

# 多段階学習とモデル内部推論に基づく 視覚言語モデルのユーモアキャプション生成

和田拓也<sup>1</sup> 植木一也<sup>1</sup>

<sup>1</sup> 明星大学情報学部情報学科

kazuya.ueki@meisei-u.ac.jp

## 概要

近年の視覚言語モデルは画像の正確な記述に優れるが、人間的なユーモアや共感性を含むキャプション生成は未だ困難である。本研究では、写真をお題とする大喜利タスクを対象に、段階的な学習により創造的なボケを生成する手法を提案する。具体的には、教師あり学習を用いた言語・視覚の基礎学習に加え、GRPOによる自己説明を通じた思考プロセスの強化と、他者評価（スター数）に基づく選好学習を導入した。これにより、単なる状況説明を超え、視点の転換や論理の飛躍を伴う人間らしいユーモアの獲得を目指した。

## 1 はじめに

近年、ChatGPTをはじめとする大規模言語モデル (Large Language Model) や、視覚情報と言語を統合的に扱う視覚言語モデル (Vision-Language Model) の発展により、AIの生成能力は飛躍的に向上している。これらのモデルは、画像の内容を正確に記述することには優れているが、「ユーモア」や「面白さ」といった人間に近い感性的なキャプションを生成することは困難である。

そこで本研究では、段階的なモデルの追加学習を行うことで、写真から創造的なキャプションを生成するモデルを構築する。本研究において「人間に近いユーモアのあるキャプション」とは、単なる画像の状況説明にとどまらず、新しい視点を提示しているか、論理の飛躍が効果的に用いられているか、複数の解釈が可能で読み手の想像を広げるような表現になっているかを指す。

## 2 関連研究

根来ら [1] は、視覚情報だけでなく言語モデルの知識活用に重きを置いたアプローチを提案した。彼

らは「ボケて」データセットを用いて VLM をファインチューニングすることで、従来の CNN+LSTM モデルでは困難であった、画像内の状況と世界知識を組み合わせた高度なボケ生成を実現した。しかし、生成されるユーモアには「共感性の欠如」や「人間と AI の面白さの基準の乖離」といった課題が残されている。

## 3 提案手法



図1 狼の写真

図1の画像に対して、人間がつけた面白いキャプションは、「3匹ともレンガの家だった」であった。そのキャプションは、狼の姿から童話「三匹の子豚」を連想させるユーモアであるのに対し、VLMモデルの出力は「目は半分、やる気はゼロ。」の視覚的特徴を出力した。このように VLM モデルは、画像から連想される知識を元にしたユーモアの生成ができていない。一方、同モデルに人間のキャプションを入力しその面白さを問うと、背景知識に基づいた解説が可能であった。

そこで本手法では、学習プロセスにおいて面白さを自己説明するステージを設けることで、既存モデルが背景知識を保持しながらも、生成時に画像内容と適切に紐づけられる出力を可能とすることを目指す。

図2に提案手法の全体概要図を示す。本研究では、日本語能力に優れた google/gemma-3-27b-it<sup>1)</sup> をベースモデルとし、4段階の学習を実施した。各段階で異なるタスクを学習させることで、人間に近い

