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　提案手法

3.1　問題設定

本研究では，画像の集合が入力されたときに，それぞれの画像を，人が一意に誤解なく特定できるラベルを生成する問題に取り組む．

n 枚の画像の集合を $X = \{x_1, x_2, \ldots, x_n\}$ とする．X に対して付与されるラベルの集合を $Y = \{y_1, y_2, \ldots, y_n\}$ とする．ラベル $y_1, y_2, \ldots, y_n$ は画像 $x_1, x_2, \ldots, x_n$ にそれぞれ対応しており，各ラベルは対応する画像の特徴を説明するものとする．

ここで，ラベル y_r（$r \in [1..n]$）と X を被験者に与えたとき，被験者が y_r に対応する画像 x_r を選択しようとして，x_r 以外を選択してしまう確率を誤解可能性 p とする．

提案手法では，p が最小となるラベルの組み合わせ Y を生成することを目的とする．

3.2　話し手モデルと聞き手モデル

本研究の手法では，語用論的アプローチに基づき問題解決を図る．

語用論とは，話し手と聞き手が互いの意図を理解する過程を研究する学問であり，皮肉表現や比喩表現などを研究対象とする．

本研究での語用論的アプローチは，話し手モデルと聞き手モデルの対話を通じて，問題への回答を改善していく．

話し手モデルは，画像を詳細かつ正確に説明するラベルを生成することに重点を置く．この過程では，画像の特徴や意味内容を考慮し，誤解を引き起こす可能性のある曖昧な表現を避けることが重要である．一方，聞き手モデルは，生成されたラベルがどの程度誤解を招く可能性があるかを評価する．このモデルは，ラベルが提供する情報が明確で，かつ他のラベルとの誤解が起きないかどうかを評価し，話し手モデルに対してフィードバックを行う．

3.2.1　話し手モデルと聞き手モデルの役割

- 話し手モデル：画像を説明するラベルを生成し，その改善を行う．
- 聞き手モデル：生成されたラベルに対する誤解可能性を評価し，各画像を誤解なく選択できるかを判断する．

3.3 Pragmatic ChatGPT（PCG）手法

PCG では，話し手モデルと聞き手モデルの両方を ChatGPT を用いて実装し，ラベル生成を行う．

この手法では，以下のステップでラベルを生成し，繰り返し改善を図る．このステップでは，1，3 が話し手モデル，2 が聞き手モデルの役割を担っている．

1. ChatGPT に対して画像の集合 X を入力する．X に含まれる各画像に対して，それぞれ誤解なく画像を区別できるラベルを出力するように，プロンプトを入力する．ここで出力されたラベルの集合を Y1 とする．

2. ラベルの集合 Y1 と画像の集合 X を ChatGPT に入力する．Y1 に含まれる各ラベルについて誤解可能性を分析し，ラベル改善のヒントを出力するように，プロンプトを入力する．出力されたラベル改善のヒントを H1 とする．

3. ChatGPT に対して画像の集合 X，ラベル改善のヒント H1，ラベルの集合 Y1 を入力する．H1 を参考にして，Y1 を改善した新たなラベルを出力するように，プロンプトを入力する．ここで出力されたラベルの集合を Y2 とする．

4. 2 と 3 を繰り返して，最終的に得られたラベルの集合 Ym（m は任意の自然数）を，誤解可能性の低いラベルとして提出する．

各モデルごとに行っている処理を以下に示す．

- 話し手モデル（生成モデル）
  - 初期ラベル生成：X に含まれる各画像に対応するラベル Y1 を出力する．
  - 改善ラベル生成：聞き手モデルによって出力されたヒント H と，X，Y1 から，Y1 を改善したラベル Y2 を出力する．
- 聞き手モデル
  - ヒント生成：X と Y1 から，ラベル改善のヒント H を出力する

なお，出力するラベルの長さに制限をかける場合は，ラベル出力処理を行っている部分のプロンプトに以下を追加する．この例では 10 単語以内でラベルを出力するように指定している．

> Tray ceiling Recessed lighting, Beige wall paint, ..., Decorative straw hats on wall, Distressed wood mirror frame

