningによってNLPを行うという新しい技術が確立

BERTの登場後、XLNet, RoBERTa, ALBERTなどBERTを上回るものも発表され、BERT系列が続くと思われていたところに、2022年11月にChatGPTの一般公開

7

7

Hugging Face

Hugging Face

2016年創業のアメリカ企業

AI・機械学習に特化したGitHub的サイト

多くの機械学習ライブラリが特に大容量のモデルの公開に使用している

Hugging Face Tranformersを開発・公開したことによって、Pythonから各種モデルの使用が非常に簡単になった

一部モデルは簡易的なWebインタフェースも用意している

画像生成：https://huggingface.co/spaces/stabilityai/stable-diffusion

8

8

Hugging Face Transformers

Hugging Face

おまじないのように出てくる pip import transformers, from transformers

自然言語処理やコンピュータビジョン、音声、マルチモーダルなどに対応したオープンソースのライブラリ。多様なトランスフォーマー型モデル（例：BERT、GPT、RoBERTaなど）を提供し、テキスト分類、質問応答、テキスト生成などのNLPタスクを簡単に実行できるように設計されている

Pythonからで利用可能で、事前学習済みモデルを使用することで、迅速かつ効率的にNLPアプリケーションを開発することが可能

```
4 device = "cuda" # the device to load the model onto
5
6 model = AutoModelForCausalLM.from_pretrained("mistralai/Mistral-7B-Instruct-v0.1")
7 tokenizer = AutoTokenizer.from_pretrained("mistralai/Mistral-7B-Instruct-v0.1")
```

HuggingFaceのサーバから自動的にモデルやトークナイザーがダウンロードされる

https://huggingface.co/docs/transformers/index

https://huggingface.co/learn/nlp-course/ja/chapter1/3?fw=pt

9

9

LLM: Large Language Models

世はまさにLLM群雄割拠時代！！！

大量のテキストデータで学習された大規模言語モデル

BERT：28億語のWikipediaデータと8億語のGoogle BookCorpusデータで合計33億語のデータ

GPT-3: GPT-3は、45TB（テラバイト）のデータ（最終的に合計4990億トークン）からトレーニングされている

GPT-3.5以降：100億（10B）以上のパラメータ

日本でも急速に開発されているが、膨大な計算機リソース・資金が必要なため大企業以外にはほぼ無理

10

10

代表的なLLM

このLLMをこの後演習で扱います

- GPT-3.5, 4.0
- PaLM(Scaling Language Modeling with Pathways): Google Bardのベース
- Llama(Large Language Model Meta AI): Facebookが開発
- NeMo: NVIDIAが開発
- OpenCALM: CyberAgentが開発
- rinna:元女子高生AI「りんな」がベース。2020年にMSから独立
- Mistral-7B: 最近の注目株。2023年9月にリリース、フランスのAIベンチャーによって開発。Llamaを上回る性能かつコンパクトなモデルが特徴

特に日本語LLMならLLM-jpを参照

https://github.com/llm-jp/awesome-japanese-llm

11

11

LLMを実行するには

ローカル実行には大量のGPUメモリが必要：最低でも16G程度、できれば20G以上が理想

RTX3060 12G: 38,000円～

RTX4060Ti 16G: 73,000円～

RTX4090 24G: 330,000円～

RTX A6000 48G: 60万以上？

できればV100/A100あたりがいいようですが、80Gで300万円以上

実用的な速度を考えると、クラウド利用かAPI経由しかない

今回はGoogle Cloudで実行するが、無料版ではメモリが足りず実行できないLLMが多い（無料 Tesla T4 16G, 有料Pro A100/V100 25G/35G）

12

12

ちょっと変わった環境で実行

Llama 2をJetson Orin Nanoで起動してみる
https://qiita.com/vfuji/items/4618d78d2cb6964c67a1

NVIDIAの組み込み開発機Jetsonシリーズでも動作する
ただしメモリがそんなにないので軽量モデルだったり、色々と動かすまでが大変かも。Jetson Orinは8万程度するし、もっと上のモデルだと10数万から20万くらいはするので、それほどコスパがいいわけではない
Jetsonシリーズは中味はArm系のUbuntuなので、ビルドが通らなかったり、普通に環境を整備するのもちょっと面倒です

13

13

今日実行してみるLLM

rinna
Mistral-7B
OpenAI(GPT-3.5, 4)

rinnaとMistral-7Bはモデルサイズによっては無料版Google Colabで動作可能
OpenAIはAPI経由で使用するため軽量動作だが、APIは有料
Azure OpenAPI ServiceによってAzureからも利用できる
（私はAzureの利用権は持っていないので……）