1) <https://huggingface.co/google/gemma-3-27b-it>

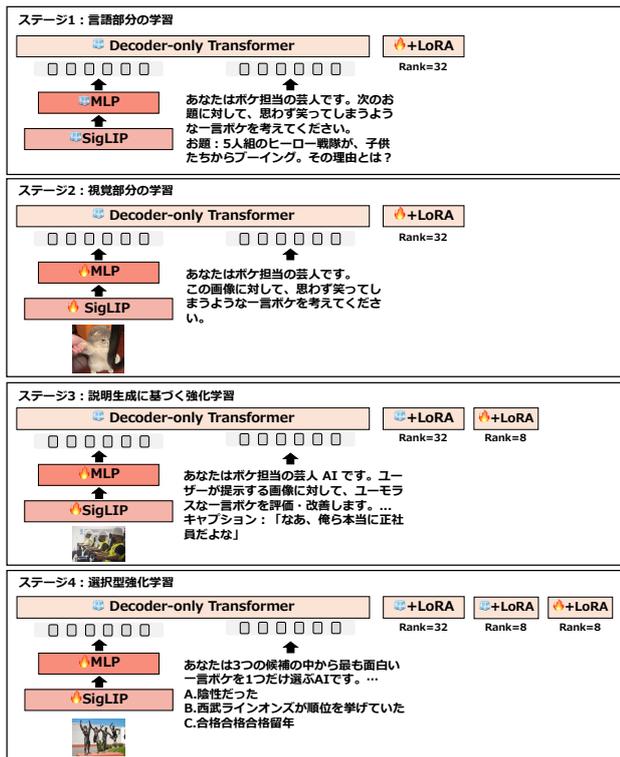


図 2 各ステージのモデル構造

ユーモア（新しい視点の提示，効果的な飛躍，多様な解釈可能性）を備えたキャプション生成能力を向上させた。

### 3.1 ステージ 1：言語部分の教師あり学習

本ステージでは，モデルの言語部分のみを対象に，大喜利データを用いた言語生成能力の強化を行った．学習データには，zhongshsh/CLoT-Oogiri-GO<sup>2)</sup> データセットおよび，NHK「着信御礼！ケータイ大喜利」のテキストデータを使用した．CLoT-Oogiri-GO データセットは，写真とキャプションを投稿・共有するプラットフォーム「ボケて」<sup>3)</sup> の投稿データを収集したものである．この中から，画像内にお題テキストが含まれる大喜利形式のデータのみを抽出し，ソフトバンク社が開発した日本語特化型視覚言語モデル sbintuitions/sarashina2-vision-14b<sup>4)</sup> を用いて，画像内の文字情報をテキスト化した．

学習手法には LoRA (Low-Rank Adaptation) [2] を採用した．本ステージでは LoRA の Rank を 32 に設定した．これは標準的な設定値よりも大きく，後続

2) <https://huggingface.co/datasets/zhongshsh/CLoT-Oogiri-GO>

3) <https://bokete.jp/>

4) <https://huggingface.co/sbintuitions/sarashina2-vision-14b>

するステージ 3 および 4 での追加学習時において，獲得した言語的知識や「ボケ」のニュアンスが失われないう，表現力を十分に確保するための設定である．学習時には，4 種類のシステムプロンプト（「芸人」や「大喜利回答者」の人格付与）からランダムに 1 つを選択してプロンプトを与えることで，出力表現の多様性を確保した．

### 3.2 ステージ 2：視覚部分の教師あり学習

本ステージでは，言語部分（ステージ 1 の LoRA を継続学習）および画像認識部分を学習対象とし，画像とテキストを対象学習させた．データセットには CLoT-Oogiri-GO および「ボケて」から収集した自作データを用い，ステージ 1 同様，4 種類のシステムプロンプト（「芸人」や「大喜利回答者」の人格付与）からランダムに 1 つを選択して入力を行った．

予備実験において全データ（約 12.8 万件）を学習させたところ，モデルが本来有していた視覚言語理解能力が著しく低下する現象が確認された．具体的には，目を閉じた狐の写真を与えて「写真の状況を説明せよ」というプロンプトを入力した際，画像内容を無視して「どうせ俺の写真は SNS にアップされるんだろ」といった，指示とは異なる出力が生成された．この問題を回避するため，学習過程全 1,398 ステップのうち，600 ステップ時点で保存された LoRA を使用した．その結果，同画像に対して「この動物は眠いたため目を閉じています」と出力されるようになり，写真内容と整合した出力がされるようになった．

### 3.3 ステージ 3：説明生成に基づくユーモア強化学習

本ステージでは，画像とキャプションに対して「なぜ面白いのか」を説明させ，その上で改良案を生成する自己フィードバック型の学習を行った．図 A.1 に本ステージで使用したプロンプトを示す．ここでは新たに LoRA (Rank 8) を追加し，画像認識側も学習させた．

学習手法には Group Relative Policy Optimization (GRPO) [3] を採用した．GRPO はグループ内の相対評価に基づいて報酬を与える方式であり，「唯一の正解」が存在しないユーモア生成タスクに適している．本研究では DeepSeek-R1-Zero [3] の方針に倣い，思考過程 (Thinking Process) を明示的に与えず，結果に対する報酬のみで自律的な思考プロセスの形成を促した．

具体的には、報酬関数の設計において、モデルの出力が <think> タグによる思考過程と、それに続く <output> タグによる最終回答という所定の構造に従っているかを正規表現を用いて判定する仕組みを採用した。この形式要件を満たす場合にのみ報酬値 1.0 を与え、満たさない場合は 0.0 とするフォーマット報酬を設定した。これにより、モデルは単に回答を出力するのではなく、キャプションの面白さを言語化（自己説明）し、その分析結果を踏まえて最終的なボケを生成するという推論プロセスを学習した。

### 3.4 ステージ 4：選択型強化学習

本ステージでは、1 枚の画像と 3 つのキャプション候補から最も面白いものを選択し、理由を述べるタスクを学習させた。図 A.2 に本ステージで使用したプロンプトを示す。

