H-020

RethinkNet を用いた料理画像からの料理名と食材の推定
Estimation of food name and its ingredients from cooking image using RethinkNet

名高 祐輔† 青野 雅樹†
Yusuke Nadaka Masaki Aono

1. はじめに

近年は健康志向が向上しており，それに伴って食事面から健康管理を行うために，食事記録アプリケーションが増加している．しかしそれらはユーザが手動で料理情報を入力するものが多く，手間がかかるという問題がある．この問題を解消するために，料理画像から料理名や食材情報を自動認識する技術の需要が高まっている．また料理画像から料理や食材情報を自動認識する技術は，食事記録以外にも料理画像からのレシピ検索や，栄養素・カロリー推定など食事関連の様々なタスクに応用可能であるという点からも重要な技術といえる．

近年，画像認識の分野では Deep Convolutional Neural Network(CNN)の登場以来，画像認識の精度が飛躍的に向上しており，ILSVRC の 1000 種類分類タスクでは人の認識精度に匹敵する精度を達成している[2][3][4]．料理画像における画像認識でも CNN を用いたモデルが提案されており，従来手法の精度よりも向上している[5]．しかし，料理画像は同じクラスの料理でも使用している材料の種類や調理方法の違いにより外見も異なってくるため，料理分類タスクは一般的な画像認識よりも難しいタスクである．更に料理画像からの食材推定タスクに関しても食材の調理方法や用いられる料理の違いがあるため，料理分類タスク同様に困難であるといえる．したがって料理画像からの料理分類と食材推定の精度向上には，一般的な画像認識モデルを用いるだけではなく，料理と食材および食材同士の関係性を考慮したモデルを設計することが求められる．

本研究ではラベル相関を考慮できる RethinkNet を導入した深層学習モデルを提案する．データセットの観察の結果，料理と食材間および食材同士に関係性があることを発見したため，この関係性を用いることでより高精度な料理分類および食材推定が期待できる．

実験では，VireoFood172 データセットを用いて料理分類と食材推定の評価を行い，従来モデルとの比較を行う．2 章では料理画像からの料理分類や応用タスクに関する論文について述べる．3 章では従来モデルと提案モデルについて説明する．4 章では比較実験におけるモデルの学習方法や評価指標の説明と，実験結果とその考察を述べる．5 章では結論および今後の課題について述べる．

2. 関連研究

深層学習による料理画像の画像認識研究は近年盛んに研究されている．河野ら[5]は CNN を用いた料理画像の画像認識モデルを提案し，ハンドクラフト特徴量を用いた手法を上回る精度を達成した．Martinel ら[6]は料理画像から料理の層構造特徴を捉える Slice Network を提案し，提案ネットワークと一般画像認識で高い精度を誇る Wide Residual Network[7]を併用したネットワークである WISeR により UEC Food100，UEC Food 256，Food-101 の三種の料理画像データセットにおいて他の深層学習モデルを上回る精度を達成した．

† 豊橋技術科学大学, Toyohashi University of Technology

図 1 本研究で用いる料理画像と料理と食材ラベルの例

料理画像の画像認識の応用タスクとして，料理分類と食材推定のマルチタスク，料理分類とカロリー推定のマルチタスク，料理画像と調理レシピのクロスモーダル検索などが挙げられる．マルチタスク学習を行う CNN として Abrar ら[8]により Multi-task CNN が提案されており，これを利用して Chen ら[1]は料理分類と食材推定を同時に学習する VGG16 をベースとしたネットワークを提案し，それぞれのタスクを独立に学習した場合よりも精度が向上することを確認した．また，伊藤ら[9]は Chen らの提案したネットワークが単純な構造であることを指摘し，全結合層部分において各タスクのネットワークの全結合層の出力を他方のネットワークに入力する改良と，DenseNet[10]のスキップ結合を導入することで，Chen らの手法を上回る精度を達成した．柳井ら[11]は料理分類とカロリー推定のマルチタスクを同時に学習するネットワークを提案し，シングルタスクで学習した場合よりもマルチタスクで学習した場合のカロリー推定の精度が向上することを確認した．料理画像と調理レシピのクロスモーダル検索に関しては料理画像とレシピ情報の Joint Embedding を深層学習で学習することにより可能にしている[12][13].

