筑波大学 情報学群 情報メディア創成学類

卒業研究論文

大規模言語モデルを用いた語用論的アプローチに基づく誤解可能性を考慮した画像情報ラベリング

清野 駿

指導教員　若林 啓

2024年2月

概要

本研究では，複数の画像情報が与えられたときに，そのうちの一つを誤解なく指し示すことのできる自然言語ラベルを導出する手法を提案する．

近年，画像情報に対して適切な自然言語ラベルを付与することのできる，マルチモーダルな学習済み大規模言語モデルが広く利用可能になっている．

しかし，大量の画像情報を一度に処理して，それぞれに特有の特徴を反映した自然言語ラベルを付与することは容易ではない．

本研究では，語用論に基づく「聞き手モデル」を導入し，内省的処理を通じてラベルの品質を向上させるアプローチを提案する．

実験ではベースライン手法と 2 つの提案手法の計 3 つの手法によってラベルを生成し，これらの手法によるラベルの誤解可能性を被験者実験を通じて比較した．

目 次

図 目 次

第1章　はじめに

画像情報に対して自然言語のテキストを生成する技術は，様々な産業的な応用への活用が期待される [1]．特に，複数の画像が与えられたときに，それらを誤解なく指し示すことのできるテキスト表現を生成することは，ロボットによる場所のナビゲーションや，人間に対する情報提供を正確に行う上で，重要なタスクである．本稿では，与えられた複数の画像のそれぞれに対して，それらを互いに区別可能な自然言語ラベルを生成する問題を，画像情報ラベリングと呼ぶ．

画像情報を説明する自然言語テキストを生成する手法は，画像キャプション生成の分野で研究されている [2, 3, 4]．特に近年では，与えられた画像情報を説明する自然言語テキストの生成や，画像情報に関する質問に回答することのできる，マルチモーダルな事前学習済み大規模言語モデルが広く利用可能になっている [5]．深層学習モデルは，モデルの規模（パラメータ数）と訓練データの規模を増加させることで品質が向上し続ける傾向があり [6]，画像キャプション生成においても有効性が示されている．このことから，可能な限り大きなスケールで事前学習した共有の大規模言語モデルを，再学習することなく活用する技術の検討が重要になってきている．

しかし，大規模言語モデルは，大量のデータをそのまま入力することはできないことから，多数のデータの関係を考慮した処理には適していない．このため，複数の画像に対して，それらを区別する自然言語ラベルを生成する，画像情報ラベリングの問題設定における性能は十分に検証されていない．似た画像と区別できるようなキャプションを生成する手法の研究は行われている [7, 8, 9, 10] が，これらの研究ではキャプション生成モデルの学習を伴うことから，大規模言語モデルの性能を活用することができないという課題がある．

本研究では，画像情報ラベリングタスクにおいて，大規模言語モデルを用いたテキスト生成手法の性能を評価すると共に，語用論的アプローチに基づいて性能の向上ができるかどうかを検討する．語用論的アプローチは，聞き手の解釈を考慮して，誤解を避けるような言語を生成するアプローチであり，複数のテキスト生成タスクにおける品質の向上が報告されている．本研究では，2 つの手法を通して，画像情報ラベリングにおける語用論的アプローチの有効性を検証する．1 つ目の手法では，大規模言語モデルに生成させた画像情報のラベルの中から，誤解の起きやすいラベルを検出し，再び大規模言語モデルに修正させる．2 つ目の手法では，大規模言語モデルに直接ラベルを生成させるのではなく，多数の特徴をテキストで挙げさせて，そのうちのいくつかを選択してラベルを構成する．この選択においては，誤解の起きやすさを評価する聞き手モデルを内部的に持つことによって，誤解可能性を最小化するような特徴の組み合わせを発見する最適化手法を提案する．

複数の部屋の画像が与えられたときに，それぞれの部屋を指し示すラベルを生成する実験を行った．生成したラベルの評価には Amazon Mechanical Turk を利用し，人間がどの程度誤解するかを検証した．実験の結果から，テキスト長が短いラベルにおいて，提案する語用論的アプローチによって性能が改善することを示す.

