6Q-03

Faster RCNN を用いた食事画像から料理部分の抽出モデルの構築

朱子宜† 戴 瑩†
岩手県立大学ソフトウェア情報学部†

1. はじめに

食は人間にとって非常に重要なものである。現在では、食材の入手が簡単で、現代人の食事生活も豊富になった。でも、この便利さとともに、他の問題が生じる。例えば、糖尿病と肥満症をかかることである。そして、毎日の食事記録によって、食の摂取を管理することが必要である。しかし、以前は記録の方法は的手録で不便であるため、自動的に記録する方法が求められる。

本研究では、Faster-RCNN に基づいて、食事画像から料理部分を検出するモデルを構築することによって、食事で使った食材を自動的識別し記録するのを目指す。本論文の課題として、料理部分を検出する時に利用する anchor box のサイズ と 学習率を grid search によってモデルのパラメターを調整する。その上で、５分割交差検証で検出モデルの性能を評価する。

2. 提案手法

2.1 システム概要

本研究では、食事画像から自動的に料理部分を抽出して料理食材認識を行うことを目指し、システムは二つのモデルから構築される。料理部分を検出するモデルと食材を識別するモデルである。本論文では主に食事画像から料理部分を検出するモデルの構築と性能評価を行う。

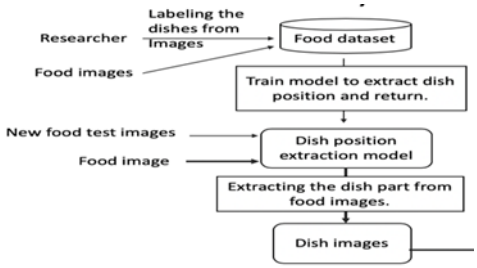

図１ モデルの構築 diagram

図１はモデルの構築 diagram である。まず、ネットで収集した食事画像に料理部分をレベル付け、学習データセットを作成する。そして、Faster RCNN に基づいて、転移学習により料理部分を検出するモデルを訓練することで、料理部分の位置座標を output する。

2.2 データセット

本研究では、レシピサイトで日常生活によく食べる 1000 枚の食事画像を収集してラベリング作業を行い、デッタセットを作成した。さらに、現存の研究と違って、料理部分のラベルつけは、今後の食材分析モデルの実現を考えるため、料理部分だけ含めるボックスをアノテーションする。最後はデータセットに K-fold 関数を利用して、分け率は８：２で学習セット、検証セット２つに分ける。その中に、料理種類が多い一方、形もいろいろある。大雑把に分けると、とうもろこし、パンのような規則ものと野菜炒めと煮物のような不規則もの二つに分ける。図２は食事画像の例である。

図２ 食事画像の例

3. 実装環境と実現方法

3.1 実装環境

本研究では、料理部分の検出モデルの構築は、pytorch フレームワークを利用する。操作環境は ubantu 18.4、GPU RTX 2080ti 11GB で実験を行った。

3.2 実現方法

Faster RCNN に基づく食事画像から料理部分の検出モデルを構築するために、Faster RCNN は CNN でデータセットの画像から特徴マップを生成する。そして領域提案ネットワーク (RPN)は特徴マップの局所領域毎に、物体らしさのスコアが付与された複数のバウンディングバックスを出力する。そして RoI (Region of interest) プーリング以降のネットワークへと進み、物体の分類が行われる。図 3 は Faster RCNN の構成である。

Dish detection from food image by Faster RCNN
Ziyi Zhu† and Ying Dai†
†Faculty of Software and Information Science, Iwate Prefectural University

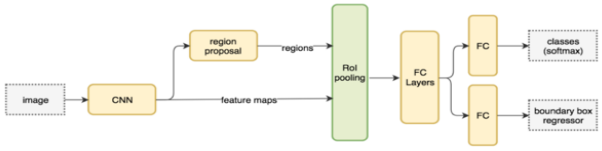

図３　Faster RCNN の構成

RPN は最初、特徴マップ上のアンカー毎に k 個のアンカーボックスを生成する。アンカーボックスはスケールとアスペクト比の組み合わせによる形状が決められる。本研究ではアンカーボックスのサイズの変化は三つに用意する。また、データセット全てのボックスのアスペクト比は統計により 0.7 から 1.3 までの範囲で分布することを分かった。図 4 は各ボックスのアスペクト比の分布である。そして本研究ではアンカーボックスのアスペクト比は「0.7, 1, 1.3」に設定する。さらにモデルの訓練段階では、学習率は(0.01, 0.001, 0.05)に設定する。

