Transformer Decoder を用いた料理画像からの料理名と食材の同時推定

名高　祐輔†　　青野　雅樹††

† 豊橋技術科学大学情報・知能工学専攻　〒 441–8580 愛知県豊橋市雲雀ヶ丘 1–1
†† 豊橋技術科学大学情報・知能工学系　〒 441–8580 愛知県豊橋市雲雀ヶ丘 1–1
E-mail: †nadaka@kde.tut.ac.jp, ††aono@tut.jp

あらまし　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．これらの記録を手動で行うのは手間がかかるため，料理画像認識による料理情報の自動認識技術の需要が高まっている．画像認識の分野では畳み込みニューラルネットの登場により認識精度が大きく向上しており，これを料理画像認識に用いた研究も多い．しかし，料理画像認識は一般画像認識と比較して難しい問題であり，更に使用されている食材推定は，マルチラベル問題となるため難しい問題である．本研究では料理画像からの料理名と食材の同時推定を行う．そのうちマルチラベル問題となる食材推定の精度を向上させるために，Transformer Decoder を用いた料理名と食材の同時推定モデルを提案する．評価実験では，先行研究で提案された深層学習モデルをベースラインとし，提案手法との比較実験を行った．その結果，提案モデルの有効性を確認することが出来た．

キーワード　料理画像認識，食材推定，マルチタスク CNN，深層学習

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．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．

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．

本研究では食材の関係性を考慮するために Transformer Decoder を導入した深層学習モデルを提案する．提案モデルの概要図を図 1 に示す．データセット観察の結果，料理と食材間および食材間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高精度な料理分類および食材推定が期待できる．

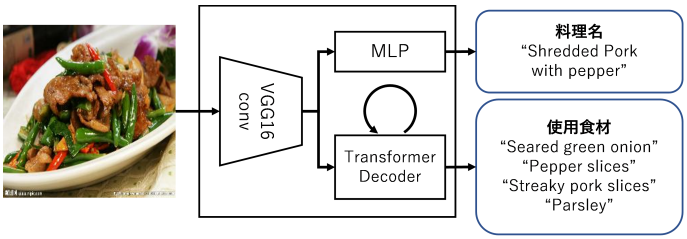

図 1: 提案モデル概要図

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 節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 節では従来モデルと提案モデルについて説明する．4 節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 節 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 [1] 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 [2] 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 を併用したネットワークである WISeR により UEC Food100 [3]，UEC Food256 [4]，Food-101 [5] 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料理名予測精度を達成した．

料理画像認識の応用タスクとして料理名と使用食材の同時推定，料理名とカロリーの同時推定，料理画像と調理レシピのクロスモーダル検索，料理画像からのレシピ生成などが挙げ

られる．マルチタスク学習を行う CNN として Abrar らにより Multi-task CNN が提案されており，これを利用して Chen ら [6] は料理名予測と食材予測を同時に学習する VGG16 [7] 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 [8] 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 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．Ege ら [9] は料理名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では，料理画像から得られる画像特徴量とレシピテキストから得られる文章特徴量との Joint Embedding を学習することで可能にしている [10] [11]．料理画像からのレシピ生成は Salvador ら [12] による研究があり，画像キャプションの自動生成技術を応用して料理画像のみから調理レシピの生成を行っている．

深層学習を用いた画像からのマルチラベル分類の研究としては，Wang ら [13] が提案した CNN とリカレントニューラルネットワーク（RNN）を組み合わせた画像とラベル間の関係性を学習するモデルなどがある．また Chen ら [14] は Wang らの提案したモデルと違い，学習時にラベルの順序を必要としない画像からのマルチラベル分類のモデルを提案した．Chen [15] らはグラフ畳み込みネットワークを用いた画像からのマルチラベル分類のモデルを提案した．