第2章　関連研究

本章では，本研究で扱う画像情報ラベリングタスクに関連する研究として，画像に基づくテキスト生成技術について 2.1 節で述べる．また，提案手法の着想に関連する研究として，語用論的アプローチに基づく機械学習手法について 2.2 節で述べる．

2.1　画像に基づくテキスト生成

画像内容に基づくテキスト生成技術の発展は，画像キャプション生成の研究によって牽引されてきた [2]．特に，深層学習を用いた画像キャプション生成では，ニューラルネットワークに画像の特徴を抽出し，得られた特徴に基づいて再帰型ニューラルネットワークでテキストを生成するモデルの枠組みが用いられる [3]．この枠組みにおいて Xu ら [11] は，注意機構を用いて画像内の注目すべき箇所の情報を抽出することで，生成されるキャプションの品質が向上することを示した．

一方近年，タスクによらない画像やテキストの表現学習によって，深層学習の性能が大幅に向上することが明らかになってきた [6]．この知見を発展させて，画像とテキストのマルチモーダルモデルを構成し，これを大規模な画像とテキストのデータセットを用いて特徴量表現を学習する手法が研究されている [1, 5]．事前学習済みの大規模マルチモーダルモデルを用いることで，画像キャプション生成の品質が向上する [4] だけでなく，画像のゼロショット分類 [12] や，画像質問応答 [13] などのタスクの精度も向上することも示されている．

さらに近年では，モデル規模と訓練データをスケールアップすることで，テキストでの指示（プロンプト）に応じた画像に基づくテキスト生成ができることが明らかになった．Li ら [14] は，事前訓練済みの画像特徴量抽出器とテキスト生成器の間を橋渡しする Querying Transformer を訓練する手法によって，画像に関する自然な会話や，柔軟な指示に応じた説明の生成ができることを示した．GPT-4[15] は，入力した画像を描画するプログラムを出力したり，グラフの解釈もできることを示している．例えば,「この画像の特徴を 10 個挙げてください」というプロンプトと共に画像を入力することで，自然言語テキストで画像の特徴を得ることができる．

これらの深層学習に基づく手法はブラックボックスであり，どこまでのタスクを実現できるかは未知数であるが，基本的には 1 枚あるいは少数の画像とテキストの間を結びつける処理を想定している．このため，多数の画像に対して，それらを区別する自然言語ラベルを生成する問題設定における性能は十分に検証されていない．本研究では，画像情報ラベリングタスクにおいて，これらのテキスト生成手法の性能を評価すると共に，画像情報ラベリングタスクを分割して再構成することによるタスク性能の向上を検討する．

2.2 語用論的アプローチに基づく機械学習

言語の生成にあたって，聞き手の解釈を考慮して，誤解を避けるような語を選ぶことは，コミュニケーションにおいて重要な要素である．このような，話し手による聞き手の解釈を考慮したアプローチは，計算語用論と呼ばれる [16]．Rational Speech Acts (RSA) フレームワークは，計算語用論の代表的な計算モデルである [17, 18]．RSA フレームワークでは，話し手と聞き手のモデルが，それぞれ相手の目的や解釈について再帰的に推論する過程を計算する．例えば，RSA フレームワークにおける話し手モデルは，内部的に持つ聞き手モデルに発話を解釈させたときに，伝達したい意図通りの解釈が得られるような発話を選択する．

この知見を反映して，RSA フレームワークに基づいてテキスト生成タスクの品質を向上させる手法が研究されている [19]．William ら [20] は，室内のナビゲーションタスクにおいて，計算語用論に基づく手法を提案した．話し手モデルによって生成された複数の指示を聞き手モデルに入力し，聞き手モデルが意図通りに解釈する指示を出力する．SAIL データセットを用いた評価により，語用論的アプローチによってナビゲーションの成功率が向上することが示された．

画像キャプション生成においても，語用論的アプローチによって，聞き手が似た画像と区別できるようなキャプションを生成する手法が提案されている [7, 8, 9]．Andreas ら [10] は，画像の参照ゲームにおいて，聞き手の誤解を低減するようなキャプションを生成する手法を提案した．これらの手法は，ニューラルネットワークでモデル化された話し手と聞き手の学習を伴うことから，近年極めて高い品質を達成している大規模事前学習済みモデルを活用することが難しい．本研究では，大規模事前学習済みモデルを再学習することなく用いる語用論的アプローチによって，誤解可能性を考慮した画像情報ラベリングを実現する手法を検討する．

