The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

条件付き変分オートエンコーダと単語分散表現による絵文字の生成

Gerenaring Emoji with Conditional Variational Autoencoders and Word Embedding

山口 篤季 *1　藤田 桂英 *2
Atsuki Yamaguchi　Katsuhide Fujita

*1東京農工大学工学部
Faculty of Engineering, Tokyo University of Agriculture and Technology

*2東京農工大学大学院工学研究院
Institute of Engineering, Tokyo University of Agriculture and Technology

Emoji are among the most widely used communication tools worldwide. Because the number of emoji increases every year and there are 82 face emoji, it might be difficult for users to select an appropriate emoji immediately. Moreover, it is troublesome to continue designing new emoji. Therefore, the aim of the present study is to generate an emoji based on input text automatically to facilitate easier communication and eliminate the process of designing new emoji. The proposed model employs conditional variational autoencoders (CVAE), quasi-recurrent neural networks (QRNN) as the text encoder, and the pre-trained word vector GloVe to the embedding layer connected to the text encoder. In the experiments described herein, it will be observed that the proposed method can generate an emoji that corresponds to an input caption, and output image quality is improved using GloVe.

1. はじめに

絵文字は 1999 年に日本で初めて公開され，携帯電話の普及とともに瞬く間に人気となった．2010 年代になると，世界中でスマートフォンが普及し，絵文字は Unicode によって標準化され，広く使われるようになった．絵文字には，顔型の絵文字（Smiley）から，食べ物，車両，建物に至るまで様々な種類が存在し，絵文字は我々の日常のコミュニケーション活動において重要な役割を果たしている．

絵文字が世界中で広く使われるようになるにつれて，絵文字を対象とした研究が盛んに行われるようになってきた．[1] は，Twitter 上における絵文字とテキスト間の関係性を調査し，どのような種類の絵文字がテキストベースのツイートに最も共起するかを予測した．この研究は，絵文字が世界中で広く扱われているにもかかわらず，自然言語処理の研究題材としてほとんど注目が集められてこなかったことに着目し，絵文字のもつ潜在的な意味を解析することを目的としていた．また，[3] は全ての Unicode の絵文字に関する単語分散表現を作成し，公開した．実験結果では，自然言語を対象とした単語分散表現と同様に，絵文字においても絵文字の意味の加減算が可能なことが示され，テキスト解析において絵文字を考慮することの重要性が示唆された．

一方，絵文字の数は年々増加しており，顔型の絵文字に限っても 82 種類存在することから，ユーザが適当な絵文字を即座に選択することは困難である．さらに，新たな絵文字を適宜デザインし，追加する作業は大きな労力を必要とする．敵対生成ネットワークを始めとした機械学習の生成モデル分野が発展してきていることから，入力のキャプションに対して対応する絵文字を自動生成することは，実現可能なタスクである．

本研究に応用可能な関連研究として “Text-to-image task” が有名である．これは，入力のテキストに基づいて画像を生成するタスクである．このタスクについては数多くの先行研究が存在する一方で，text-to-image synthesis task を応用した絵文字生成に関しての先行研究はほとんど存在せず，次の二つの先行研究のみが挙げられる．[8] と [6] は，深層畳み込み敵対生成ネットワーク（DCGAN）[7] を絵文字生成に活用し，入力のキャプションから絵文字を生成することに成功した．しかし，これらの先行研究は実験に使用しているデータセットに関する説明や内容が不十分であり，実装に関する記述が省かれているため，再現性に欠ける研究となっている．

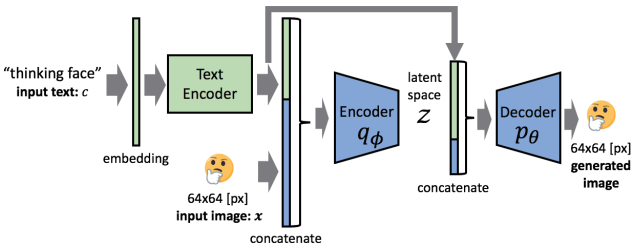


図 1: CVAE ベースの提案手法

そこで，本研究では絵文字を入力文から自動生成するモデルを提案する．これにより，ユーザのコミュニケーション活動の促進が期待されるほか，新たに絵文字をデザインする手間を省くことができる．提案手法は，CVAE [9] と事前学習済み単語ベクトル GloVe [5] を活用している．CVAE を活用することで，従来手法の GAN ベースのモデルよりも高速でかつ安定した学習を実現できる．さらに，事前学習済みの単語ベクトル GloVe を活用することで出力画像の精度に関して，定量的評価により示す．

以下に，本論文の構成を示す．まず，絵文字を入力文から自動生成するモデルを提案する．次に，ベースラインと提案手法の定性的評価の結果を示す．最後に，本論文のまとめと今後の課題を示す．

連絡先: 山口 篤季，東京農工大学工学部情報工学科，東京都小金井市中町 2-24-16，y-atsuki@st.go.tuat.ac.jp

- loudly crying face
- a sad face with tears streaming down both cheeks
- this face is distraught and inconsolable