3. 提案手法

我々はデータセットを観察した結果，特定の料理にはある特定の食材が多用されている，またある食材と別の食材は同時によく出現していることを発見した．つまり料理と食材の間，および食材同士には関係性があるということである．そこで先行研究において精度の良かった深層学習モデルにマルチラベル問題を考慮できる RethinkNet[14]を導入したモデルを提案した．この章では RethinkNet と先行研究の深層学習モデル，提案モデルについて説明する．

3.1 RethinkNet

RethinkNet は繰り返しマルチラベル予測を行うことで，入力に対するマルチラベル予測結果を改善していく手法である．この処理をモデル化するために，以前の予測結果で

ある内部状態を保持できる RNN を用いている．この RethinkNet は画像に対するマルチラベリングの既存手法である CNN-RNN[15]や Order-Free RNN[16]といった手法よりも高精度であることが比較実験で示されている．

次に，RethinkNet の構造について説明する．図 2 のように RethinkNet は RNN 層と全結合層からなる．RNN 層は入力された特徴ベクトル$\boldsymbol{x}$に対するマルチラベル分類器の役割を持ち，出力を全結合層と次ステップの RNN 層に渡す．全結合層は RNN 層の出力をラベルベクトル$\boldsymbol{y}$に変換する役割を持ち，RNN 層の各ステップの出力に対して変換を行い，それぞれをラベルベクトル$\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots \boldsymbol{y}^{(n)}$として出力する．この構造で RethinkNet はn回のステップ数だけ内部の RNN 層の処理を繰り返し，全結合層を通してマルチラベル予測を出力する．RethinkNet のステップ数は内部で行うマルチラベル予測の繰り返し回数を示している．ステップ数=1 のときは前ステップの予測情報が無いため，通常のマルチラベル予測に等しい．ステップ数が 2 以上の場合では，RNN 層に前ステップで予測された出力が特徴ベクトル$\boldsymbol{x}$と同時に入力されるため，他のラベル情報を考慮した予測がされる．ステップ数が増えるに連れ，以前の多くの予測結果を考慮できるため，最終ステップの予測では困難なラベルに対しても正確なラベル付けが期待される．そのためマルチラベル予測の評価時には，RethinkNet の最終ステップでの出力$\boldsymbol{y}^{(n)}$を用いる．

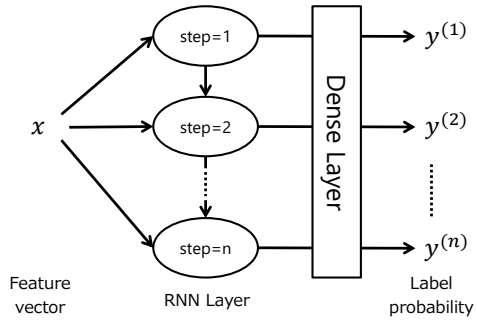

図 2　RethinkNet のモデル図

3.2 提案モデル

3.1 で説明した RethinkNet を先行研究の深層学習モデルである Arch-D に導入したものを提案モデルとする．本節では従来モデルと 3 つの提案モデルについて図 3 を用いて説明する．なお図 3 では提案モデルの提案部分を赤色で示してある．

3.2.1 Arch-D

まず先行研究の深層学習モデルである Arch-D について説明する．Chen らは料理画像からの料理名と食材の予測において VGG16 をベースにした深層学習モデルを複数考案しており，いずれも VGG16 の全結合層部分を改変したものであった．その内で最も高精度なモデルが図 3(a)に示す Arch-D であったため，本研究ではこの Arch-D を従来モデルとして比較実験および提案モデルのベースに用いる．

Arch-D は図 3(a)のように VGG16 の fc1 層の後に料理予測と食材予測の 2 つのネットワークに分岐させ，その後 1 つの全結合層を経由して料理もしくは食材の予測確率を出力している．料理側の全結合層の次元数は 4096，食材側の全結合層は 1024 である．