第3章　形式

ここでは、論文の表紙および本体の記述方法について述べる。

3.1　表紙

表紙は、\maketitle によって作成するため、以下の項目に相当する文字列をそれぞれ記述する。

題目: 題目は \title に記述する。行替えを行う場合は \\ を入力する。ただし、題目の最後に \\ を入力するとコンパイルが通らなくなるので注意する。なお、4 行以上の題目の場合、表紙ページがあふれるためスタイルファイル “mast-jp-xxx.sty” を変更する必要がある（xxx は使用文字コードに合わせて euc, sjis, utf8 のいずれかになる）。

著者名: 著者名は \author に記述する。

指導教員名: 指導教教員は \advisor に記述する。

年月: 年月は \yearandmonth に記述する。
年月は別途指示された場合はそれにしたがう（指示がなければ提出時のものを記述する）。

3.2　本体

本体は 1 段組で記述する。

図表には番号と説明 (caption) を付け、文章中で参照する。表 3.1 と図 3.1 はそれぞれ表と図の例である。表の説明は上に、図の説明は下に書くことが多い。図の挿入に用いるパッケージについては使用環境に合わせて自由に選択してほしい。

表 3.1: 表の例

年 度	1 年次	2 年次	3 年次	4 年次
1995	85	92	86	88
1996	83	89	90	102
1997	88	87	91	112
1998	144	93	90	115

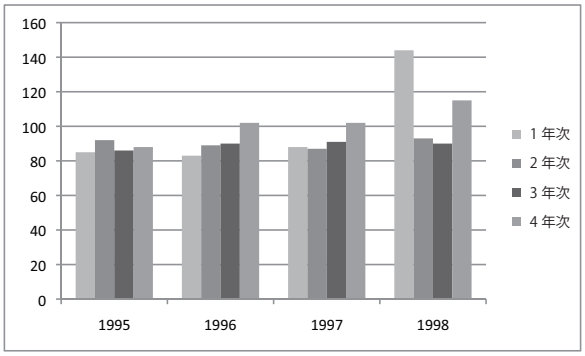

図 3.1: 図の例 （graphicx パッケージを使用）

詳しくは参考書 [**?**, **?**] などを参照のこと。奥村晴彦氏の「TEXWiki」 https://texwiki.texjp.org/ は日本語の TEX に関する情報が充実している。具体的な文献の参照例として本の例 [**?**]、雑誌論文の例 [**?**]、予稿集の例 [**?**] を挙げておく。

謝辞

　必須ではないが、書くことが望ましい。
研究補助を受けている場合、他に指定がなければここに書く。

参考文献

[1] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv [cs.CV] 2306.13549*, 2023.

[2] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: generating sentences from images. In *Proceedings of the 11th European conference on Computer vision: Part IV*, pp. 15–29, 2010.

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 39, No. 4, pp. 652–663, 2017.

[4] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in Vision-Language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2136–2148, 2023.

[5] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-Language pre-training: Basics, recent advances, and future trends. *arXiv [cs.CV] 2210.09263*, 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[7] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. 2017.

[8] Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. Pragmatically informative image captioning with Character-Level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 439–443, 2018.

[9] Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. Pragmatic Issue-Sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1924–1938, 2020.

[10] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1173–1182, 2016.

[11] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of Machine Learning Research*, Vol. 37, pp. 2048–2057, 2015.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research*, Vol. 139, pp. 8748–8763, 2021.

[13] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are Few-Shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6088–6100, 2022.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image pre-training with frozen image encoders and large language models. 2023.

[15] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv [cs.CL] 2303.12712*, 2023.

[16] Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12619–12640, 2023.

[17] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, Vol. 336, No. 6084, p. 998, 2012.

[18] Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.*, Vol. 20, No. 11, pp. 818–829, 2016.

[19] Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. In *North American Chapter of the Association for Computational Linguistics*, pp. 1951–1963, 2017.

[20] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going beyond literal command-based instructions: extending robotic natural language interaction capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1387–1393, 2015.