1U-07

CLIPを用いた細粒度分類食事画像データセットの構築

渡部 光貴† 山肩 洋子† 相澤 清晴†
東京大学†

1 はじめに

画像処理分野における研究が進められる中で, 注目されている分野の一つとして食事画像を用いた研究が挙げられる. 特に, 食事画像分類タスクは食事管理やレストラン業界, ソーシャルメディアなど多岐にわたる応用可能性を持つ. このタスク特有の難しさとして, 「クラス数の多さ」と「クラス内の多様性」が挙げられる. 実際, 世界各地には多種多様な料理や食材が存在し, 一つの料理においてもレシピや調理法, 盛り付け方などで見え方はさまざまである. そのため, 従来の数百クラスでの食事画像の学習・分類では粒度が粗く, 人間の想定した料理名とは違うものが予測結果として表示されてしまうなどの問題がある.

食事の粒度の高さによって生じる問題を解決するためには, クラス数が従来のものよりも数倍多い食事画像データセットで学習を行い, 粒度の高い分類に対応した食事画像分類器を作成することが望ましい. しかし, 1,000 クラスを超える食事画像データセットは現状限られており, 我々が知る限りでは Food2K [1] が存在するが, データの偏りやラベリングの精度に関してはさらなる検証が必要であると考えられる. 特に, 地域特有の郷土料理に偏ったデータ構成や, 視覚的に区別が困難な類似クラスの存在, 直訳による不適切なラベリングが, 画像分類タスクの精度に影響を与えている.

そこで本論文では画像とテキストを同一特徴量空間にマッピングすることを可能とする Vision-Language 事前学習モデル CLIP [2] を活用して, 細粒度分類食事画像データセットを構築する手法を提案する. また, 本研究室で開発・運営を行っている食事管理アプリケーション FoodLog Athl [3] を通して得られた食事画像データを用いて 2,000 クラス近くの多クラス食事画像データセット「FoodLog-1921」を構築し, その過程を示す.

Development of an Fine-Grained Food Image Dataset Using CLIP
†Mitsuki Watanabe, The University of Tokyo
†Yoko Yamakata, The University of Tokyo
†Kiyoharu Aizawa, The University of Tokyo
本研究の一部は, JST JPMJCR22U4 の支援を受けた

2 手法

2.1 データ概要

FoodLog Athl とは, スマートフォンを使用して撮影した食事画像を選択すると, 画像認識技術により画像中の料理が検出され簡単に食事記録をつけることができるアプリである. さらに, アプリ内でユーザーと管理栄養士が会話することを可能としており, FoodLog Athl を用いて食事の記録を取ることで, 栄養士から食事に関するフィードバックを受けることができる.

本研究でデータセット作成に用いたデータは 2018 年 10 月から 2023 年 9 月までの間に FoodLog Athl 内で集められた食事記録で, その中でも栄養士によって管理されていた信頼性の高い食事記録約 20 万件を使用した. 集められたデータの特徴として以下の点が挙げられる.

- バウンディングボックス付きのデータである
- ユーザーのほとんどが日本人であるため, 日本人の食事の傾向を反映している
- ユーザーの日々の食事に基づいているため, 購入された調理食品など, いわゆる「中食」も数多く含まれている

また, データに含まれる食事記録は, 単一の食事画像内に複数の食品が存在する場合が多い. このため, 各食品は個別のエントリとして識別され, 1 枚の画像に対して複数のエントリが対応することがある.

2.2 データ整理

データ整理の手順を表 1 に示す. データ整理の過程で, 各クラス内の全画像に対して CLIP 特徴量を抽出し, その平均ベクトルを算出した. 平均ベクトルから大きく逸脱する特徴量を持つエントリを「外れ値」と定義し, 手順 5, 6 において削除・修正の対象とした. さらに, 同一画像内で同じ料理が重複している画像も手順 7 における検証の対象とした. このような重複は, 画像認識モデルの誤識別やユーザーの誤入力によって生じる可能性が高く, CLIP 特徴量を用いた手法では検出できなかった誤りを発見するのに有効であった.