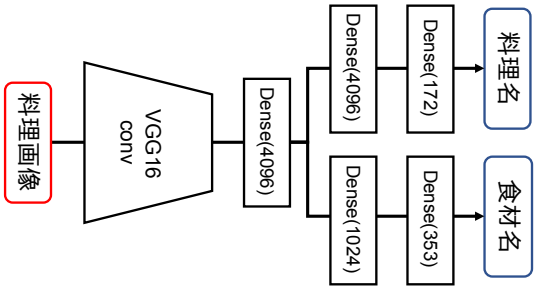

図 2: ベースラインモデル

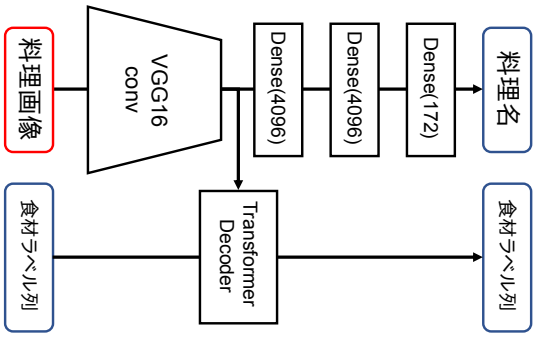

図 3: 提案モデル 1

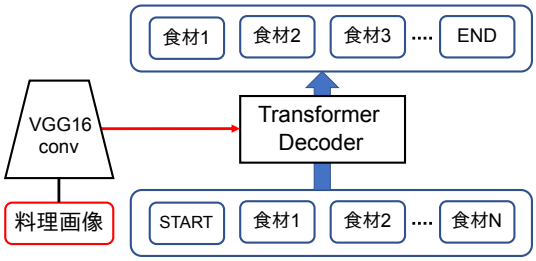

図 4: 訓練時の Transformer Decoder の図

3 提 案 手 法

提案手法は先行研究の深層学習モデルをベースに，Salvador ら [12] の研究で用いられた Transformer Decoder による食材推定ネットワークを導入する．先行研究のモデルではマルチラベル問題である食材推定部分において，食材ラベル間での関係性を考慮する構造は提案されていなかった．そこで食材推定部分に Transformer Decoder を導入することにより，食材ラベル間の関係性の学習をさせることで精度の向上が期待できる．以下よりベースラインモデルと Transformer Decoder の導入部分について示す．

3.1 ベースラインモデル

ベースラインには先行研究の Chen [6] らが提案した Arch-D モデルを用いる．これは CNN の 1 つである VGG16 をベースにしたモデルであり，全結合層部分においてネットワークを 2 つに分岐させて料理名予測と食材予測を同時に行う．ベースラインモデルを図 2 に示す．

3.2 Transformer Decoder の導入

提案モデル 1 としてベースラインモデルの食材推定部分に Transformer [16] の Decoder 部分を導入する．Inverse Cooking では料理画像から調理レシピを生成する過程で食材推定を行っており，Transformer Decoder で自己回帰的に食材推定を行っている．これを本研究にも導入する．食材ラベルを順序付けた食材ラベル列とみなし，Transformer Decoder で食材ラベル列を学習させることでラベル間の関係性を考慮する．この学習を行うにあたって教師データとして正解の食材ラベル列の順序が必要となるが，元々がマルチラベル問題であるため正解の順序は存在しない．そこで本研究では訓練データにおける食材ラベルの頻度を集計し，その降順，昇順，ランダム順序を用いてモデルの訓練を行った．ベースラインモデルに Transformer Decoder を導入したモデルの全体図を図 3 に示す．

訓練時の様子を図 4 に示す．訓練時には料理画像と START から始まる食材ラベル列をモデルに入力する．料理画像は CNN に入力することで画像特徴量に変換され，料理名予測として用いられると同時に Transformer Decoder 内で単語ベクトルとの Attention にも用いられる．教師データとして料理名ラベルと末尾に END を持つ食材ラベル列を用いる．START と END は食材ラベル列の始まりと終わりを示すクラスであり，これに伴って Transformer Decoder が持つ Embedding 空間の語彙数も拡張する必要がある．

