CE-003

Sentence-BERTの文ベクトルによる画像生成
Image Generation by Latent Variables in Sentence-BERT

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

概要

我々はこれまでに Sentence-BERT が生成する文章の分散表現である文ベクトルが文章の意味をどの程度捉えているかを　k-means や UMAP などを用いて検証し，生成される文ベクトルが文章の意味を極めて高くとらえていることを確認した．本稿では Sentence-BERT で生成された文ベクトルから画像生成を行い，文章の意味に合わせた画像生成ができるかどうかについて検討を行う．

1　まえがき

自然言語処理において 2018 年に Google が発表した BERT[1] は文章の文脈を捉えることができるようになったモデルとして注目を集めた．非常に大規模なデータで事前学習された BERT を元モデルとして Siamese Network[2] によって入力された文章から文脈の意味をとらえた非常に高精度な文ベクトルを生成できる BERT を改良したモデルとして 2019 年に発表されたモデルが Sentence-BERT[3] である．我々はこれまでに Sentence-BERT が生成する文章の分散表現である文ベクトルが文章の意味をどの程度捉えているかを　k-means や UMAP などを使い検証をしてきた [4]．その結果，Sentence-BERT は単語が異なる場合や本来の意味を持たない隠語なども 似ている文章と判別できていることから文の意味を理解した文ベクトルが生成できることを確認した [4]．このような高品質の文ベクトルを用いることで文章分類，文章生成，文章校正などに応用することが可能である．

本研究では生成された文ベクトルの性質そのものに注目する．BERT では入力文書のトークン単位で文ベクトルが生成されるのに対し，Sentence-BERT では文章単位で文ベクトルが生成される．このことから Sentence-BERT で生成される文ベクトルは文章の意味を包含しているとみなすことができる．そこでこの生成された文ベクトルを潜在変数ベクトルとして画像を生成するシステムを考える．与えられた学習用データの文章とその文章が表現する画像との関係を文ベクトルを介在するように学習を行う．そして文ベクトルに含まれる文章の意味と画像に含まれる要素との関係を学習することができるのであれば，未学習データからも文章の意味を理解した画像を生成することが可能であるかを検証する．このような文章の意味を理解し，その文意から画像を生成する試みは最近，OpenAI が "DALE·E 2"[5] を，Google が "Imagen"[6] を公表している．これらでは文章から高精細な画像を生成するために様々な工夫がされているが，本稿で提案する方法はファインチューニング無しの Sentence-BERT が生成する文ベクトルから直接画像を生成し，文ベクトルに包含される意味を画像として表せるかどうかを検討している点が異なる．

1）　Department of Intelligent Systems, Faculty of Knowledge Engineering Tokyo City University

2　BERT[1]

自然言語処理では文章の単語を分散表現ベクトルに置き換えることが基本である．文章は単語データの並びである．Google が提案する BERT (Bidirectional Encoder Representations from Transformers)[1] は入力された単語データの並びである文章から別の文章を予測する．元の文章から別の文章を予測する際に BERT では単に元の文章から文章を予測するだけではなく，予測された文章から元の文章を予測するという双方向の学習を Attention 機構を用いて行っている．Attention 機構は離れた位置のトークンに対しても注意を向けることができるため，この Attention 機構を用いた双方向の学習が可能な Transformer モデルによって各単語の意味を包含した文章の文脈情報を学習することが可能となった．この双方向の学習によって BERT は入力された文章をその文脈の意味を包含した文ベクトルに符号化することを実現し，高精度な文ベクトルが生成できる．

入力する文章はまず形態素解析によってトークン化し，トークン列に変換する．トークン列をベクトル化することで BERT の入力にする．BERT はこの入力されたトークーンベクトルを文ベクトルに変換する．BERT の学習には非常に大規模な学習データセットが必要となる．このため事前学習モデルが公開されている Hugging Face が公開している Transformer モデルを使用した事前学習済 BERT モデルでは入力された文章を 768 次元の文ベクトルに変換する．適切な文ベクトルに変換できるようにするためには大規模な文書コーパスで学習をする必要がある．大規模コーパスから分散表現の学習はラベル無し学習である．そこで大規模な文書コーパスから汎用的な言語パターンは図 1 に示すような 2 種類の事前学習によって実現している．

