レストラン検索・予約サイトの投稿画像分類における マルチモーダルモデルの適用検証

柳沢勇気

株式会社カカクコム

yanagisawa_yuki@kakaku.com

概要

レストラン検索・予約サイトの投稿画像を「料理」や「ドリンク」といった種別に分類することは，ユーザーが目的の写真を効率よく閲覧することに貢献する．大量の画像を人手で分類するのは多大な時間と労力を有するため，自動分類モデルを用意するのが望ましい．しかし，従来の教師あり学習モデルは学習データの作成にコストがかかる．そこで，大量の画像–テキストペアで事前学習されたマルチモーダルモデルを使うことで，学習データ作成にかかるコストを無くし，画像をゼロショットで自動分類することを試みた．その結果，複数の画像種別において実用的な分類性能を達成し，人手による作業を大幅に削減できることを確認できた．

1 はじめに

レストラン検索・予約サイトにおいて，ユーザーが投稿した画像を「料理」や「ドリンク」といった種別ごとに分類することは，ユーザーが目的の写真を効率よく閲覧することに貢献する．

株式会社カカクコムが運営するレストラン検索・予約サイト「食べログ」 [1] においては，ユーザーが投稿した画像を「料理」，「ドリンク」，「メニュー」，「内観」，「外観」，「その他」の種別へ分類している．これにより，ユーザーは飲食店の画像を効率よく探すことができるようになる．食べログでは，日々大量の画像が投稿されており，これを 24 時間体制で適切な種別へ人手で分類している．

大量の画像を人手で分類するのは多大な時間と労力を有するため，自動分類モデルを用意するのが望ましい．解決方法としては，自前の学習データセット（画像と画像に付与したいラベルがペアになったデータを集めたもの）を使い，教師あり学習によって学習した自動分類モデルを利用する方法が考えられる．

しかし，これにはモデルを学習させるためのデータを作成する工程に多大な時間と労力がかかる．特に，実サービスにおける運用では，分類ラベルの追加・変更をすることがある．そのたびにデータの収集とラベリングを繰り返し実施する必要があるという問題がある．

そこで本研究では，大量の画像–テキストペアで学習したマルチモーダルモデルである Contrastive Language-Image Pre-training（CLIP） [2] を活用することで，モデルの学習にかかる時間や労力を抑えつつ，レストラン検索・予約サイトにおける投稿画像の人手での分類作業を削減することを試みた．その結果，複数の画像種別において実用的な分類性能を達成し，人手による画像分類の作業を大幅に削減できた．

2 導入方法

食べログでは，ユーザーが投稿した画像を 24 時間体制で適切な種別へ人手で分類している．分類作業は，品質向上の観点から 1 画像につき 2 回実施しており，一次作業で種別の振り分け，二次作業で再確認して確定という流れである．

人手での分類は高品質な分類を実現する一方で，多大な労力とコストを要する．本研究では，CLIP によるゼロショット分類を導入し，一次作業の振り分け作業を「人」から「CLIP での分類」に置き換える方法で人手による一次作業の削減を図った．な

お，置き換えにあたっては，分類確度が高い結果のみを置き換える方針とした．分類確度が低い結果を置き換えると，分類品質の低下や二次作業での修正増加を招くリスクがあるためである．本研究では，適合率（特定の種別に分類された画像全体のうち，実際にその種別に属している画像の割合）が 80%以上の種別を人手での一次作業の対象から CLIP への分類に置き換える方針とした．

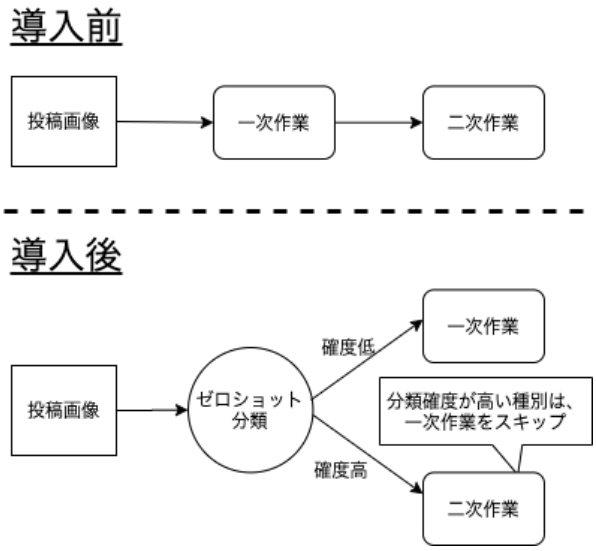

図 1　ゼロショット分類導入前後の比較

3 実験

実際にユーザーが投稿した画像を評価データとして，各種別に分類できるかを評価する．また，提案手法に記載した一次作業の削減率も評価する．

3.1 評価データセット

評価データセットは，実際にユーザーから投稿があった画像から 1 万枚をサンプリングしたものを使用した．種別ごとのデータ数は以下のとおりである．

- 料理: 6,365 件（63.7%）
- ドリンク: 636 件（6.4%）
- メニュー: 1,002 件（10.0%）
- 内観: 233 件（2.3%）
- 外観: 603 件（6.0%）
- その他: 1,161 件（11.7%）