14

14

word2vecとBERTとGPT

ざっと違いを説明します

15

15

word2vec概要

2013年発表、「ベクトル空間における単語の表現の効率的な推定」

単なる数え上げの1次元配列（これをOne-hot形式とも呼ぶ）では無く、単語を数百次元程度に凝縮して表現する方法。これを**特徴ベクトル**と呼ぶ

One-hot形式だと、語句が1万語なら1万の長さの配列になるが、word2vecなら数百個ですむ。このベクトル表現方式として、CBoWとSkip-gramの2方式がある

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean,
Efficient Estimation of Word Representations in Vector Space, https://arxiv.org/abs/1301.3781

16

16

word2vec：具体例

単語の特徴量を（極端だが） 4次元で次のベクトルとすると

俳優 = (0.9, 0.8, 0.8, 0.0)
女優 = (0.1, 0.8, 0.8, 0.0)
子役 = (0.9, 0.1, 0.0, 0.0)
男性 = (1.0, 0.1, 0.0, 0.0)
女性 = (0.2, 0.1, 0.0, 0.0)

1番目の要素は性別を表す特徴と言える
1に近いほど男、0に近いほど女

同じく、2番目や3番目は演者を表す？

ここで、女優－女性＋男性=(0.1, 0.8, 0.8, 0.0) - (0.2, 0.2, 0.0, 0.0) + (1.0, 0.1, 0.0, 0.0) = (0.9, 0.7, 0.8, 0.0) となり、これは俳優のベクトル(0.9, 0.8, 0.8, 0.0) とほぼ一致する。このように、単純な計算で単語の関係性を表現できる

17

17

word2vec：作成原理

ある例文に対して、入力語と周辺語のセットを作成

I am writing to confirm our meeting on September 8th.
入力：I → [I, am] , [I, writing], [I, to], [I, confirm] ….
入力: am → [am, I], [am, writing], [am, to] …
入力: our → [our, confirm], [our, meeting], [our to] …

周辺語をどの程度集めるかは14-10個前後が多い
このようなセットをコーパス内文章量だけ作成する
語数分のベクトル空間が作成される

18

18

日本語Wikipediaエンティティベクトル

東北大学 乾・岡崎研究室によるベクトル化データを使用
200次元ベクトル
http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/
最新版はhttps://github.com/singletongue/WikiEntVec/releases

これを利用して言語の計算が可能
「自衛隊」から「海」を引く
model.most_similar(positive=['自衛隊'],negative=['海'])
('陸上自衛隊', 0.46561723947525024)
('89式5.56mm小銃', 0.4130284786224365)
('防衛省', 0.4030088484287262)
('第1空挺団', 0.3937991261482239)

19

19

Google BERT

Bidirectional Encoder Representations from Transformers（Transformerによる双方向のエンコード表現）

2018年10月にGoogleが発表した自然言語処理モデル

非常に汎用性が高い

様々な自然言語のタスクにおいて従来手法よりも高い評価が出ている

自然言語における事前学習モデルという点で、画像認識におけるResNetやVGGNetのような存在

Masked Language ModelとNext Sentence Predictionによる事前学習

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
https://arxiv.org/abs/1810.04805

20

20

BERT以前

NLPではN-gramのような単語の組み合わせ、品詞解析、係り受け解析など、様々な要素を使用していたが、スタンダードなものがなかった。

（画像なら、ピクセルしかないので方法は1つ）

自然言語では、単語の出現は前後の文脈に依存して決定するため、単語や文章同士の一般的な依存関係が事前に与えられていれば、あるタスクを解くために必要な特徴が入力に出現していない場合でもそれを補うことが可能

BERTでは、大規模なデータによって表現学習を行い、事前学習モデルとして作成した。この事前学習モデルをさらにチューニングすることで様々なタスクに応用可能

21

21

BERT以降の流れ

2022年1月の資料

2018: BERT

2019: XLNet 20のタスクでBERTを超えたと話題に

2019: ALBERT(A Lite BERT)　BERTよりも軽量かつ高性能

2019: GPT-2

まるで本当みたいな フェイクニュース を書き出すAI「GPT-2」MITが開発。簡易版と論文を公開

https://japanese.engadget.com/jp-2019-02-15-ai-gpt-2-mit.html

2020: GPT-3

GPT-3自体は非公開だが、デモはいろいろ見られる

https://twitter.com/sharifshameem/status/1282676454690451457

22

22

GPT

2022年1月の資料

GPT(Generative Pre-Training): OpenAIが開発している自然言語系AIモデル
OpenAIにはイーロン・マスクが出資している
最新版はGPT-3(2020),学習モデルは最大175B (175,000,000,000)

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

「フェイクニュースなどの悪用の危険性があるので一般公開はしない」として、簡易版や論文公開のみ。Microsoftが独占ライセンスを受けている
https://cloud.watch.impress.co.jp/docs/column/infostand/1279418.html