１つ目の事前学習は Masked Language Model（MLM）である．MLM は入力トークンの 15%を Mask トークン

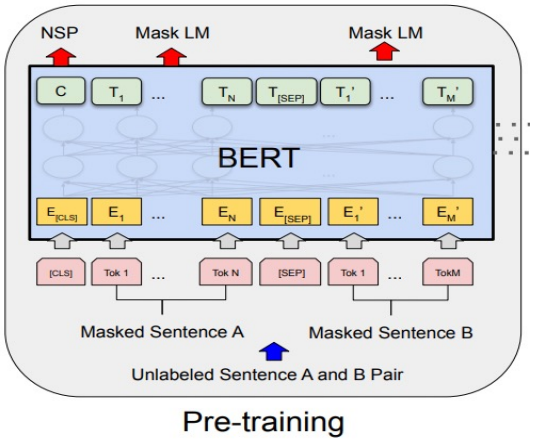

図 1　BERT pre-training[1]

でマスクし，元のトークンを当てるタスクである．これはある単語を周辺の単語から予測させるタスクであり，これによって文章中の単語間の関係を学習させることができる．

2 つ目の事前学習は Next Sentence Prediction（NSP）である．MLM では単語同士の関係性のみに着目するため文章間の関係を考慮する問題に対して対処できない．そこで文章間の関連性を学習できるようにするため，関係性のある 2 つの文章と関係性の無い 2 つの文章を与え，それらの関係性を識別できるように学習させる．

これらの事前学習によって BERT は文章の双方向理解を実現している．

3　Sentence-BERT[3]

Sentence-BERT[3] は事前学習済みの BERT に対しファインチューニングを行いより意味的に正しく，似た文章は似た文ベクトルになるようにした自然言語モデルである．具体的にはある文章とそれに類似する文章のペアを学習データとして選び，これら似た文章から生成される文章ベクトルが似たベクトルになるように BERT をファインチューニングしている．これは BERT によって生成された文ベクトルを図 2 に示すように Triplet Network[7] と Siamese Network[2] を用いてファインチューニングすることで実現している．Triplet Network によって 2 つの文ベクトルの差分を連結し文章のラベルに応じた学習を行う．また Siamese Network によって 2 つの文ベクトルの cos 類似度を求め，類似した文章の文ベクトルが類似するように学習を行う．

3.1　Triplet Network[7]

Triplet Network を用いた学習では「文 A」,「肯定文 p」「否定文 n」の 3 つの文章を入力しベクトル化する．それぞれのベクトルに対し「文 A」,「肯定文 p」の距離を「文 A」,「否定文 n」の距離より近づけるように学習する．また,「文 A」,「否定文 n」の距離を「文 A」,「肯定文 p」の距離より遠ざけるように学習させている．3 つの特徴量ベクトルに対しそれぞれの位置関係をよりわかりやすく，大袈裟すぎないように特徴量ベクトルを調整することで文章間の関係性をより明確にさせる．

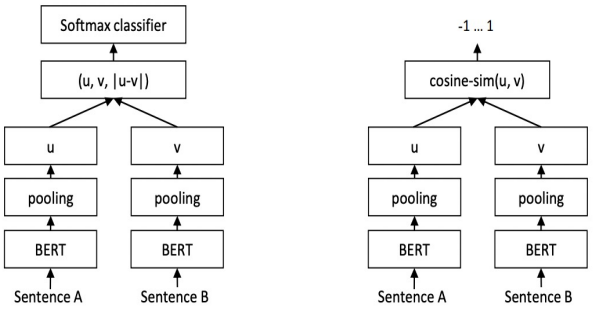

図 2　Sentence-BERT[3]

3.2　Siamese Network[2]

Siamese Network では 2 つの文章に対し学習を行う．「文章 A」と「文章 B」に対し 2 つの文章が同一ラベルであれば 2 つの文章を近づけるように学習させる．2 つの文章のラベルが異なる場合遠ざけるように学習させる．よって，似た文章は似たベクトルに，似ていない文章は異なるベクトルになる．

4　テキストから画像を生成するモデル

本研究では画像とその画像を説明する文章をペアを複数用意し，画像を説明する文章から画像が生成されるように学習を行う．具体的には画像を説明する文章を Sentence-BERT で文ベクトル化する．次にこの文ベクトルを入力として，逆畳み込み演算を用いて画像を合成する CNN を用意する．文ベクトルの意味の通りの画像を出力することができればこれは「復号」と呼べるためこの CNN を以降 Decoder と呼ぶ．画像を説明した文章から生成された文ベクトルを入力にして，これに対応した画像が出力されるように生成された画像と教師データの画像との平均二乗誤差を損失関数として学習を行う．提案するモデルを図 3 に示す．