正解ラベルには、CLoT-Oogiri-GO データセットに含まれる「スター数 (Star 数)」を用いた。スター数とは、大喜利プラットフォーム「ボケて (Bokete)」における評価指標である。ユーザは 1 つのボケに対して最大 3 つの星を付与することができ、その獲得総数が当該ボケに対する「面白さ」の尺度となる。本研究では、多くのユーザーによって評価されたこのスター数を、一般的な「面白さの度合い」を表す客観的な数値として扱った。

学習データ構築の際は、同一の画像に対して投稿された複数のキャプションから 3 つを抽出し、その中で最もスター獲得数が多いものを「正解」として設定した。評価関数では「正解の選択」およびステージ 3 と同様の「タグ構造の整合性」に報酬を与えた。この学習により、モデルは人間の嗜好や感性に基づくユーモアの基準を自律的に学び取り、出力の最適化を行えるようになった。

## 4 検証実験

### 4.1 実験条件

提案手法の有効性を検証するため、「ボケて」から抽出した多様なジャンルの画像 20 枚を用いて評価実験を行った。比較対象として、未学習モデル (gemma-3-27b-it) に単純な指示のみを与えて生成を行うベースライン条件と、ステージ 1~4 を通じて学習したモデルに思考誘導プロンプト (図 A.3) を適用し、内部推論を経て生成を行う提案手法の 2 条

件を設定した。

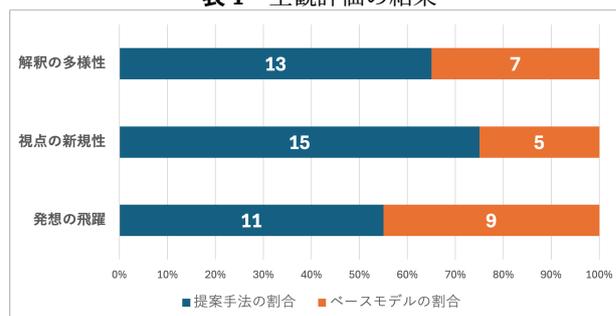
21 名の被験者に対し、同一画像のキャプション 2 種を提示し、「視点の新規性」「発想の飛躍」「解釈の多様性」の 3 観点から、どちらが優れているかを選択させた。

### 4.2 評価結果

20 枚の評価画像に対し、21 名の被験者による投票を行った。その結果、提案手法の得票数が過半数を超えた画像の枚数を集計した結果を表 1 に示す。

表より、「視点の新規性」および「解釈の多様性」においては、それぞれ 75%、65% の画像で提案手法が支持されており、ベースラインと比較して明確な優位性が確認された。一方で、「発想の飛躍」に関しては、全 20 枚中 11 枚 (55%) の支持にとどまり、ベースラインと比べても大きな改善は確認できなかった。

表 1 主観評価の結果



### 4.3 評価結果の分析

**高評価事例** 視覚情報の意味を大胆に再解釈することで高評価を得た事例を、図 3 に示す。対象画像は、審判がイエローカードを提示している場面である。ベースモデルは「試合中の会話」という、画像本来の支配的な文脈に留まった。一方、提案手法は内部推論において、カードを「ハンバーガー」等の金銭的価値に見立て、「手取り 12 万円の提示」という社会的リアリティのある文脈へと転換させた。この「見立て」による意外性は被験者に高く評価され、「視点の新規性 (76.2%)」や「発想の飛躍 (66.7%)」において高い支持を得た。

**低評価事例** 一方、評価が振るわなかった事例を図 4 に示す。画像は穏やかな表情の大仏である。ベースモデルは「晩ご飯の悩み」と解釈し、仏と日常の落差を用いた共感しやすいボケを生成した。対して提案手法は、「昔カノに仏像と褒められた」と

入力写真	モデル	プロンプト	内部推論	最終出力
	学習済みモデル	プロンプトチューニング後	<ol style="list-style-type: none"> <li>1) 競技とは関係ないことをして、審判が口で注意をしている。</li> <li>2) マックのアルバイトで、お客にハンバーガーを強く勧めている。</li> <li>3) オチ構造</li> <li>4) 1枚のハンバーガーの写真で「これ12万円払うなら1万円」と強引に上乘せする。 【画像との整合性チェック】ハンバーガーを1個の値段と「月給」に置き換えてボケにする。</li> </ol>	これ、手取りで12万円です
	ベースモデル	通常プロンプト		審判「ちょっと、その13番！ ホッケーの鍵、どこに置いたか覚えてます？」