カッコ内は評価データ全体に対する比率を指しており，これは実際の投稿実績と同等である．評価データの画像の正解種別は人手で付与した種別を使用した．なお，学習データはゼロショット学習のモデルを採用しているため不要である．

3.2 分類手法

CLIP による投稿画像のゼロショット分類の手法（以下，本手法）についての概要を図 2 に示した．CLIP による投稿画像のゼロショット分類にあたっては，分類を行う画像および，分類先である各種別を説明する適切なテキスト（プロンプト）を用意する．

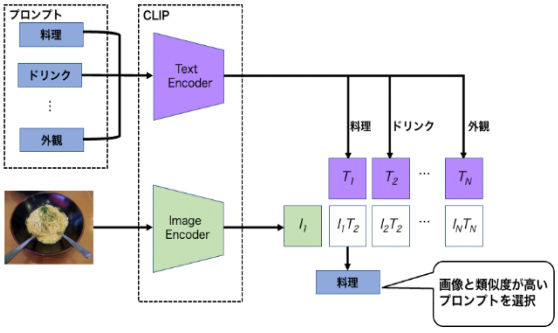

図 2　CLIP を使ったゼロショット分類

上記のプロンプトと画像に対し CLIP の Image Encoder と Text Encoder を使用して，個々のプロンプトと画像をそれぞれ特徴ベクトルに変換する．

画像の特徴ベクトルと各プロンプトの特徴ベクトルとの間のコサイン類似度を計算し，ソフトマックス関数を適用することで，各プロンプトがその画像に対してどれだけ「適合しているか」という確率を得る．画像と各プロンプトとの確率を画像がそのプロンプトで説明される種別に分類される確率とみなし，最も高い確率値を示した種別に画像を分類する．ただし，種別ごとに閾値を設定しておき，確率値が「その他」カテゴリを除く全ての種別の閾値を下回る場合には画像は「その他」カテゴリに分類する．

使用した CLIP の事前学習済みモデルは，rinna 株式会社によって提供された japanese-cloob-vit-b-16 [3] である．このモデルは，Conceptual 12M [4] の 1,200 万の言語・画像ペアのデータを日本語に翻訳し学習データとして使用している [5] ．

3.3 プロンプト設計

本研究では，日本語テキストで学習されたモデルを使用した．そのため，プロンプトも日本語テキストで準備した．プロンプトは，基本的に種別の名称をそのまま使用したが，2 つの観点で工夫をした．

3.3.1 メニュー

「メニュー」に分類したいのは，料理やドリンクの名称と価格が記載された，いわゆる食事メニューである．プロンプトを「メニュー」とした場合，食事メニューではない張り紙や看板などの画像が，「メニュー」に分類されてしまう傾向があった．これらの画像は「その他」種別へ分類したい．そこで本研究では，プロンプトを「フードメニュー」に修正し，張り紙や看板などの画像と食事メニューが同じ種別に分類される確率が下がるよう工夫した．

3.3.2 その他

「その他」は，「料理」，「ドリンク」，「メニュー」，「内観」，「外観」のどれにも該当しない画像を分類する．「その他」へ分類する画像は多岐に渡るため，プロンプトを網羅的に設計することは難しい．そこで本研究では，その他を表すプロンプトは設定せず，「その他」以外の種別のプロンプトと画像の確率値が一定より下回る場合に「その他」へ分類するようにした．

3.4 確率値の閾値調整

「2. 導入方法」において，適合率が 80%以上の種別を人手による一次作業から CLIP での分類に置き換えることとした．そのため，各種別の適合率が 80%以上になるように，3.2 で述べた確率値の閾値を調整した．図 3 に，各種別の適合率と確率値の閾値の関係を示した．