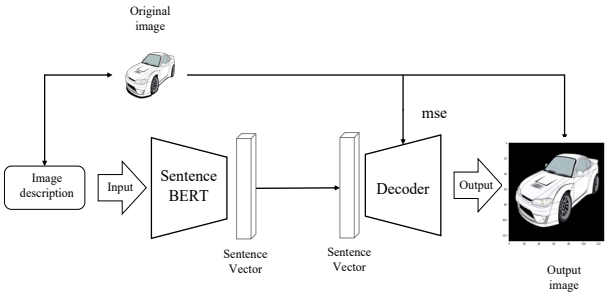

図 3　model

画像を説明する文章は形態素解析ツール fugashi を用いて「分かち書き」を行った．これをトークン化して Sentence-BERT に入力する Sentence-BERT は東北大学乾・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済みの BERT[8] を基に，Hugging Face がファインチューニングしたもの [9] を用いた．事前学習は 2020 年 8 月 31 日現在の日本語版ウィキペディアで約 3000 万文をコーパスとして学習されている．上記の Sentence-BERT で画像を説明する文章を文ベクトル変換する．得られた文ベクトルを入力として対応した画像が生成できる Decoder を出力画像の平均二乗誤差 (MSE) を損失関数として学習を行った．用いた Decoder を図 4 に示す．

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
decoder_input (InputLayer)   [(None, 768)]             0

batch_normalization_5 (Batc  (None, 768)              3072
hNormalization)

dense_1 (Dense)              (None, 768)               590592

batch_normalization_6 (Batc  (None, 768)              3072
hNormalization)

reshape_1 (Reshape)          (None, 2, 2, 192)         0

batch_normalization_7 (Batc  (None, 2, 2, 192)        768
hNormalization)

conv2d_transpose_3 (Conv2DT  (None, 16, 16, 256)      3981568
ranspose)

batch_normalization_8 (Batc  (None, 16, 16, 256)      1024
hNormalization)

conv2d_transpose_4 (Conv2DT  (None, 64, 64, 256)      5308672
ranspose)

batch_normalization_9 (Batc  (None, 64, 64, 256)      1024
hNormalization)

conv2d_transpose_5 (Conv2DT  (None, 128, 128, 3)      6915
ranspose)