表 1: データ整理の手順と各手順におけるエントリ数とクラス数の推移

番号	手順	エントリ数	(差)	クラス数	(差)
0	(生データ)	189,794		9,546	
1	画像パスのないエントリの削除	154,238	(-35,555)	7,905	(-1,641)
2	バウンディングボックスの情報がないエントリの削除	132,702	(-21,536)	6,733	(-1,172)
3	バウンディングボックスのサイズが 0 のエントリの削除	132,693	(-9)	6,733	(-0)
4	クラス内エントリ数が 3 未満のエントリを削除	127,693	(-5,000)	2,680	(-4,053)
5	外れ値が多いユーザーのエントリをすべて削除	123,830	(-3,863)	2,663	(-17)
6	外れ値を確認し, 修正・削除	123,610	(-220)	2,663	(-0)
7	同一画像内に同じ料理が重複しているものを確認し, 修正	123,495	(-415)	2,663	(-0)
8	ラベルがおかしなものを修正・削除	123,472	(-23)	2,663	(-0)

2.3 クラスタリング・ラベリング

FoodLog Athl を通して集められたデータの各クラスラベルはユーザーが食事記録時に入力した食品名や商品名である. この状態では以下の問題が生じる.

- 「目玉焼き」「めだま焼き」などの表記ゆれ
- 「ご飯」「ご飯大盛り」など, 食品分類とは無関係な表現を含むラベル
- 「セブン-イレブン　たんぱく質が摂れるチキン&たまご」など, 商品名をそのまま使用したラベル

ここで, ユーザーが記入した食品ラベルに対してルールに基づく処理を行い, ラベルの修正やクラスの統合, クラスタリングを行う. まず, データセット内の各食品ラベルから食品メーカーの名称や分量・栄養価に関する記述を全て削除し, 食品の種類に関する情報のみをラベルに残す. その後, CLIP 特徴量を用いて食品クラスのクラスタリングを行う. クラスタリング手法には, 以下の 2 つが考えられる.

画像特徴量に基づくクラスタリング

画像分類タスクに適したクラスタリングが行えるが, 「豆腐」と「白米」のように見た目は似ているが食品分類上は大きく異なる食品を誤認識しやすい

食品ラベルのテキスト特徴量に基づくクラスタリング

食品部類を間違う可能性は低いが, 文字列としての近さなど食事とは関係のない情報が含まれてしまう

これらの方法には各々の長所, 短所があるため, 両者を統合してクラスタリングを行う. あるエントリに対して, バウンディングボックス内の画像特徴量, 食品ラベルのテキスト特徴量をそれぞれ F_{img}, F_{text} とし, ハイパーパラメータ α と正規化係数 n を用いて,

$$F = n\{\alpha F_{img} + (1-\alpha)F_{text}\}$$

で求められる総合的な特徴量 F を用いてクラスタリングを実施した. 予備実験の結果に基づき $\alpha = 0.7$ とし, k-means 法によるクラスタリングを行い, 1,921 種類の食品カテゴリを得た.

3 結果・展望

以上の提案手法に基づきデータの処理を行い, 最終的に 123,472 枚の食品画像を含み, 1,921 クラスもの詳細な分類が行われた食事画像データセット「FoodLog-1921」を構築した. また, データセット構築の要所において CLIP 特徴量を用いたルールに基づく手法を用いたため, 今後データが増えた時や他のデータセット構築の際にも応用が容易であると考えられる. 今後は高い表現力を有する CLIP モデルに対して, FoodLog-1921 を用いて追加学習を行い, 従来のものよりも詳細な食事画像の分類が行える分類器の開発を行う.

参考文献

[1] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang. Large scale visual food recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. 2021.

[3] K. Nakamoto, K. Kumazawa, H. Karasawa, S. Amano, Y. Yamakata, and K. Aizawa. Foodlog athl: Multimedia food recording platform for dietary guidance and food monitoring. In Proceedings of the 4th ACM International Conference on Multimedia in Asia (MMAsia), pp. 1–2, 2022.