図 3.1: プロンプトの例

3.4　特徴量に基づく組み合わせ最適化手法

3.4.1　全体の流れ

特徴量に基づく組合せ最適化手法（FCO）では，画像から特徴を表す単語群を生成し，これを基に最適なラベルを導き出していく．このプロセスは，話し手モデルと聞き手モデルの相互作用によって実現される．以上の処理の流れをまとめた図を図 3.2 に示す．

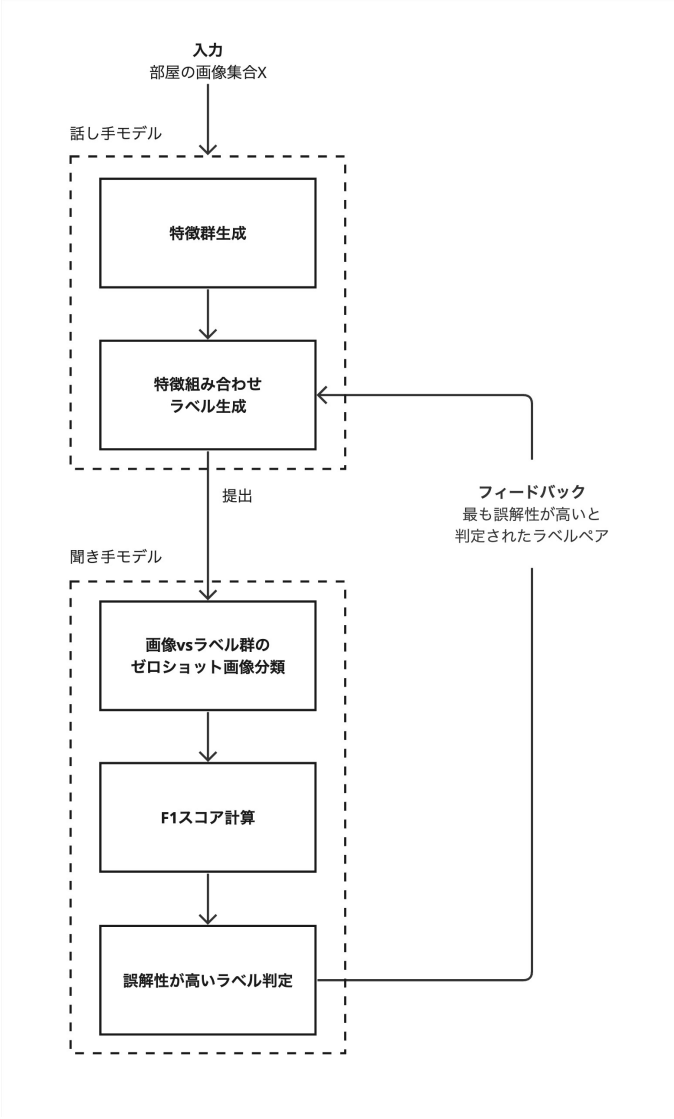

図 3.2: ラベル生成の流れ

3.4.2 特徴群生成とラベル生成

初めに，話し手モデル（ChatGPT-4）に画像群 X を入力し，各画像から特徴を表す単語群 Z を生成する．特徴群 Z の例は図 3.3 に示す．この際，各画像からは 50 個の特徴 z が得られる．次に，これらの特徴群 Z からランダムに特徴を選び，組み合わせてラベル y を生成する．これを画像群 X 全体で行い，ラベル集合 Y_1 を作成する．

Tray ceiling Recessed lighting, Beige wall paint, ..., Decorative straw hats on wall, Distressed wood mirror frame

図 3.3: 部屋 X1 の特徴群 Z1

続いて，聞き手モデルに Y_1 と X を入力する．このモデルは，各ラベルの誤解可能性を分析し，最も誤解性が高いラベル組をヒント H_1 として出力する．

3.4.3 ゼロショット画像分類

生成されたラベルの誤解可能性を評価するため，ゼロショット画像分類を実施する．このプロセスでは，生成されたラベル群 Y_1 と単一の画像 x_1 を入力として与える．ゼロショット画像分類を適用し，画像 x_1 が各ラベルに属する確率を確率分布として出力する．この手順は画像 x_1 から x_n に至るまで同様に行われる．実際にゼロショット分類をした際の結果を図 3.4 に示す.