Transformer Decoder の詳細な構成を図 5 に示す．下部か

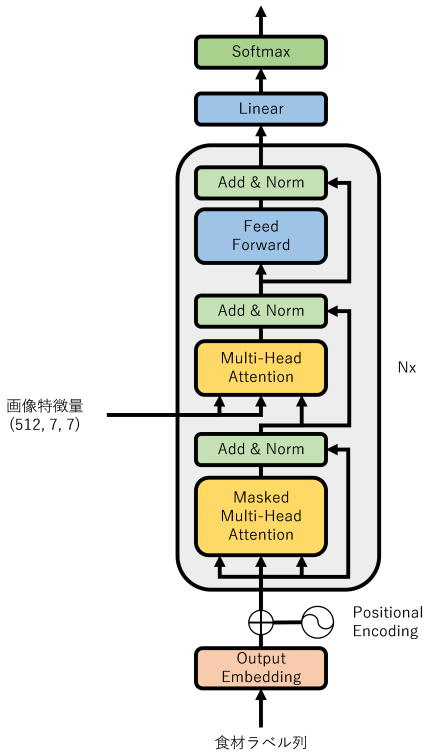

図 5: Transformer Decoder の構成図

ら食材ラベル列が入力されると Embedding 層により単語ベクトル化される．単語ベクトルは Self-Attention の後，料理画像の画像特徴量との Attention が行われる．Attention に用いる画像特徴量は，VGG16 の最終畳み込みブロックの出力を用いる．その後，位置毎の単語に対してのフィードフォワードネットワークを通る．Self-Attention からフィードフォワードネットワークまでの一連の処理はまとめて一つのブロックとして扱われており，ブロックの数だけ同様の処理が行われる．最後に Softmax 活性化関数を通して，各位置の食材ラベルに対して食材ラベルの多クラス分類が行われる．

推論時の様子を図 6 に示す．図中の TD は Transformer Decoder を示している．推論時は料理画像と START のみをモデルに入力する．食材推定時は最初に Transformer Decoder に START のみを入力し，出力された食材ラベルから自己回帰的に食材ラベル列の出力を行う．推論は END を出力するまで繰り返す．画像特徴量は推論の毎ステップ入力される．

3.3 推論時の探索手法

推論時は Transformer Decoder を用いて自己回帰的に食材ラベル列を出力する．自己回帰的に推論を行う手法は貪欲法とビームサーチの 2 つが存在する．本研究では両方の手法を試したため，以下に具体的なアルゴリズムの説明を示す．

3.3.1 貪 欲 法

Transformer Decoder の出力は各食材ラベルの確率値である．貪欲法では推論ステップで常に最大確率の食材ラベルのみを予測食材ラベルとして扱う．貪欲法の欠点として，初期に間違ったラベル予測を行った場合，後続の予測は間違ったラベル