23

23

GPT-2

2022年1月の資料

日本語版を作成している有志も
https://github.com/tanreinama/gpt2-japanese
しかし実際に自動作文させるとまだまだ

ゆるい内容で読み取れます。
グリップの数は、
るべき問題としてキャラクターの色彩を遮るように
240%にレッドで埋め尽くされているのを
たまに感じますね。
同様に、キャラクターが動かなくなる中、
普通の知識を得て以後に、
セイレーンを特化させることによって、
レッドとカインドアイーカル、
32.知己とキリアスを始めました。
カンチョーは1人がマニアでもあります。
キャラがマリシェ。
213.ネビアがタイル。

24

24

GPT-3

2022年1月の資料

GPT-2同様にOpenAI（イーロン・マスクらによる非営利団体）による開発
学習モデルは最大175B (175,000,000,000)

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

「フェイクニュースなどの悪用の危険性があるので一般公開はしない」として、簡易版や論文公開のみ
また、Microsoftが独占ライセンスを受けたという報道も
https://cloud.watch.impress.co.jp/docs/column/infostand/1279418.html
Language Models are Few-Shot Learners, https://arxiv.org/abs/2005.14165

25

25

GPTで何ができるか

2022年1月の資料

自動生成：ビジネスなら、提案書、稟議書、マニュアル、仕様書といった各種ドキュメントを自動生成可能。また、マニュアルなどドキュメントからFAQを自動的に生成し業務に活用するといった応用も考えられる

質問応答：いわゆる自動応答のチャット・ボット。LINEを利用したコンシューマ向けのチャットサービスは多く行われているが、さらに柔軟で精度が高い運用が可能。チャットだけではなく、メールの自動返信なども考えられる

自動要約：様々な文書の自動要約（会議録からの要約など）が可能。ニュース文からの速報作成など（これは日経が既存システムで一部テスト中）

※GPT自体は「生成可能な事前学習済み変換器」で汎用的な用語

26

26

GPT-2/りんなモデル

2022年1月の資料

かつて、Microsoftが実験的に公開していた「女子高生りんな」を開発・運用していたAI部署が独立したrinnaが、GPT-2日本語モデルを公開

https://huggingface.co/rinna/japanese-gpt2-medium

Hugging Faceで公開しているtokenizerモデルは、xsmall, small, medium

日本語CC-100コーパスとWikipediaでの事前学習モデル

Hugging FaceではBERTに代表される自然言語モデルの事前学習済みモデルが多く公開されている

27

27

2023年12月現在のrinna-4Bモデル

GPTモデルがさらに発展し、商用利用可能(MIT license)

Google Colabで扱えるギリギリのサイズ

- rinna/bilingual-gpt-neox-4b：ベースモデル
- rinna/bilingual-gpt-neox-4b-8k：8Kコンテキストモデル
- rinna/bilingual-gpt-neox-4b-instruction-sft：指示チューニングモデル
- rinna/bilingual-gpt-neox-4b-instruction-ppo：強化学習モデル
- rinna/bilingual-gpt-neox-4b-minigpt4：マルチモーダル会話モデル

さらにFine-Turningも可能だが、それなりの計算資源が必要

28

28

BERT

Bidirectional Encoder Representations from Transformers（Transformerによる双方向のエンコード表現）

2018年10月にGoogleが発表した自然言語処理モデル

非常に汎用性が高い

様々な自然言語のタスクにおいて従来手法よりも高い評価が出ている

自然言語における事前学習モデルという点で、画像認識におけるResNetやVGGNetのような存在

Masked Language ModelとNext Sentence Predictionによる事前学習

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
https://arxiv.org/abs/1810.04805

29

29

BERTとGPTの違い

双方向の文脈理解

BERT

- 与えられたテキストの「左右両方の文脈」を同時に考慮して理解する。文中の各単語が周囲の単語とどのように関連しているかをより深く理解できる

事前学習と微調整

- BERTは大量のテキストデータで事前に学習（Pre-training）され、特定のタスクに合わせて微調整（Fine-tuning）される
- 事前学習では、マスク付き言語モデル（MLM）と次文予測（NSP）という二つのタスクを使用

用途

- 文書分類、質問応答、名前付きエンティティ認識など、多くのNLPタスクに適す

一方向の文脈理解:

GPT

- GPTは、「左から右」または「右から左」の一方向の文脈のみを考慮してテキストを生成する。これにより、文の次の部分を予測することに特化している

自己回帰言語モデル

- GPTは自己回帰言語モデルであり、与えられたテキストに基づいて次の単語を予測することに焦点を当てている

用途

- GPTは、テキスト生成、文書要約、対話システムなど、生成的なタスクに特に適している

30

30

Transformer

2017年にVaswaniら(Google)によってTransformerという手法が発表される

翻訳タスクにおいて、seq2seqよりも高速かつ高精度

seq2seq：RNNによるEncoder-Decorderモデル

BERT：Attention（注意機構）を用いたEncoder-Decoderモデル

RNNをはじめとする時系列処理は、時刻tと時刻t+1における逐次処理になる。しかし、逐次処理ゆえに並列処理ができない。そこで、BERTは時系列方向を集積しない

Attentionとは、文中のある単語の意味を理解する時に、文中の単語のどれに注目すれば良いかを表すスコアのこと。英語のthis, it, thatなどはその単語だけでは翻訳できない。これらの語を含む文章中の、どの単語にどれだけ注目すべきかというスコアを表す

