E-007

異なるアーキテクチャを持つ自然言語処理モデルの潜在空間の可視化と比較

Visualization and Comparative Analysis of Latent Spaces in Natural Language Processing Models with Different Architectures

泉 諒音[1)]　神野 健哉[1)]
Masato Izumi　Kenya Jin'no

概要

本稿では，異なるアーキテクチャを持つ自然言語処理モデルの潜在空間を視覚化し，それらの違いを詳細に比較，検討を行った．分析の結果，これらのモデルは潜在空間に類似性が見られるものの，微細な違いも観察された．Sentence-BERT はエンコーダーモデルであり，GPT-2 は生成モデルであるため，これらの構造的違いが結果に影響を与えていると考えている．このアプローチは，自然言語処理モデルの固有の特性や強みを明らかにするものであり，モデル選択や進化の貢献を目指している．

1　まえがき

自然言語処理（NLP）技術は，近年，BERT や GPT といったモデルの急速な進化により大きく変貌している．これらのモデルは，大規模なデータセットを使用してトレーニングされ，そのアーキテクチャは広く公開されている．しかし，これらのモデルがどのようにして具体的な出力結果を導出しているかについてはしばしば不明であり，多くがブラックボックスとしての性質を持っている．この問題に対処するため，本研究ではモデルが生成する潜在変数空間の性質に焦点を当て，良い出力結果がその適切な構成によるものではないかという仮説を立てている [1].

自然言語処理モデルが生成する潜在変数は実数値で表され，これらを直接評価することは困難である．したがって，潜在変数の視覚的な解釈を試みることで，モデルの決定過程をより深く理解することを目指している．本研究では，異なるアーキテクチャを持つモデル，具体的にはエンコーダーモデルの Sentence-BERT[2] と生成モデルの GPT-2[3] を用いて，これらの潜在変数から調査を行う．

2　Sentence-BERT[2]

Sentence-BERT は，事前学習済みの BERT モデルを基に，より意味的な正確さを追求し，似た文章が似た文ベクトルになるようにファインチューニングを施した自然言語処理モデルである．

元の BERT は，Transformer のエンコーダ部分を使用し，マスクされた言語の予測（Masked Language Model, MLM）と次の文予測（Next Sentence Prediction, NSP）の 2 つの事前学習タスクにより，双方向の Transformer を構成している．

日本語モデルは，東北大学の乾・鈴木研究室が公開している大規模日本語コーパスで事前学習された BERT[4] を基に，Hugging Face[5] によるファインチューニングが施されている．事前学習には，2020 年 8 月 31 日時点の日本語版ウィキペディアから抽出された約 3000 万文が使用されている．英語版のモデル [6] では，Microsoft が開発した事前学習済みの BERT に，10 億件以上のデータセットを用いてファインチューニングが行われている．

3　GPT-2[3]

GPT-2 は OpenAI によって開発された自然言語処理モデルであり，卓越した文章生成能力を持つ．このモデルは入力文に基づいて次の単語を予測することで事前学習が行われる．具体的には，入力された文のトークンを一つずらし，続く単語を予測する単方向の Transformer アーキテクチャを使用する．

英語版の GPT-2[7] のトレーニングには，OpenAI が独自に作成した WebText データセットが使用される．このデータセットは，約 4,500 万のリンクから抽出された 800 万件以上の文書を含み，多岐にわたるトピックをカバーする．一方，日本語版のモデル [8] は，日本語 CC-100 と日本語版ウィキペディアを使用して学習される．これにより，日本語特有の表現や語彙に対する理解を深めることが可能となる．

4　画像生成モデル

本研究では，画像とそれを説明する文章のペアを複数用意し，文章から画像を生成する過程を学習する．本稿で用いるモデルを図 1 に示す．

画像を説明する文章を自然言語処理モデルで潜在変数にし，この潜在変数を入力として使用する．その後，逆畳み込み演算を用いる畳み込みニューラルネットワーク（CNN）を通じて画像を合成する．

1）東京都市大学大学院 総合理工学研究科 情報専攻 Informatics, Graduate School of Integrative Science and Engineering, Tokyo City University