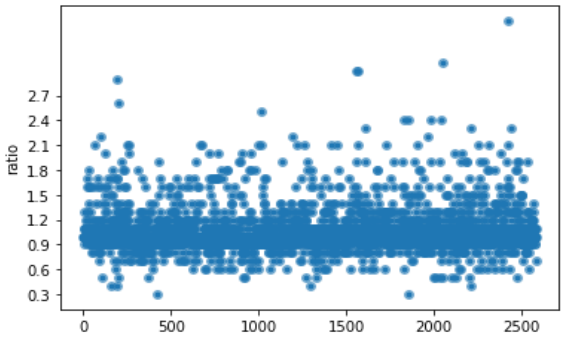

図４　ボックスのアスペクト比の分布

3.3 検出モデルの訓練

近年では ImageNet で事前学習済みモデルから特徴を転移して、新たなタスクで活用する方法が認識のパフォーマンスが大幅に向上した。本研究では、COCO で事前学習済み fasterrcnn_resnet50_fpn モデルを食事データセットで微調整する。最適なモデルを構築するために、anchor のスケールと学習率をパラメターにして手動的 grid search 調査実験を行う。結果により、モデルの性能は一番いいのが学習率は 0.005、アンカーサイズは(64, 128, 256)ときのモデルであった。また、そのモデルを訓練するとき 10 回左右で loss 値が収束したので、学習試行回数は 10 回に設定した。

4. 性能評価

検出モデルの性能は mAP(mean Average Precision)精度で評価する。mAP の計算式は以下の通りである。

$$mAP=\frac{1}{M}\sum_{i=1}^{M} APi \qquad (1)$$

ここで、M は食事画像の総数である。

$$AP=\frac{1}{N}\sum_{j=1}^{I}\frac{[num\_of\_Hit]j}{nj}$$
$$=\frac{1}{N}\sum_{j=1}^{I} Precision \qquad (2)$$

ここで、N は正しく認識した正解ラベルの総数、num_of_Hit は正解ラベルの中、正しく認識できた数、ni は正解ラベルの数である。

そして５分割交差検証で検出モデルの mAP 精度を測る。実験結果を表２に示す。IoU (intersection-over-union) は正解領域と予測領域の共通部分である。

表２　５分割交差検証の結果(%)

精度	f1	f2	f3	f4	f5
IoU=0.95	71	71.2	69.9	70.3	70.4
IoU=0.75	86.4	87.5	88.7	88.1	88.3

表２より、IoU=0.75 と IoU=0.95 の場合は、モデルの平均テスト精度がそれぞれ 87.8%と 70.56%である。新たな食事画像あらうまく料理部分を抽出することができて、構築したモデルの有効性を示した。検出された料理部分の例を図５で示す。

図５　検出例

また、本研究では、二つの現状が発見した。まず、もしデータセットのラベル付けは規則な食べ物と不規則な食べ物が一つの種類に設定した場合、モデルは不収束で学習をうまくできなかった。そして、データセットの画像が 1000 枚から増やすと検出精度があまり上がらないことである。

5. まとめ

本稿では、食事画像から理部分を抽出するモデルの構築を行った。統計結果により最適な学習率とアンカーボックスのアスペクト比とサイズを見つけた。そして、5 分割交差検証で検出モデルの平均テスト精度について、IoU=0.75 と IoU=0.95 の場合はそれぞれ 87.8%と 70.56%であった。今後は抽出した料理部分の画像から食材分析モデルを構築することを目指す。

6. 参考文献

[1] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497

[2] TORCHVISION OBJECT DETECTION FINETUNING TUTORIAL, https://pytorch.org/tutorials/intermediate/torchvision_tutorial.html