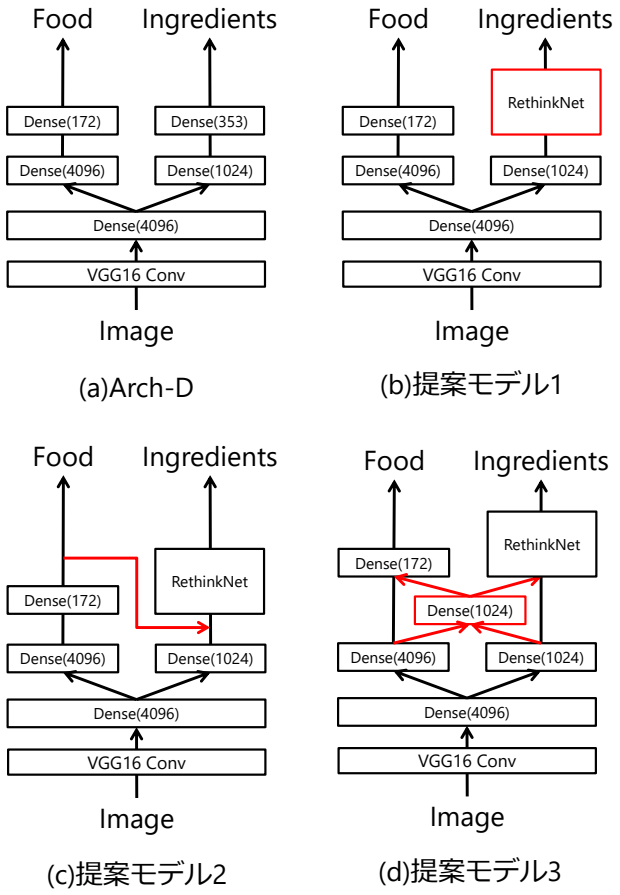

図 3　従来モデルおよび提案モデル図

3.2.2 提案モデル 1

提案モデル 1 を図 3(b)に示す．これは従来モデルの Arch-D モデルに 3.1 で説明した RethinkNet を導入した提案モデルである．マルチラベル予測を行う食材推定ネットワーク部分に RethinkNet を導入することで，マルチラベル問題である食材推定の精度向上を図っている．RNN 層には RethinkNet[14]の研究結果より LSTM を採用し，またステップ数は 5 回とした．

3.2.3 提案モデル 2

提案モデル 2 を図 3(c)に示す．これは 3.2.2 の提案モデル 1 の料理名予測の出力を食材推定ネットワーク内の RethinkNet に入力した新たな提案モデルである．先行研究から食材推定よりも料理名予測の方が高精度であることは確認されているため，高精度な料理名予測結果を食材予測に活用することで精度向上を図っている．

3.2.4 提案モデル 3

提案モデル 3 を図 3(d)に示す．これは 3.2.2 の提案モデル 1 の分岐したネットワークを交差するように全結合層を追加した提案モデルである．料理名予測と食材予測の両方に用いる特徴量を，全結合層を介して共有することで両方の予測の精度向上を狙っている．

4. 比較実験

提案モデルの有効性を確認するために，従来モデルとの比較実験を行った．

4.1 データセット

比較実験には先行研究でも用いられている VireoFood-172 を用いた．これは Chen らの研究で作成されたデータセットであり，中華料理の料理画像と料理名，使用食材のラベルからなるデータセットである．料理名クラス数は 172，各クラスに 100 枚以上の画像データが存在する．使用食材のラベル数は 353 で，1 枚の画像には平均 3 ラベル付与されている．また食材クラスは料理画像を見て分かるものが選択されている．総画像枚数は 110,241 枚で，訓練用データ，検証用データ，テスト用データから構成される．各データ数を表 1 に示す．

表 1　実験に用いたデータセット

訓練用データ	検証用データ	テスト用データ
66,071	11,016	33,514

4.2 学習方法

各モデルは ImageNet で学習済みの VGG16 のパラメータを初期値としてファインチューニングを行った．最適化手法には MomentumSGD を用い，学習率は 0.001 とした．学習はバッチサイズ 50 で 100 エポック行った．

モデルを学習する損失関数は，先行研究と同様に料理名予測に対する損失関数L_1と食材予測に対する損失関数L_2の重み付き和を用いた．これを以下の式(1)で表す．

$$L = -\frac{1}{N}\sum_{n=1}^{N}(L_1(n) + \lambda L_2(n)) \quad (1)$$

ここでNは全訓練画像枚数，λはL_1とL_2のバランスを調整する係数，nは入力画像のインデックスである．本研究では従来モデルと提案モデルでλの値を変えており，従来モデルでは先行研究の設定同様にλ =0.2 または 1，提案モデルではλ =1 としている．これは先行研究ではλ =0.2 で実験を行っているが，RethinkNet を用いて実験するときにはλ =1 の場合で結果が良かったことから，比較のために従来モデルは両方の値を用いて実験を行ったからである．

