E-027

LLM を用いた画像に対する面白キャプション生成
Funny Caption Generation for Image using LLMs

根來 勇太† 森田 和宏† 泓田 正雄†
Yuta Negoro Kazuhiro Morita Masao Fuketa

1. はじめに

電笑戦[1]など，これまで数多くの「AI に人を笑わせることができるのか」という取り組み・研究が行われてきた．美濃口らの先行研究[2]では bokete[3]のデータから，Convolutional Neural Network (CNN)と Long Short-Term Memory (LSTM) を連結した深層学習モデルを用いて画像に対するボケた文章の生成を行った．この研究では，STAIR Captions をデータセットとした場合よりは面白いボケを生成できているが，人間のボケより面白くない，出力された文章が破綻していることがあるなどの改善するべき問題点がいくつか存在する．また，ChatGPT[4]をはじめとした Large Language Model (LLM) は質疑応答などのタスクは得意だが，ジョークや謎かけを生み出すことは苦手である．そこで本研究では，LLM を bokete のデータでファインチューニングすることで先行研究の問題点の改善と LLM による面白い文章生成の実現を目指す．

2. 関連研究

AI を用いた画像に対する面白キャプション生成に関する研究と LLM について紹介する．

2.1 Neural Joking Machine (NJM)

美濃口らは bokete のデータから CNN と LSTM を連結した Image to Text 深層学習モデルを作成し，AI によるボケた文章を生成している．

ここで，bokete とは「写真に一言ボケる」Web サービスのことであり，数多くのユーザーが画像に対する面白いボケを投稿している．

評価実験では，121 人の被験者が各手法で生成されたキャプションについて 0 (面白くない)から 3 (面白い)の点数で評価を行っている．評価実験の結果，平均スコアは 0.59 となり，画像と説明文がセットになっている STAIR Captions を用いて作成された深層学習モデルの生成したキャプションよりは面白いという評価になったが，人が投稿したボケよりは面白くないという評価であった．また，出力された文章が破綻していることがあることや未知語を示す UNK トークンが出力に含まれているといった課題が存在している．

本研究では，AI による面白キャプション生成の精度を向上させるとともに上記で上げた問題点を改善することを目的とする．

2.2 Large Language Model (LLM)

LLM とはパラメータ数が非常に多い言語モデルのことであり，ChatGPT などが例に挙げられる．LLM はプロンプトと呼ばれる自然言語で与えた指示に従って，要約や対話などの多様なタスクを高精度で行うことが可能である．しかし，Hallucination などの課題・苦手なタスクも存在する．Jentzsch らの研究[5]によると ChatGPT がユーモアの認識に一定の理解はあるものの，ダジャレやユーモアを含んだ文章生成や認識には課題があるとされている．また，大量の計算資源を必要とすることも課題である．

本研究では，ユーモアを理解させるために LLM に bokete のデータを学習させる．さらに効率の良いファインチューニングやローカル LLM を用いることで計算資源の問題解消に取り組んでいる．

3. 提案手法

LLM を用いた面白キャプション生成の手法について述べる．

3.1 データセットの作成

データセットにはボケ缶データセットを使用した．ボケ缶データセットとは，bokete に収録されている画像とボケの集合のことであり，ボケについた星評価数の範囲によって缶の種類が分けられている．ここで，星評価とはボケてユーザーが一人当たり一つのボケに星 1-3 をつけたものである．今回は星評価数の範囲が 101 以上のデータから約 1 万 2 千個のボケと画像を学習用データに用いた．

次にデータセットの画像に対して説明文の付与を行った．これは LLM がテキスト入力しかできないためである．画像説明文の作成には MLLM(Multimodal Large Language Models)の一つである LLaVA[6]を使用した．MLLM とは，テキストだけでなく画像などの入力にも対応した LLM のことである．具体的には，データセットの画像を MLLM に入力し，プロンプトとして「この画像を説明してください」と指示を与えることで画像説明文を作成した．なお，入力する際の画像は 224×224pixel にリサイズ処理を行った．

最後に LLM に面白いキャプションを生成させるためのプロンプトをデータセットに付与した．プロンプトは「入力された画像説明文からユーモアあふれる文章を生成してください」とした．

作成したデータセットの例を表 1 に示す．

3.2 LLM のファインチューニング

3.1 節で作成したデータセットを用いて LLM をファインチューングする．本手法では，Japanese StableLM Alpha 7B[7]と Swallow 7B -instruct[8]の 2 種類のローカル LLM モデルを採用した．どちらも 70 億パラメータの LLM であり，StableLM は汎用言語モデル，Swallow は指示チューニングを施した言語モデルを使用した．また，両モデルとも 8bit に量子化して使用した．