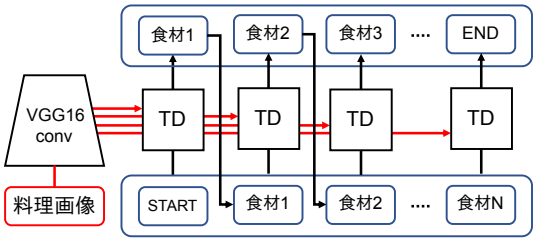

図 6: 推論時の Transformer Decoder の図

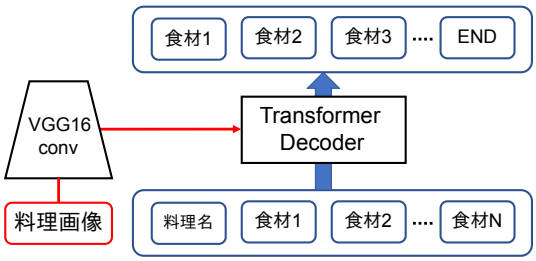

図 7: 提案手法 2 の訓練時の図

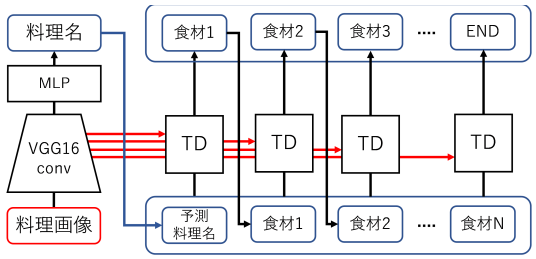

図 8: 提案手法 2 の推論時の図

の影響を受けるという問題がある．

3.3.2 ビームサーチ

ビームサーチは貪欲法の欠点を補う探索手法である．ビームサーチでは各推論ステップにおいて複数の食材ラベル候補を保持し続ける．この候補の数をビーム幅と呼ぶ．推論ステップでは前ステップの候補すべてを用いて推論を行い，各食材ラベルの確率を取得する．得られた確率と候補の食材ラベル確率から現在ステップまで予測した食材ラベル列の平均確率を算出してランキングを行い，上位の食材ラベルを候補として次ステップの推論に移る．

3.4 料理名を考慮した食材推定ネットワーク

提案手法 2 として食材推定部分で学習させる食材ラベル列の初期ラベルとして料理名を使用する手法を提案する．提案手法 2 の訓練時の様子を図 7 に示す．比較的高精度である料理名予測の情報を食材予測に用いることで，食材予測の精度向上を期待できる．

提案手法 2 の推論では，食材ラベル列の初期単語として料理名部分で予測された料理名情報を用いる．提案手法 2 の推論時の様子を図 8 に示す．その際に料理名を初期単語としてどのような単語ベクトルとするかで更に 2 つの手法を考案した．1 つ目は料理名部分で予測された最大確率の料理名の単語ベクトルのみを用いる手法である．これは最大確率の料理名のみの情報を用いているため，料理名予測が正しくない場合は食材ラベ

ル列予測に悪影響を及ぼすと考えられる．2 つ目は料理名部分で予測された各料理名クラスの確率で，それぞれの料理名の単語ベクトルを重み付けし，その総和を初期単語として用いる手法である．これは様々な料理名情報が入ってしまっている一方で，料理名予測で最大確率のものが正しくない場合でも，2 番目に高い料理名ラベルがあっている場合はその単語ベクトルの情報もある程度含んだ初期単語が生成できるため，料理名予測が失敗してもある程度食材ラベル列の予測精度を高める効果が期待できる．本研究では 1 つ目の料理名単語ベクトル生成手法を提案手法 2A，2 つ目を提案手法 2B と定義する．

4 比較実験

提案モデルの有効性を確認するために，ベースラインとの比較実験を行った．

4.1 評価用データセット

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数

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 評価方法

評価方法は料理名予測には Accuracy を，食材予測には Macro-F1, Micro-F1 を用いた．Macro-F1 と Micro-F1 は式 (1)，(2)，(3)，(4) から計算される presicion と recall のマイクロ平均とマクロ平均を用いて式 (5) 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．

$$PRE_{micro} = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k + FP_k} \tag{1}$$

$$PRE_{macro} = \frac{\sum_{k=1}^{N} PRE_k}{N} \tag{2}$$

$$REC_{micro} = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k + FN_k} \tag{3}$$

$$REC_{macro} = \frac{\sum_{k=1}^{N} REC_k}{N} \tag{4}$$

$$F1_l = 2 \cdot \frac{PRE_l \cdot REC_l}{PRE_l + REC_l} (l = micro, macro) \tag{5}$$