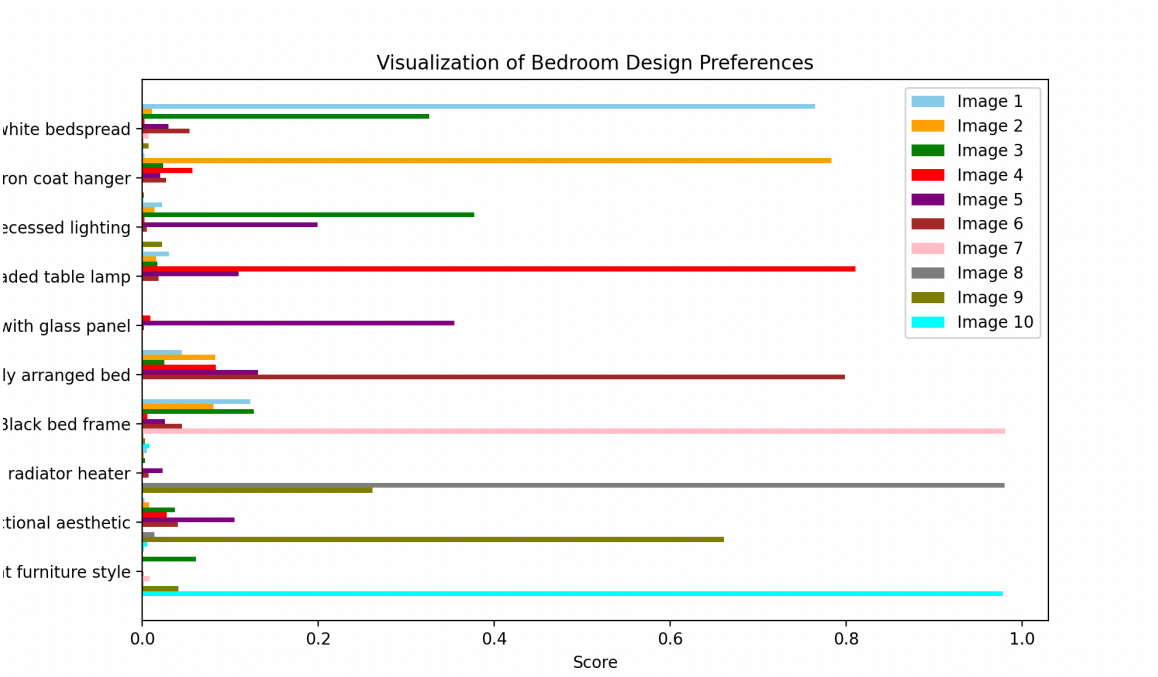

図 3.4: ゼロショット分類の結果

3.4.4 F1 スコア計算

ラベルセットの誤解可能性はゼロショット画像分類の結果を用いて計算された F1 スコアで評価される．ゼロショット画像分類を行ったときに，i 番目の画像に対して，j 番目のラベルに与えられた確率を $p_{i,j}$ とすると，ラベル j の適合率は $\frac{p_{j,j}}{\sum_i (p_{i,j})}$，再現率は $p_{j,j}$ である．

ラベル j の F1 スコアは，ラベル j の適合率と再現率の調和平均で求められる．そして全ラベル Y に対してそれぞれ F1 スコアを求め，全 F1 スコアの平均をラベルの性能とする．

3.4.5 誤解性の高いラベルペア判定

ゼロショット画像分類の結果を分析して改善すべきラベルのペアを特定する過程では，画像 x_k を説明したラベル y_k とし，x_k 以外の画像の中で y_k への所属確率が最も高いものを y_1 から y_n で取得する．自分の画像以外の画像の所属確率が最も高いラベルと，最も高い画像の真の正解ラベルを改善ペアとして出力する．このペアは，ゼロショット分類において最も誤解されやすい，つまり最も性能が低いと判断されるラベルと，そのラベルが誤って関連付けられた画像の正しいラベルである．