ファインチューニングの手法としては Lora[9]を採用した．

† 徳島大学 Tokushima University

Lora の優れている点として，メモリ効率が良い，推論コストが増加しないことがあげられる．

3.3 面白キャプション生成

3.2 節で学習した 2 つのモデルを用いて面白キャプション生成を行う．

4. 評価実験

提案手法を評価するために，先行研究と同じく人手での比較実験を行った．

4.1 評価実験の設定

評価実験の詳細について示す．画像枚数は 10 枚であり，それぞれの画像に対して，2.1 節の従来手法で生成したキャプション，3 章で紹介した提案手法の 2 つのモデルで生成したキャプション，人間のボケの合計 4 つを被験者に比較評価してもらった．その際，画像は全ての手法で同一かつ学習時に使用していないものを採用した．被験者は 13 名であり，各手法によって生成されたキャプションについて 0(面白くない)から 3(面白い)の点数で評価を行う．なお，比較の際はどのキャプションがどの手法であるかはわからないようにした．

4.2 評価結果

表 2 にアンケートでの評価結果として，点数 0 から 3 の割合及び平均を示す．結果として，提案手法のほうが先行研究の平均スコアより高いと評価された．また，UNK トークンが出力される問題が解消されており，破綻した文章が生成されることが少なかったことから，先行研究の問題点を改善することができた．これは，LSTM のモデルに比べ LLM(7B)は大規模な事前学習を行っており語彙が豊富なことが理由の一つであると考えられる．

一方，人間の投稿に比べ，提案手法のキャプションは文章が長くなる傾向にあり，特に文章が長く「オチ」だけでなく「フリ」まで出力しているボケは評価が低い傾向にあった．

最後に，表 3 に図 1 に対して提案手法によって生成されたキャプションの一例を示す．

5. おわりに

本研究では，LLM を bokete のデータを用いてファインチューニングすることにより，面白キャプション生成を行う手法を提案した．

今後はプロンプトチューニング，MLLM のみを用いたファインチューニング，学習パラメータの調整などを行うことでさらなる精度向上を試みる．

謝辞

本研究を進めるにあたり，ボケ缶データセットを提供して頂きました株式会社オモロキ様に感謝いたします．

参考文献

[1] 電笑戦 : https://github.com/aws-samples/bokete-denshosen
[2] 美濃口宗尊, 吉田光太, 螺良和樹, 池谷拓夢, 片岡裕雄 , 中村明生, “Neural Joking Machine —面白キャプションの生成及び評価に関する基礎検討—“, 精密工学会誌, VoL.85, No.12, pp.1151-1156 (2019)
[3] 写真で一言ボケて(bokete), 株式会社オモロキ: https://bokete.jp/

図 1　使用した画像例

表 1　図 1 から作成したデータセットの例

項目	例
プロンプト	入力された画像説明文からユーモアあふれる文章を生成してください
画像の説明文	この画像は、オフィスでの仕事に集中している5人の男性のグループ写真です。彼らは、オフィスの机や…
ボケ(正解)	言葉は通じないけど楽しい職場です！

表 2　アンケート評価の結果

Score	先行研究	提案手法 (Stable LM)	提案手法 (Swallow)	人間の投稿
0	59.2	44.6	46.2	14.6
1	29.2	25.4	30.0	20.8
2	6.9	23.8	16.2	32.3
3	4.6	6.2	7.7	32.3
Avg	0.57	**0.92**	0.85	1.82

表 3　各手法で生成されたキャプション例(図 1)

先行研究
「あの娘なら、今日の<UNK>の面接、誰これ？」「いや、俺はそのくらいの<UNK>だよ」

提案手法(StableLM)
「いいか、この中にスパイがいる。」と言われたけど、どう見ても自分がそのスパイ。

提案手法(Swallow)
社長が「私の代わりに出頭してきてくれ」と意味不明な遺言を残して他界した

人間の投稿
言葉は通じないけど楽しい職場です！

[4] ChatGPT : https://openai.com/chatgpt/
[5] Sophie Jentzsch and Kristian Kersting. “ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models.” In the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pp.325– 340 (2023)
[6] Liu, Haotian and Li, Chunyuan and Wu, Qingyang and Lee, Yong Jae, “Visual Instruction Tuning” In NeurIPS (2023)
[7] Lee, Meng and Nakamura, Fujiki and Shing, Makoto and McCann, Paul and Akiba, Takuya and Orii, Naoki, “Japanese StableLM Base Alpha 7B”, https://huggingface.co/stabilityai/japanese-stablelm-base-alpha-7b
[8] K. Fujii and T. Nakamura and M. Loem and H. Iida and Masanari Ohi and Kakeru Hattori and Hirai Shota and Sakae Mizuki and Rio Yokota and Naoaki Okazaki, “Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities” (2024)
[9] Edward J Hu and Yelong Shen and Phillip Wallis and Zeyuan Allen-Zhu and Yuanzhi Li and Shean Wang and Lu Wang and Weizhu Chen, “LoRA : Low-Rank Adaptation of Large Language Models”, In International Conference on Learning Representations (2022)