図3 高評価事例：提案手法とベースラインの生成プロセス比較

入力写真	モデル	プロンプト	内部推論	最終出力
	学習済みモデル	プロンプトチューニング後	<ol style="list-style-type: none"> <li>1) 仏様の顔がリアルな映像に、ほかし加工がされていない</li> <li>2) イケメンだと思ってたら…元カノの写真に「仏様でも無理です」と書いてあった</li> <li>3) 逆張り</li> <li>4) 仏像に似てる元カノとの写真に対して、ほかしをかけないという大胆な発想でこの仏像の写真と関連付けてみた。</li> </ol>	昔カノに仏像と褒められた
	ベースモデル	通常プロンプト		悟りを開いても、晩ご飯のメニューは悩むらしい。

図4 低評価事例：提案手法とベースラインの生成プロセス比較

いう個人的エピソードへの飛躍を試みたものの、画像の持つ荘厳な雰囲気とは乖離しており、唐突な印象を与えた。結果として、「発想の飛躍」の支持率は28.6%に留まり、ベースラインに大きく劣る結果となった。

## 5 まとめと今後の課題

本研究では、視覚言語モデルにおけるユーモア生成能力の向上を目的として、自己説明を通じた思考プロセスの強化と、スター数に基づく選好学習を統合した多段階学習手法を提案した。

評価実験の結果、提案手法は「視点の新規性」および「解釈の多様性」においてベースラインを大きく上回り、人間の感性に近い多角的な視点の獲得に成功した。特に、モデルが背景知識を柔軟に活用し、画像単体からは得られない文脈を補完することで、ユーモアの質が向上することが確認された。

一方で、「発想の飛躍」と「画像との整合性」の両立には課題が残った。定性分析の結果、モデルが生成した「飛躍」が画像の視覚的根拠から乖離し、読み手の納得感を損なう事例が確認された。これは、

モデルが内部知識の活用を優先するあまり、画像の状況説明をおろそかにした結果であると考えられる。

今後の課題としては、単なる面白さ（スター数）だけでなく、読み手の共感を得るための文脈との整合性を明示的に考慮した報酬関数の設計が挙げられる。また、画像内の状況をより詳細に認識可能にすることで、意外性と納得感を両立したユーモア生成を目指す。

## 参考文献

- [1] 勇太根来, 和宏森田, 正雄泓田. Llm を用いた画像に対する面白キャプション生成. 第23回情報科学技術フォーラム (FIT2024) 講演論文集, 2024. E-027.
- [2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Low-rank adaptation (lora) of large language models. **arXiv preprint arXiv:2106.09685**, 2021.
- [3] Zhihong Shao, Peiyu Wang, Qiuju Zhu, Runxin Xu, Jun Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. **arXiv preprint arXiv:2402.03300**, 2024.

## A 付録

あなたはボケ担当の芸人 AI です。  
ユーザーが提示する画像に対して、ユーモラスな一言ボケを評価・改善します。  
- <think> には次を含める：  
1) 与えられたキャプションがなぜ面白いのか  
2) どこを改善すればさらに面白くなるのか  
- <output> には改善後の一言ボケを 30 文字以内で記述すること。  
キャプション：  
caption  
出力は<think></think>の中に思考を書き、その後<output></output>の中に回答を作成してください。  
例えば以下のように出力してください。  
<think>…</think>  
<output>…</output>

図 A.1 ステージ 3 で使用したシステムプロンプト

あなたは 3 つの候補の中から最も面白い一言ボケを 1 つだけ選ぶ AI です。  
以下の画像を見て、A/B/C のいずれか 1 つを選んでください。  
- 出力形式：  
<think>ここに考えを書く</think>  
<output>A|B|C</output>  
【候補】  
A) optA  
B) optB  
C) optC

図 A.2 ステージ 4 で使用したシステムプロンプト

あなたはボケ担当の芸人 AI です。  
ユーザーが提示する画像に対して、ユーモラスな一言ボケを作成します。  
【厳格な出力規約】  
- 出力は \*\*タグのみ\*\*。  
- タグは小文字で、順序は厳守：  
<think>…</think><answer>…</answer>  
- <think> は 4 行構成、各行はプレフィックス番号を含める：  
1) 画像から注目した要素  
2) そこからの文化的/社会的連想  
3) 12 指標から今回ねらう指標名（例：語呂、駄洒落、逆張り、ひねり、オチ構造、共感、未来予測、文化的背景、擬人化、独創性、シンプルさ、過剰表現）  
4) Creative Leap-of-Thought（大胆な飛躍）のアイデアと、最後の「画像との整合性チェック」ひとこと  
- <answer> は 30 文字以内の一言ボケ。終助詞や語感も工夫する。  
出力は<think></think>の中に思考を書き、その後<answer></answer>の中に回答を作成してください。  
例えば以下のように出力してください。  
<think>…</think>  
<answer>…</answer>

図 A.3 推論時に使用したプロンプト