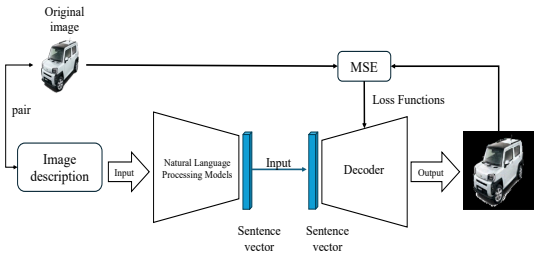

図 1　model

生成された文ベクトルを入力として，生成画像と教師データの画像との間の平均二乗誤差を損失関数として用い，対応する画像が出力されるように学習を行う．

さらに，本研究ではエンコーダー部分を異なる自然言語処理モデルに入れ替え，同一のデコーダーおよび学習データを用いて画像の生成を行う．エンコーダーとして様々な自然言語処理モデルを採用し，それぞれのモデルが生成する潜在変数の特性と効果を検証する．

5　データセット

本研究で使用される画像生成モデルは，車の画像のみに限定している．車はその形状が特徴的であり，一つの車種に複数のカラーバリエーションが存在することが一般的であるため，潜在変数空間での色情報と形情報の解析に特に適していると考えられる．これにより，車の画像を研究の対象として選定した．「色 + の + 車の形」というフォーマットの文章と，それに対応する画像のペアを使用している．画像は 128×128 ピクセルの解像度で，車の背景は透明である．合計で 30 色のカラーバリエーションを用いた．総計 600 枚の画像がこの研究で使用されている．

6　結果

結果を図 2 に示す．

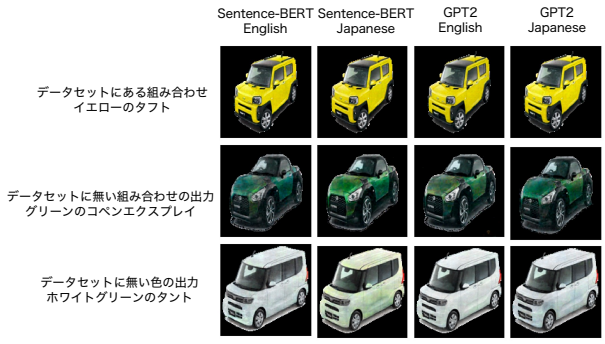

図 2　Result

本研究では，データセットに存在する組み合わせの画像が全モデルを通じてクリアに出力された．これは各モデルが学習目標を効果的に達成していることを示している．ただし，データセットに存在しない色と車の組み合わせ，特にコペンエクスプレイのグリーン色のケースでは，モデルごとの出力の違いが顕著に表れた．

データセット外の色の出力を評価した結果，日本語で学習された Sentence-BERT だけが優れた結果を示した．これに対し，英語モデルでは日本語の入力が適切に処理されなかったため，望ましい結果が得られなかった．さらに，日本語で学習された GPT-2 は単方向の Transformer 構造を採用しているため，入力された最初の単語の影響が出力に強く反映される．この動作特性により，ホワイト色が強調される傾向が見られたと考えられる．

7　まとめ

実験結果からは，各モデルが基本的な画像生成タスクにおいては同様に良好な結果を示したことが観察された．しかし，色の複雑さが増すにつれて，モデル間での出力差異が顕著になった．この差異は，モデルの構造的違いやそれぞれのデータセットの特性に起因していると考えられる．具体的には，モデルが目指す目的と訓練に使用されるデータの違いが，画像生成の性能に直接的な影響を与えている．

今後の研究で，潜在変数空間のさらなる調査と，調査範囲の拡大を行っていく．

謝辞

本研究の一部は JSPS 科研費 23K11266，23H03387，24K15115，東北大学電気通信研究所共同プロジェクト研究，東京都市大学重点推進研究未来知能ユニットの助成によるものです．

ダイハツ工業株式会社様には多くのデータを提供頂きました．厚く御礼申し上げます．

参考文献

[1] Masato Izumi, Kenya Jin'no, "Investigation of the structure of the latent variable space in Sentence-BERT sentence vectors using an image generation model", NOLTA, IEICE, Vol. 15. E14-N, No.2, pp. 376-388, Apr. 2024.

[2] N. Reimers, I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", Proc. EMNLP 2019, pp. 3982–3992, 2019.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, "Language Models are Unsupervised Multitask Learners", OpenAI blog 1.8 (2019): 9.

[4] acl-tohoku/bert-japanese: https://github.com/cl-tohoku/bert-japanese,

[5] sonoisa / sentence-bert-base-ja-mean-tokens-v2 https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2, 2021.

[6] sentence-transformers/all-mpnet-base-v2 https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[7] openai/gpt-2 https://github.com/openai/gpt-2, 2019.

[8] rinna/japanese-gpt2-small https://huggingface.co/rinna/japanese-gpt2-small, 2021.