また食材推定の評価を行う際は，モデルが出力した食材ラベル列の内，END が出力されるまでの食材ラベルを食材推定結果として扱い，評価を行う．

4.3 実験設定

実験設定はベースラインモデルの訓練のものと同様とした．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 0.01 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，料理名予測と食材予測のクロスエントロピーの和を用いた．これを以下の式で表す．

$$L = -\frac{1}{M} \sum_{n=1}^{M} L_{food}(n) + L_{ingr}(n) \tag{6}$$

ただし，M は全料理画像枚数である．損失関数の詳細な説明を以下に述べる．料理名予測の損失関数は式 (7) で定義される．

$$L_{food}(n) = \log p_{c,n} \tag{7}$$

なお，$p_{c,n}$ は料理画像 x_n の持つ正解料理名クラス c の予測確率である．食材予測の損失関数は式 (8) で定義される．

$$L_{ingr}(n) = \sum_{t=1}^{L} \log p_{c,n}^{t} \tag{8}$$

なお，L は食材ラベル列長，$p_{c,n}^{t}$ は料理画像 x_n の持つ正解食材ラベル列の t 番目の食材ラベル c の予測確率である．

また Transformer Decoder のパラメータは Salvador ら [12] の研究と同様のものを用いた．ブロック数は 4，マルチヘッドアテンション数は 2，Embedding 次元数は 512 である．

4.4 実験結果

各手法による実験結果を示す．実験結果の表で各評価尺度で最良の値を太字で示している．表 2 に評価実験の結果を示す．ベースラインモデルの λ は学習時の食材予測の損失関数の重み係数を示す．

ベースラインモデルと提案手法モデルの精度を比較すると，いずれの評価尺度においても提案モデルがベースラインモデルの精度を上回っていることを確認できる．料理分類の Accuracy においてはベースラインモデルの論文内の精度が最も良い値だが，再現実験した結果と比較すると提案手法 1 が最も良い精度である．食材推定のいずれの評価尺度においては，ベースラインモデルの論文の値と比較して提案モデルの精度は大きく上回っている．

次に提案手法同士の結果を比較を行う．貪欲法とビームサーチの精度を比較すると，ビームサーチの方が高精度であることが確認できる．これは貪欲法の欠点をカバーした手法がビームサーチであるため，精度が向上したと考えられる．また提案手法 1 と提案手法 2 を比較すると，食材推定の精度は提案手法 2 が上回っている一方で，料理名推定の精度は少し低下している．

表 2: 実験結果

	サーチ手法			食材ラベル列順序			料理予測	食材予測	
model	貪欲法	ビーム幅 2	ビーム幅 3	昇順	降順	ランダム	accuracy	macro-f1	micro-f1
Arch-D(論文)							0.8206	0.4718	0.6717
Arch-D(再現, $\lambda = 0.2$)							0.7988	0.3960	0.6096
Arch-D(再現, $\lambda = 1.0$)							0.7896	0.4708	0.6271
	✓			✓			**0.8154**	0.5002	0.6510
		✓		✓			**0.8154**	0.5026	0.6791
			✓	✓			**0.8154**	0.4974	0.6727
	✓				✓		0.8124	0.4495	0.6561
提案 1		✓			✓		0.8124	0.5157	0.6773
			✓		✓		0.8124	0.5088	0.6730
	✓					✓	0.8120	0.4478	0.6358
		✓				✓	0.8120	0.5082	0.6779
			✓			✓	0.8120	0.5023	0.6720
	✓			✓			0.7942	0.5244	0.6404
		✓		✓			0.7942	0.5498	0.7126
			✓	✓			0.7942	0.5482	0.7082
	✓				✓		0.7980	0.3155	0.5279
提案 2A		✓			✓		0.7980	0.5518	0.7156
			✓		✓		0.7980	0.5436	0.7107
	✓					✓	0.7992	0.3737	0.5534
		✓				✓	0.7992	0.5354	0.7151
			✓			✓	0.7992	0.5303	0.7083
	✓			✓			0.7942	0.4232	0.3804
		✓		✓			0.7942	0.5505	0.7133
			✓	✓			0.7942	0.5487	0.7078
	✓				✓		0.7980	0.1233	0.3798
提案 2B		✓			✓		0.7980	**0.5522**	**0.7175**
			✓		✓		0.7980	0.5446	0.7135
	✓					✓	0.7992	0.2466	0.3718
		✓				✓	0.7992	0.5368	0.7154
			✓			✓	0.7992	0.5322	0.7092