図 2: データセットの画像とキャプションの対応の一例

2. 提案手法

図 1 は提案するモデルを示している．このモデルは，CVAE をベースモデルとして活用し，GloVe を埋め込み層に用い，さらに QRNN [2] をテキストエンコーダとして利用している．モデルは式 (1) で定義される変分下限 $\mathcal{L}$ を最大化することで学習できる．

$$\mathcal{L}(\boldsymbol{x}, c; \theta, \phi) = \mathbb{E}_{q_\phi(z|\boldsymbol{x},c)}[\log p_\theta(\boldsymbol{x}|c, z)] \\ -D_{KL}(q_\phi(z|\boldsymbol{x}, c)||p_\theta(z)) \quad (1)$$

単語分散表現

単語分散表現は自然言語処理の研究において，単語の特徴量を抽出するのに広く活用されている．提案手法は 300 次元の事前学習済み単語ベクトル GloVe [5] を活用している．事前学習済みの単語ベクトルを活用することで，モデルは入力文の意味をより効率的に捉えることができるようになり，出力画像の質が向上することが期待される．

テキストエンコーダ

テキストエンコーダには QRNN [2] を用いる．QRNN は畳み込みニューラルネットワーク (CNN) [4] をベースにしたモデルであり，再帰ニューラルネットワーク (RNN) の構造を CNN で再現したものである．RNN とは異なる特徴として，CNN をベースとしたモデルのため並列処理に特化していることが挙げられる．このため，LSTM ベースのネットワークと同等の性能を示しながらも，より高速な演算を実現している．

3. 評価実験

3.1 データセット

データセットは絵文字画像とキャプションデータで構成され，合計 260 個のキャプションデータが存在する．絵文字画像は EmojiOne から収集し，キャプションデータは，Unicode の定める CLDR Short Name と Emojipedia の対応する絵文字の説明文からなる．絵文字あたりのキャプション数は，平均値で 3.17，中央値で 3 となっている．また，一文あたりの単語数は平均値で 9.33，中央値で 8 である．図 2 に，絵文字画像データとキャプションデータの対応を示した例を示す．

3.2 実験設定

データセットの 10%をテストセットとした．提案手法は Adam にしたがって最適化され，学習率とバッチサイズはそれぞれ，0.001 と 26 に設定した．学習は 200 エポック実行した．

3.3 定量的評価

提案手法を CNN ベースの分類器 *1 による出力画像の分類精度と，Inception スコア [10] の 2 つの評価指標に基づいて定量的に評価した．各指標は 100 回ずつ計算を行った．表 1 は 3 モデル間の定量的比較結果を示している．なお，DCGAN による従来手法の実装が公開されていないため，本論文では [7] による DCGAN の実装を基に，条件付き DCGAN を実装して比較対象とした *2．

Model	Accuracy	IS	Runtime (s)
CVAE w/ GloVe	0.869	1.05	**20.2**
CVAE w/o GloVe	0.262	1.02	/
Cond-DCGAN	**0.885**	**1.32**	33.6
Dataset	/	1.38	/

表 1: 3 モデル間の定量的比較結果: “Accuracy” は CNN ベースの分類器の分類精度を示している．“IS” は Inception スコアの略である．太字は提案手法（CVAE w/ GloVe）と従来手法（Cond-DCGAN）間でマン・ホイットニーの U 検定により有意な差が認められたことを表す．($p < .05$)

提案手法における GloVe の有無が与える影響　表 1 より，二つの指標において，事前学習済みの単語ベクトルを用いたモデルは，事前学習済みの単語ベクトルを用いないモデルよりも優れていることが確認できる．CNN ベースの分類器による分類精度に着目すると，比較対象の二つのモデル間では 0.6 以上の有意な差がある．したがって，事前学習済みの単語ベクトルを用いたモデルの大半の出力画像は，入力のキャプションに従った画像を出力できている．一方，Inception スコアにおける二つのモデル間の差は 0.04 よりも小さく，データセットの値よりも 0.3 以上の差がある．これは，二つのモデルの出力画像が共に色彩面のノイズを含み，ぼやけていることが原因である．