BERTは双方向Transformerを使用している

A. Vaswani, N. Shazeer, et.al ,Attention Is All You Need, https://arxiv.org/abs/1706.03762

31

31

BERTとGPTのTransformerの使い方の違い

エンコーダーのみを使用: BERT

- BERTはTransformerのエンコーダー部分のみを使用
- B入力テキスト全体を一度に見て、各単語が文脈の中でどのように相互作用するかを理解する。これは「双方向」の文脈理解と呼ばれ、文の全体的な意味を捉えるのに有効

マスク付き言語モデル:

- トレーニング中にランダムに単語を「マスク」し、その隠された単語を予測することにより学習する。これにより、BERTは言語の深い理解を学ぶことが可能

デコーダーのみを使用: GPT

- GPTはTransformerのデコーダー部分のみを利用する。GPTは左から右へとテキストを生成する「一方向」のアプローチを採用しており、生成する各単語はそれまでに生成された単語に基づいている

自己回帰言語モデル:

- GPTは自己回帰モデルを使用し、与えられた単語シーケンスに基づいて次の単語を予測する。これにより、GPTは効果的なテキスト生成タスクに適している

32

32

やってみようrinna

ギリギリGoogle Colabで動くモデルを使ってrinnaを体験
昔のモデルに比べるとかなり良くなりました

rinna4B-instruction.ipynb: 4Bモデルによる基本対話
rinna4B-image.ipynb: ※かなり時間がかかります

33

33

やってみようOpenAI

OpenAIのAPIが必要
課金確認はここ https://platform.openai.com/usage

OpenAI-Basic.ipynb: APIの基本的な使い方
OpenAI-PDFChat.ipynb: OpenAIの知識にPDFの情報を追加させる
OpenAI-Image.ipynb: OpenAIの画像解析・生成APIを使用

最初はChatGPTにコードを書いて貰ったが、そのままのコードではダメ
（PDFをまるまる与えるとデータが大きすぎるので、仕様に沿う形で分割しないといけない）
結局PDFを適当に意味ブロックで分割してそれで食わせるようにした

34

34

Japanese StableLM

紹介

画像生成AIでおなじみ(?)のStability AIが開発
現在はJapanese StableLM Instruct Alpha 7B v2がリリース済み。商用利用可
https://ja.stability.ai/blog/japanese-stablelm-instruct-alpha-7b-v2

モデルは2つ
Base：汎用言語モデル。テキストの生成や理解などの一般的なタスク
Instruct：指示応答言語モデル

※そこそこ良さそうなので使ってみたかったのですが、アクセストークンやgitへのリクエスト権限が必要で間に合わず

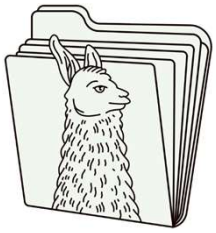
35

35

超お手軽ローカルLLM その1

たった単一4Gのファイルで Win/Mac/Linux にフル対応
llava-v1.5-7b-q4-server.llamafile をダウンロードして、実行権を付与するか拡張子を.exeに変更するのみ
https://gigazine.net/news/20231210-llamafile/
https://github.com/Mozilla-Ocho/llamafile

日本語も通るが、実行はWindowsではCUDAなしではかなり遅い
Macだとかなり速いが、Apple Siliconに最適化されている？
ライセンスはApache 2.0/MIT License



36

36

超お手軽ローカルLLM その2

Text generation web UI
https://github.com/oobabooga/text-generation-webui
この手のツールでは老舗(?)だが、環境整備があれこれ必要

LM Studio https://lmstudio.ai/
環境整備不要＆GUIでLLMをインストール＆使用できる夢のようなツール
日本語LLMも豊富に使用できる。例えば東大松尾研が開発したELYZAも利用可能
対話ならELYZA-japanese-Llama-2-7b-instruct

37

37

今日のまとめ

LLMはまさに群雄割拠時代。すぐに新しいモデルが登場するので、常に目を光らせておかないといけない
LLMの多くは商用利用も可能だが、動作環境自体がシビア
Google Colab無料版でも動くが、可能ならば有料Proがよい
本格的な検討・開発はクラウドで
OpenAIに課金すれば、APIであれこれできる。触っておいた方がよい

38

38