損失関数L_1を以下の式(2)に示す．

$$L_1(n) = \log(\hat{q}_{n,y}) \quad (2)$$

ここで$\hat{q}_{n,y}$は入力画像x_nに対する正解料理ラベルyの予測確率である．損失関数L_2を以下の式(3)に示す．

$$L_2(n) = \sum_{c=i}^{I}\left(p_{n,c}\log(\hat{p}_{n,c}) + (1-p_{n,c})\log(1-\hat{p}_{n,c})\right) \quad (3)$$

ここで$p_{n,c}$は入力画像x_nに対する材料 c の正解ラベルを表す二値変数である．$\hat{p}_{n,c}$は入力画像x_nに対する材料 c の予測確率である．またIは食材クラス集合である．

4.3 評価指標

評価指標は料理名分類には Accuracy を，食材推定には mAP と Micro-F1,Macro-F1 を用いた．Micro-F1 と Macro-F1 は式(4)，(5)，(6)，(7)から計算される$precision$と$recall$のマイクロ平均とマクロ平均を用いて式(8)から算出される．ここで，PRE_kは食材クラスkにおける$precision$，REC_kは食材クラスkにおける$recall$，Nは食材クラス数，TP_k，FP_k，FN_kはそれぞれ食材クラスkにおける真陽性，偽陽性，偽陰性のサンプル数である．

$$PRE_{micro} = \frac{\sum_{k=1}^{N}(TP_k)}{\sum_{k=1}^{N}(TP_k + FP_k)} \quad (4)$$

$$PRE_{macro} = \frac{\sum_{k=1}^{N}(PRE_k)}{N} \quad (5)$$

$$REC_{micro} = \frac{\sum_{k=1}^{N}(TP_k)}{\sum_{k=1}^{N}(TP_k + FN_k)} \quad (6)$$

$$RCE_{macro} = \frac{\sum_{k=1}^{N}(REC_k)}{N} \quad (7)$$

$$F1_l = 2\frac{PRE_l\, REC_l}{PRE_l + REC_l}(l = micro, macro) \quad (8)$$

また食材推定の評価を行う際は，しきい値 0.5 を設定し，食材推定の予測確率がしきい値以上である場合はその食材が使用されていると判別し，しきい値を下回る場合は使用されていないと判別する．判別の結果を二値変数として食材推定結果とし，評価を行う．

4.4 実験結果

先行研究モデルと提案モデルの推定精度の比較を表 2 に示す．表 2 を見ると，料理分類の Accuracy と食材推定の mAP と Micro-F1 において，提案モデルは先行研究モデルの精度を上回っており，提案モデルの有効性を確認できる．

特に料理名予測の精度が従来モデルよりも大きく向上しており，これは RethinkNet を用いて料理と食材間のラベルの関係性を学習したことが料理名予測で強く影響したためだと考えられる．

表 2　実験結果

手法	料理	食材		
	Acc	mAP	Micro-F1	Macro-F1
Arch-D(論文)	0.8206	-	0.6717	0.4718
Arch-D(再現, λ =0.2)	0.7672	0.3918	0.5352	0.3155
Arch-D(再現, λ =1)	0.7574	0.4370	0.5789	**0.4370**
提案モデル 1	0.7978	0.3866	0.5442	0.2600
提案モデル 2	**0.8017**	0.4349	0.5846	0.3038
提案モデル 3	0.7697	**0.4552**	**0.5963**	0.3542

次に VireoFood172 データセットの中で，従来モデルでは予測に失敗し，提案モデル 3 で予測に成功した具体例を図 4 に示す．図 4 は左上の料理画像がモデルへの入力画像であり，正解ラベルと各モデルの予測結果を表している．また各モデルの予測結果の内，太字で表しているのが正解している予測ラベルを示している．

図 4 の成功例 1 では，従来モデルの Arch-D では入力画像に対して料理名の予測は正しいが，食材の予測が全くできていない．これに対して提案モデル 3 では料理名予測が正解し，食材予測も余分なラベルを予測しているものの，正しいラベルの食材を全て予測できている．これは料理名と食材との関係や，食材同士の関係性を，RethinkNet を用いた提案モデル内で学習できたためだと考えられる．

図 4 の成功例 2 では，Arch-D は食材予測で正解ラベルである Noodles を予測できており，また画像内に写っている肉や野菜と同じようなラベルを予測できているが，料理名予測では Noodles とは関係のない料理名を予測している．

一方，提案モデル 3 では食材の正解ラベルも予測できており，かつ料理名予測も正しい予測ができている．これは提案モデル 3 の各予測に用いる特徴量を，中間の全結合層で共有したため，料理名と食材の関係性を学習できたためだと考えられる．