3.4.6 聞き手モデルから受け取ったヒントをもとにラベルを改善

次に，話し手モデルに H_1 を入力し，Y_1 を改善した新たなラベル集合 Y_2 を生成する．この際，画像とテキストのペアを大規模に学習することで，画像の内容を説明するテキストとの類似度を計算できるモデルである CLIP（Contrastive Language–Image Pre-training）を使用する．CLIP は，画像とテキストの両方を同じ空間に埋め込むことで，その類似度を測定する．ここでの類似度は画像とラベルの特徴の重要度として扱う．各特徴と画像との重要度の例は表 3.1 に示す．

表 3.1: 各特徴と画像との重要度の具体例

特徴	重要度
Tray ceiling Recessed lighting	28.051437
Two-tone wood flooring	25.21447
White ceiling	25.036436
...	
Mountainous landscape view	17.63242
Glass vase with white flowers	17.470116
Striped throw on the bed	16.694729

与えられた改善すべき 2 つのラベルセット（`current_label1` と `current_label2`）に対して，それぞれのラベルセットが表す画像（`image1` と `image2`）との重要度を参考にしながら改善する処理を行う．以下にその処理をステップごとに説明する．

1. **current_label1 vs image1**
   `current_label1` の中から，`image1` との類似度が最も低いラベルを確率的に選択．これは，類似度の逆数をソフトマックス関数にかけることで確率分布を作成し，その分布に基づいてランダムに選択することで行う．

2. **all_label1 vs image1**
   `all_label1` の中から `current_label1` に含まれていないラベルを対象に，`image1` との類似度が高いラベルから確率的に選択する．

3. **current_label1 vs image2**
   `current_label1` の中から，`image2` との類似度が高いラベルを確率的に選択．これは，類似度をソフトマックス関数にかけることで確率分布を作成し，その分布に基づいてランダムに選択することで行う．

4. **all_label1 vs image2**
   `all_label1` の中から `current_label1` に含まれていないラベルを対象に，`image2` との類似度が低いラベルを確率的に選択する．

5. **ラベルの入れ替え**
   current_label1 からステップ 1 で選んだラベルとステップ 4 で選んだラベルを除外し，ステップ 2 と 3 で選んだラベルを追加して新しいラベルセット new_label1 を作成する．

6. **重複の削除**
   最後に，new_label1 と new_label2 から重複しているラベルを削除する．これにより，各ラベルセットがユニークなラベルのみを含むようになる．

7. **curent_label2 でも同様の処理をして new_label2 を作成する．**

8. **改善されたラベルセットの返却**
   処理を経て改善された new_label1 と new_label2 を関数の出力として返却する．

これにより，current_label1 と current_label2 はそれぞれの画像に対する識別性が改善され，画像とラベルの関連性が高まるように調整される．このプロセスを繰り返し，ラベル集合 Y_m を誤解可能性が低くなるように改善していく．最終的に，全試行の中で最も平均 F1 スコアが高い時のラベル組み合わせを出力する．全試行内の平均 F1 スコアが以下の図 3.5 の場合，赤丸の時のラベルセットを最終的な出力とする．