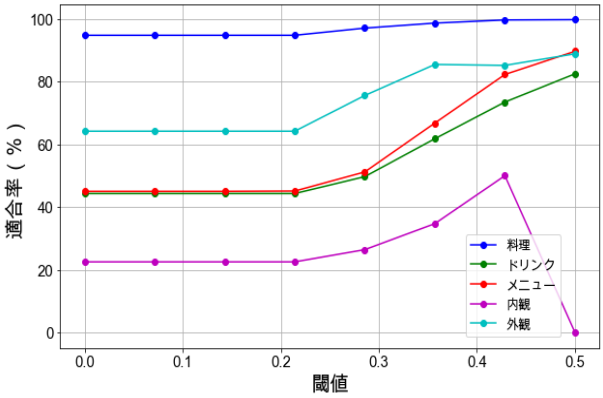

図 3　閾値ごとの適合率の推移

図 3 によると，「料理」に関しては閾値 0 で適合率 80%以上を満たしていることがわかる．一方で「ドリンク」については，閾値 0 では 80%を満たさず 0.48 付近で適合率の 80%を満たすことがわかる．よって，「ドリンク」に分類された場合には 0.48 を下回る画像を「その他」へ分類する．同様の手順で「メニュー」，「外観」の閾値も設定した．「内観」の適合率は閾値 0.4～0.5 で 50%程度まで上昇しているが，0.5 に到達したところで 0%になる．これは分類種別が「内観」であるデータで，0.5 以上の確率値を持つ結果が 0 件であることを示している．よって，「内観」については，閾値の調整で適合率 80%を満たすことはできなかった．

3.5 評価結果

3.5.1 性能評価

本研究による画像分類の種別ごとの性能評価を表 1 に示した．「料理」，「ドリンク」，「メニュー」，「外観」にて適合率 80%以上を達成することができた．対して「内観」と「その他」については，適合率 80%を達成することができなかった．そのため，参考として閾値を 0.3 にしたときの結果を記載した．なお，「その他」については全種別に対して閾値 0.3 を設定（閾値 0.3 を下回った際に「その他」の種別を付与）したときの性能を記載した．

表 1　各種別の性能評価(%)

種別	閾値	適合率	再現率	F1
料理	0	94.8	82.3	88.1
ドリンク	0.48	80.4	33.5	47.3
メニュー	0.41	80.2	66.5	72.7
内観	0.3	27.6	24.9	26.2
外観	0.315	80.4	44.3	57.1
その他	0.3	30.7	46.5	37.0

3.5.2 一次作業の削減率

人手での種別の振り分けにおける一次作業の削減数の合計は，6,949 件であった．評価データは 1 万件のため，削減率は 69.49%を達成できた．削減の内訳は，料理: 5,522，ドリンク: 265，メニュー: 830，外観: 332 である．

3.6 考察

本研究の手法で，全体として 70%近くの一次作業を削減できることがわかった．一次作業の削減の内訳としては「料理」が多くを占めている．これは食べログにおける投稿画像の 60%以上が「料理」に該当する画像でありかつ「料理」の種別の分類性能が高かったことに起因する．分類対象のデータセットの比率の高いデータに対する分類性能が高く出たことから，本研究の手法は本タスクに対して有効であったと結論付けることができる．

種別ごとの結果を見ると「内観」の性能が低いことがわかる．分類結果を詳細に見たところ，「ビルなどの建物内にある店舗の入口」など「外観」へ分類すべき画像が「内観」とへ誤分類される傾向があった．この問題を解決する手段としては，プロンプトの変更が考えられる．たとえば，「内観」のプロンプトを「飲食店の内部」，「テーブルと椅子が配置された飲食店」などのように「内観」へ分類したい画像の特徴を具体的に説明するようなテキストを設定するなどが考えられる．

4 おわりに

本研究では，マルチモーダルモデルである CLIP を使用して，レストラン検索・予約サイトにおける投稿画像の人手での分類作業の削減に取り組んだ．その結果，人手による分類作業における一次作業をおよそ 70%削減することに成功した．CLIPのゼロショット分類を活用することで，モデル学習のためのデータ収集や再学習などの運用コストを低減した自動分類システムの構築が可能であることがわかった．

今後の課題として，プロンプトの最適化や事前学習モデルの変更などによる，分類性能の向上が考えられる．また，今回設定した種別以外の分類タスク（ガイドラインに反する画像へタグ付けするなど）への適用検証にも取り組みたい．

参考文献

1. レストラン検索・予約サイト「食べログ」. https://tabelog.com/, 2024/1/10 閲覧

2. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020, 2021

3. 沢田 慶, シーン 誠, 趙 天雨. 日本語における AI の民主化を目指した事前学習モデルの公開．人工知能学会研究会資料 言語･音声理解と対話処理研究会. 2022/12/13 - 2022/12/14

4. Soravit Changpinyo, Piyush Sharma, Nan Ding, Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. arXiv:2102.08981, 2021

5. rinna 社、日本語に特化した言語画像モデル CLIP を公開. https://rinna.co.jp/news/2022/05/20220512.html, 2024/1/10 閲覧