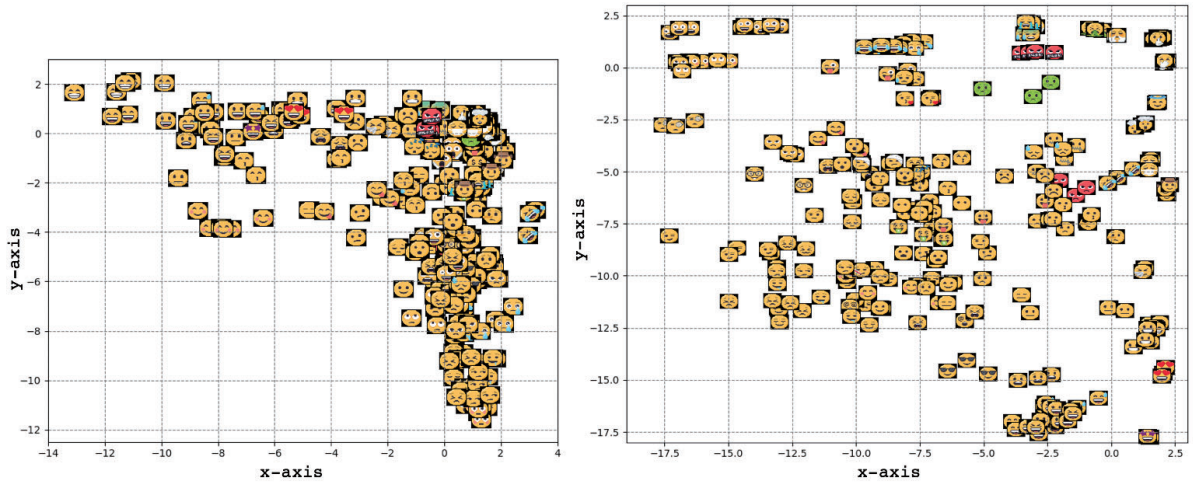
提案手法と従来手法の比較　CNN ベースの分類器による分類精度は，提案手法と従来手法で有意ではあるが 2%程度の差しか見られない．一方，従来手法の Inception スコアは，データセットのスコアとかなり近い数値を示しており，提案手法と従来手法の Inception スコアには 0.27 の大きな差が生じている．対照的に，エポック単位での実行時間に着目すると，提案手法は従来手法よりも 1.67 倍高速に動作している．また，従来手法は生成精度を最大化するのに 1400〜1500 エポック程度を要するが，提案手法は 200 エポックで十分である．以上から，提案手法は従来手法よりも画質の観点では劣るが，入力のキャプションに従った画像をより高速に生成できることが示された．

3.4 潜在表現の可視化

テキストエンコーダ (QRNN) と事前学習済み単語ベクトル GloVe を用いて条件付けられた，エンコーダネットワークがどのように入力画像を潜在変数に落とし込んでいるのかを，潜在表現を可視化することにより検証した．図 3 は，GloVe を用いて学習を行ったエンコーダモデルと，GloVe を用いないで学習をしたエンコーダモデルの出力の散布図を示している．図 3 において，同種の絵文字は，事前学習済み単語ベクトルの有無に関わらず，互いに近い位置にプロットされている．したがって，埋め込み層とテキストエンコーダは，入力のキャプションの意味を捉えて，同種の絵文字に対して似たような数値を持つベクトルを生成できている．図 3(1) にプロットされている絵文字の意味に着目をすると，x の値が小さければ小さいほど，より “positive” な絵文字がプロットされている．対照的に，y の値が小さければ小さいほど，より “negative” な絵文字が現れている．さらに，“positive” や “negative” の指標では分類の難しい絵文字は，原点: $(x, y) = (0, 0)$ 付近にプロットさ

*1　CNN ベースの分類器を実装し，82 種類の絵文字を分類できるようにデータセットで学習させた．データセットのうち 10%をテストセットとした．分類器は 20 エポックの時点で，学習用セットにおいて，0.995，テストセットにおいて，1.00 の精度を記録した．

*2　実装は `https://github.com/gucci-j/emoji-gan` に公開した．



(1) With pre-trained word vector　　(2) Without pre-trained word vector

図 3: エンコーダネットワークの出力散布図: 全データセットを学習済みのエンコーダネットワークに入力し，その出力を散布図上にプロットした．なお，マーカーを対応する入力の絵文字に置き換えた．

れている．一方，図 3(2) においては，各絵文字はカテゴリ毎に分類されているものの，図 3(1) のような大域的な絵文字の関係性を見出すのは困難である．図 3(1) と (2) の散布図における絵文字の大域的な分布の違いは，事前学習済み単語ベクトルの影響によるものが大きい．以上の実験結果より，分布が異なる絵文字間の大域的な関係性を含むときには，出力画像のノイズが軽減されるものと示唆される．

4. おわりに

本研究では，入力のキャプションに対応する絵文字を生成できる CVAE ベースのモデルを提案した．また，事前学習済みの単語ベクトル GloVe を活用することで出力画像の画質が向上することが示された．

今後の展望としては，未知のキャプションデータに対して，適切な絵文字を合成できるようにモデルを改良することが挙げられる．これにより，モデルが豊富な種類の絵文字を生成できるようになり，平文でのコミュニケーションがより促進されることが期待される．

参考文献

[1] Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. Are emojis predictable? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 105–111. Association for Computational Linguistics, 2017.

[2] James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. Quasi-recurrent neural networks. In *arXiv*, Vol. abs/1611.01576. 2016.

[3] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pp. 48–54. Association for Computational Linguistics, 2016.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. 2012.

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[6] Marcel Puyat. Emotigan : Emoji art using generative adversarial networks. `http://cs229.stanford.edu/proj2017/final-reports/5244346.pdf`, 2017.

[7] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *arXiv*, Vol. abs/1511.06434. 2015.

[8] Dianna Radpour and Vivek Bheda. Conditional generative adversarial networks for emoji synthesis with word embedding manipulation. In *arXiv*, Vol. abs/1712.04421. 2017.

[9] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28*, pp. 3483–3491. 2015.

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.