提案手法 2A と 2B を比較すると 2B の方が僅かに食材推定の精度が向上しているものの，大きな違いは無かったといえる．食材ラベル列の頻度昇順，降順で精度を比較すると，貪欲法を用いた場合は昇順だと Macro-F1 の精度が向上し，また降順だと Micro-F1 の精度が向上していることが確認できる．Macro-F1 は食材クラス平均の評価尺度であり，低頻度の食材クラスの精度の影響を受けやすく，また Micro-F1 は全体で評価を行うため頻度数の多い食材クラスの影響を受けやすい．そのため昇順だと低頻度の食材クラスから予測を始めるために低頻度の食材クラスの精度が向上して Macro-F1 が向上し，降順だと逆に高頻度の食材クラスの精度が向上して Micro-F1 が向上したと考えられる．ただし，ビームサーチを用いると食材ラベル列の順序に関係なく精度が同程度まで向上しているため，ビームサーチを用いる場合においては順序は精度に大きく影響しないといえる．

5 おわりに

本研究では，Transformer Decoder を用いることで食材ラベル間の関係性を考慮できる学習モデルと，料理名予測の情報を Transformer Decoder の入力に用いて料理名と食材ラベルの関係性を考慮するモデルを提案した．先行研究の深層学習モデルをベースラインとおいて比較実験を行った結果，提案モデルの精度はベースラインモデルを上回ったことを確認できた．

本研究の課題点としては，学習する際の食材ラベル列の順序をどのように決定するかという問題がある．本研究では食材ラベルの出現頻度の昇順，降順，ランダムで実験を行ったが，ビームサーチを用いた場合に実験結果は大きくは変動しなかった．一方で多少の精度の差は見られたため，最適な順序で学習することでより高い精度を達成できると考えられる．しかし，最適な順序を求めるには考えられる順序すべてで学習して評価を行う必要がある．この順序の問題を解決するためには，学習時に順序を必要としない手法 [14] や，グラフ畳み込みネットワークを用いた手法 [15] などを用いる必要があると考えられる．また，他のデータセットでも有効性が見られるかを検証する必要がある．

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．

文　　献

[1] Yoshiyuki Kawano and Keiji Yanai. Food image recognition with deep convolutional features. pp. 589–593, 09 2014.

[2] N. Martinel, G. L. Foresti, and C. Micheloni. Wide-slice residual networks for food recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 567–576, March 2018.

[3] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2012.

[4] Y. Kawano and K. Yanai. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[6] Jingjing Chen and Chong-wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pp. 32–41, New York, NY, USA, 2016. ACM.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[8] 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. 信学技報, 第 117 巻 of *BioX2017-38, PRMU2017-174*, pp. 13–18, 東京, 3 月 2018. 2018 年 3 月 18 日 (日)-3 月 19 日 (月) 青山学院大学青山キャンパス (PRMU, BioX).

[9] T. Ege and K. Yanai. Simultaneous estimation of food categories and calories with multi-task cnn. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pp. 198–201, May 2017.

[10] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[11] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 35–44, New York, NY, USA, 2018. ACM.

[12] Amaia Salvador, Michal Drozdzal, Xavier Giro-i Nieto, and Adriana Romero. Inverse cooking: Recipe generation from food images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2285–2294, June 2016.

[14] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Order-free rnn with visual attention for multi-label classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[15] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5177–5186, 2019.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017.