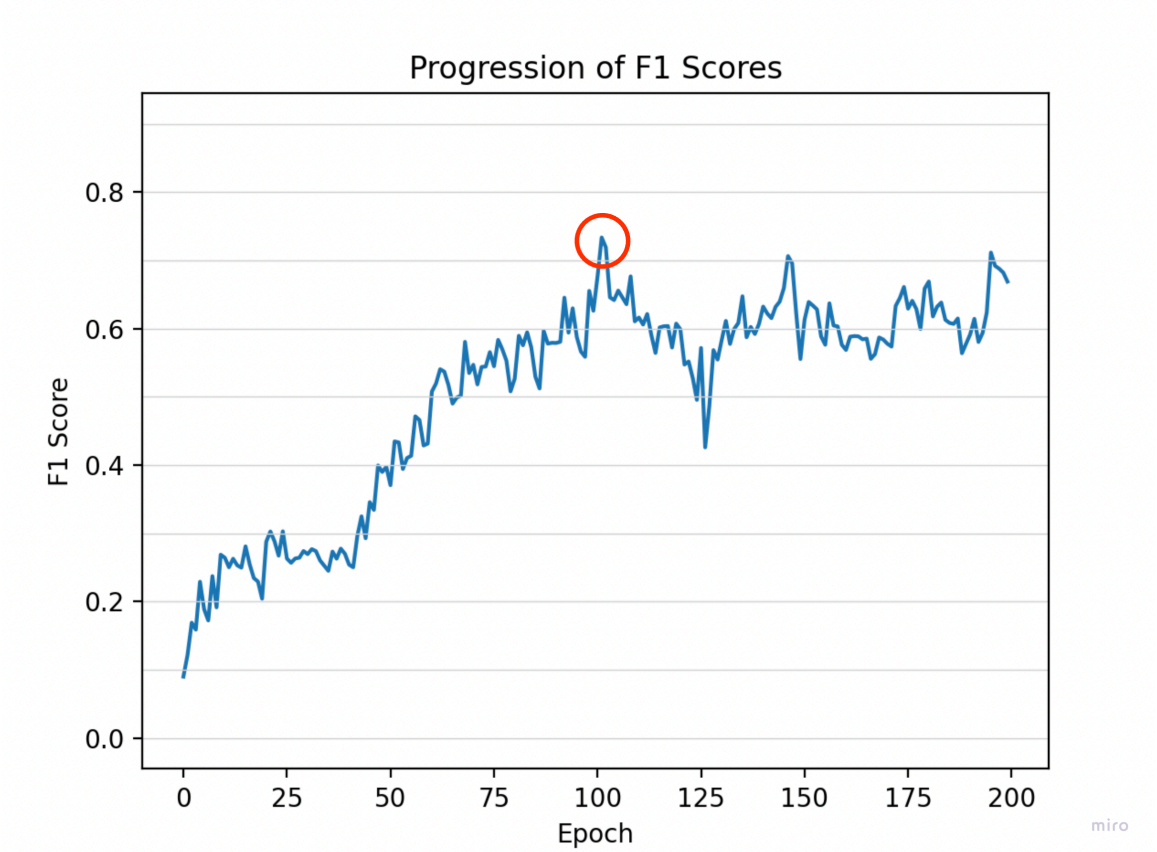

図 3.5: 200 回試行内の F1 スコアの推移

この手法では，話し手モデルが特徴群の生成，特徴の組み合わせによるラベル生成，ヒントに基づくラベルの改善を担い，聞き手モデルが各画像の選択の正確性評価，ラベルの誤解可能性の分析，改善すべきラベル組のヒント出力を行う．この相互作用により，画像に対する最適なラベルを効果的に生成することを試みる．

3.5 特徴量に基づく組み合わせ最適化手法 (New)

本節では，LLM に直接ラベルを生成させるのではなく，画像情報の特徴を列挙させ，そのうちのいくつかを選択してラベルを構成する提案手法について述べる．提案手法は，聞き手がラベルを解釈し，指し示している画像の候補を絞り込む過程を明示的に考慮する．

画像 x_i に対して，特徴量 $Z_i = \{z_1, \ldots, z_m\}$ を生成させる．m は 1 枚の画像情報あたりの特徴量の数である．例えば，図 x の部屋であれば，$z_1 =$ 'red bed', $z_2 =$ '...' など

第4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。

4.1　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）。

4.2　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 4.1 と図 4.1 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 4.1: 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

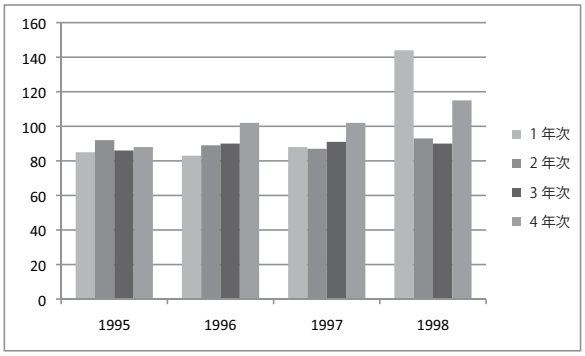

図 4.1: 図の例 （graphicx パッケージを使用）

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.