=================================================================
Total params: 9,896,707
Trainable params: 9,892,227
Non-trainable params: 4,480
_________________________________________________________________
```

図 4　Decoder

Sentence-BERT で得られた 768 次元の文ベクトルから 128×128 のカラー画像を生成するため，我々は図 4 に示すような Decoder を設計した．この Decoder は 3 種の逆畳み込み層と Batch Normalization 層で構成される．入力の 768 次元のベクトルを 192 チャネルの 2×2 のデータに変換する．これを最初の逆畳み込み層で 256 チャンネルの 16×16 に拡大，2 回目の逆畳み込み層で 256 チャンネルの 64×64 に拡大，そして 3 回目の畳み込み層で 3 チャンネルの 128×128 に変換している．それぞれの逆畳み込み層の入力は Batch Normalization 層を用いて正規化を行なっている．学習の際の MSE を図 5 に示す．

5　実験

データセットには画素数 128 × 128 の車の画像を使用した．80 車種の背景透過画像を用意した．車の画像を

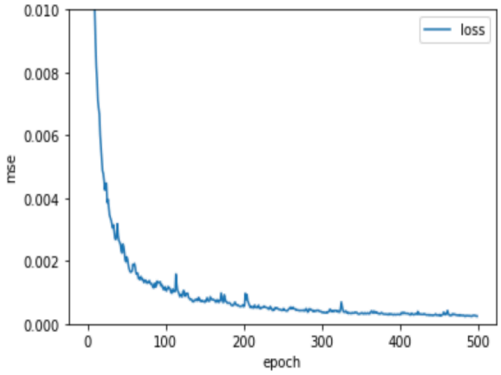

図 5　MSE

説明する文章は「色＋の＋車名」とした．車色は車種によって 5 種類から 8 種類である．車の色はブラック，レッド，グリーン，ブルーなどを含む 30 色のバリエーションを用意し，計 500 枚の画像を用いて実験を行う．色のバリエーションを図 6 に示す．

ブラック		オレンジ		リーフグリーン	
グレー		レッド		グリーン	
シルバー		ダークレッド		スカイブルー	
ルナシルバー		レッドブラウン		ライトブルー	
ホワイト		バイオレット		ターコイズブルー	
イエロー		ベージュ		ブルー	
ゴールデンイエロー		カーキ		ダークブルー	
ピンクベリー		ブラウン		レーザーブルー	
ピンク		アイスグリーン		インディゴブルー	
シャイニングオレンジ		ライムグリーン		ネイビーブルー	

図 6　色のバリエーション

5.1　データセットに含まれる画像の生成

本稿の結果を表す指標に「オリジナルカー」という車種を用いる．「オリジナルカー」の車色はホワイト，レッド，グリーン，ブルー，シルバー，グレー，ピンクの 7 色である．学習後にデータセットに含まれる画像を説明する文章を入力すると，文の意味に沿った画像が生成できるか実験を行った．結果を図 7 に示す．

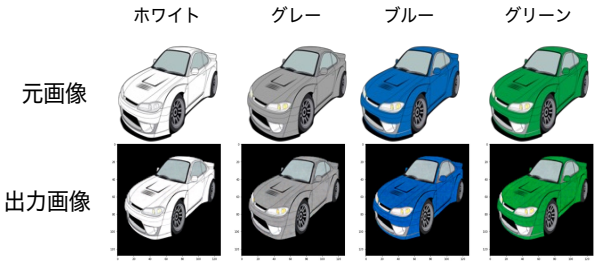

図 7　データセットに含まれる画像

データセットに含まれる画像を説明する文章に対し元画像と変わりない画像の出力を確認した．

5.2 データセット含まれない画像の生成

「オリジナルカー」に存在していない色の着色の実験を行う．入力文章は「存在していない色＋の＋オリジナルカー」とした．結果を図 8 に示す．

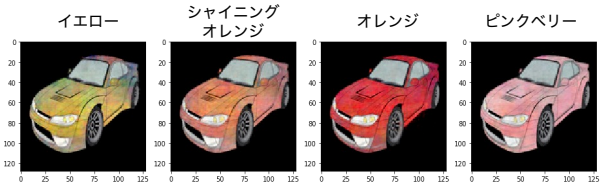

図 8 データセットに含まれない画像

元画像に無い色と車名をつなげた入力文章でも車種判別ができる精度で画像が生成できた．着色結果は，イエローに対しては色にノイズが入っているがピンクベリーは良い画像が生成できている．このことから元画像に近い色が存在するとうまく生成できることを確認した．またオレンジが赤に近い色になっていることから元画像の影響もあると考えられる．以上から元画像に含まれていない色の文章に対しては元画像に含まれているその他の車種の車色を参照し生成することを確認した．

5.3 色の濃度の比較

車種にない色と車名の組み合わせの文章に対しても画像が生成できることからその他の色名でどのような画像が生成されるか検討に入力文章に応じて生成された色の濃度がどのように変化するか実験を行った．結果を図 9 に示す．

図 9 色の濃度の比較

結果，文章を変化させることで色の濃度が変化することを確認した．薄い色の生成は良くできており，細かい濃度変化が観察できる．しかし濃い色では変化は見られるものの薄い色ほどの変化は見られなかった．これは学習時の「オリジナルカー」にホワイトは存在するが，ブラックが存在しないことに起因して薄い色の生成はうまくでき，濃い色は参照する色が少ないためうまく生成できなかったと考える．

6 まとめ

Sentence-BERT が生成した文ベクトルから文の意味を汲んだ画像が生成できるかの実験を行った．ある車種に存在するが自車に存在しない車色の車画像を生成できることを確認した．

今回の実験では車色情報だけを取り扱った．ある車種に存在するが自車に存在しない車色の車画像を生成できることを確認した．学習データ中に少ない色の着色を行なった際，データセットに多く含まれている色が強く反映していることも確認した．そのためデータセット内で使用されている色の偏りなどが学習結果に強く反映することを確認した．今後はデータセットの情報量の増加，平均化と入力の文章の文法の違いや単語，言葉遣いによる出力結果の違いについて検討する．

謝辞

本研究の一部は JSPS 科研費 JP20K11978 の助成，および東北大学電気通信研 究所共同プロジェクト研究によるものです．

参考文献

[1] J. Devlin, M. W. Chang, K. Lee, K.Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proc. NAACL-HLT2019.

[2] D. Chicco, ”Siamese neural networks: an overview”, Artificial Neural Networks, Methods in Molecular Biology, vol. 2190 (3rd ed.), Springer Protocols, Humana Press, pp. 73–94, doi:10.1007/978-1-0716-0826-5_3, ISBN 978-1-0716-0826-5, PMID 32804361.

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

[4] 泉 諒音, 神野 健哉, ”Sentence-BERT の文ベクトルの UMAP による特徴解析”, 電子情報通信学会 NOLTA ソサイエティ大会, NLS-17, 2022.

[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, ”Hierarchical Text-Conditional Image Generation with CLIP Latents,” arXiv:2204.06125, 2022.

[6] C. Saharia, W. Chan, S. Saxena, and et. al., ”Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” arXiv:2205.11487, 2022.

[7] E. Hoffer, N. Ailon. ”DEEP METRIC LEARNING USING TRIPLET NETWORK.” In ICLR workshop, 2015. [2] G. Koch, R. Zemel, R. Salakhutdinov, ”Siamese Neural Networks for One-shot Image Recognition”,ICML deep learning workshop, vol. 2. 2015.

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

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