5. おわりに

本論文では，料理と食材間や，食材同士での関係性を考慮するために，ラベル相関を考慮できる RethinkNet を導入した料理名と食材推定の深層学習モデルを複数提案した．

VireoFood172 データセットを用いた比較実験の結果，料理名予測と食材推定の精度が従来モデルを上回り，提案モデルの有効性を確認した．このことから，予測モデルの出力である料理名ラベルや食材ラベル間の関係性を利用することが精度向上に寄与することが分かった．

今後の課題としては，今回の提案モデルでは食材推定の精度である Macro-F1 が従来モデルよりも低下したため，これを改善できる方法を検討することや，RethinkNet とは別の Attention などの手法を用いて，料理名と食材間などの関係性を学習できる深層学習モデルの考案，また食材ラベルはクラス数が不均衡である傾向が見られるため，これを考慮できる手法の考案，可能であれば他のデータセットを用いての実験などが挙げられる．

謝辞

本研究の一部は，科研費基盤（B）（課題番号 17H01746）の支援を受けて遂行した．

参考文献

[1] Chen, Jingjing, and Chong-Wah Ngo. "Deep-based ingredient recognition for cooking recipe retrieval." Proceedings of the 24th ACM international conference on Multimedia. ACM, (2016).
[2] Simonyan, Karen, and Andrew Zisserman."Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[3] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) .
[4] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. (2018).
[5] Kawano, Yoshiyuki, and Keiji Yanai."Food image recognition with deep convolutional features." Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. ACM, (2014).
[6] Martinel, Niki, Gian Luca Foresti, and Christian Micheloni. "Wide-slice residual networks for food recognition." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, (2018) .
[7] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," in Proc. of BMVC, (2016) .
[8] H. A. Abrar, W. Gang, L. Jiwen, and J. Kui. Multi-task CNN model for attribute prediction. IEEE Transactions on Multimedia, Vol. 17, No. 11, pp. 1949–1959, (2015).
[9] 伊藤晃洋, 山中高夫. "料理画像認識と料理材料推定の同時学習モデル (パターン認識・メディア理解)." 電子情報通信学会技術研究報告= IEICE technical report: 信学技報 117.514 (2018) .
[10] Huang, G., Liu, Z., Weinberger, K. Q., & van der Maaten,L. "Densely connected convolutional networks." Proceedingsof the IEEE conference on Computer Vision and Pattern Recognition. Vol. 1. No. 2. (2017) .
[11] Ege Takumi, and Keiji Yanai. "Multi-task learning of dish detection and calorie estimation." Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. ACM, (2018).
[12] Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).
[13] Carvalho, Micael, et al. "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings." The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, (2018) .
[14] Yao-Yuan Yang, Yi-An Lin, Hong-Min Chu, Hsuan-Tien Lin, “Deep Learning with a Rethinking Structure for Multi-label Classification”, arXiv:1802.01697 (2018) .
[15] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. CVPR, (2016).
[16] Chen, Shang-Fu, et al. "Order-free RNN with visual attention for multi-label classification." Thirty-Second AAAI Conference on Artificial Intelligence. (2018).

Ground Truth
- Food
  - Yu-Shiang Shredded Pork
- Ingredients
  - Julienned carrot
  - Black fungus
  - Shredded pork
  - Shredded bamboo shoots

Arch-D
- Food
  - **Yu-Shiang Shredded Pork**
- Ingredients
  - Minced green onion
  - Crushed pepper

提案モデル3
- Food
  - **Yu-Shiang Shredded Pork**
- Ingredients
  - **Julienned carrot**
  - Shredded pepper
  - **Black fungus**
  - **Shredded pork**
  - **Shredded bamboo shoots**

(成功例1)

Ground Truth
- Food
  - Beef noodles
- Ingredients
  - Green vegetables
  - Noodles
  - Beef chunks
  - Water

Arch-D
- Food
  - Saute Spicy Chicken
- Ingredients
  - Seared green onion'
  - Pepper slices
  - **Noodles**
  - Hob blocks of potato
  - Bullfrog
  - Chicken chunks

提案モデル3
- Food
  - **Beef noodles**
- Ingredients
  - Pepper slices
  - Chinese Parsleycoriander
  - **Green vegetables**
  - **Noodles**
  - **Beef chunks**
  - Hob blocks of potato
  - **Water**

(成功例2)

図 4　提案手法での成功例