Vol.2021-NL-248 No.21
Vol.2021-CVIM-226 No.21
2021/5/21

食事画像に対する 少数およびゼロショット領域分割

本部 勇真[1,a)]　柳井 啓司[1,b)]

概要：健康管理アプリケーションが流行し，食事管理の意識が高まっている．料理のカロリー量計算をする際には食事領域の判別が大事な要素である．しかし，深層学習を用いる際，学習には大量のデータが必要となり，無数に存在する食事カテゴリのデータ収集は非実用的であるといえる．近年では，少数学習データを用いて領域分割モデルを学習する Few-shot Segmentation という方法が研究されている．本研究では，食事ドメインの画像をターゲットとした Few-shot 及び Zero-shot Segmentation を適用することで，食事学習データの量の不十分さを解消し，新たな食事クラスに対する領域分割の有効性を示す．また，従来の Few-shot Segmentation モデルに CookPad のレシピで学習した word2vec による単語埋め込み加えた新しい手法を提案し，従来手法よりも精度が向上する結果となった．

1. はじめに

近年のセグメンテーションタスクでは，CNN ベースのモデルによって，セグメンテーションの性能を大幅に進歩させている．既存の領域分割食事データセットとして 102 種類のカテゴリ，合計 1 万枚で構成される UEC-FoodPix Complete [1] がある．しかし，無数に存在する食事カテゴリに対して，分類する新規クラスの学習データを大量に必要とする従来の深層学習モデルにはデータ量が不十分であるといえる．近年研究されている Few-shot Segmentation は，ターゲットドメインクラスに関しての大量の学習画像を使用できる場合，数枚のサポートセット画像の情報を用いることで，未学習のクラスを正しくセグメンテーションすることを目的としている．そのため，データ量の問題を解決するとともに，既存データセットを少量の追加データで拡張することができると考えられる．Few-shot 及び Zero-shot Segmentation タスクでは，学習データと検証データおよびテストデータ間に共通カテゴリは存在しないため，学習データとテストデータの分布の違いが大きいデータセットでは，推論の際に学習クラスに偏りのある誤った領域を推論する場合があり通常の Segmentation タスクと比較すると難解なタスクである．そのため，モデル学習の手法やサポートセットの扱い方が重要なタスクとなっている．

本研究では，学習と検証カテゴリ間の分布変化が小さいと考えられる食事ドメインの画像をターゲットとして Few-shot，Zero-shot Segmentation を適用することで，新たな食事クラスに対するセグメンテーションの有効性を示し，食事学習データセットの量の不十分さを解消すると共に，Zero-shot タスクで使用されている単語埋め込みの手法を組み込んだ新たな Few-shot Segmentation モデルの提案をする．

2. 関連研究

2.1 問題設定

Few-shot Segmentation では少数学習データによる領域分割をタスクとしている．学習時と検証時のカテゴリには共通部分が存在しない．そのため，検証時の入力には未知のカテゴリのクエリ画像と，同カテゴリのサポート画像と，そのマスク画像がサポートセットとして与えられる．また，単語埋め込みの特徴量や，事前学習済みの特徴量同士の類似度を用いて，未知のカテゴリを領域分割する Zero-shot Segmentation というタスクも存在する．

3. Few-shot Learning

Few-shot Learning とは，学習データには存在しない未知のカテゴリの教師データが少ない場合に，事前に既存クラスで学習した知識と少数の同一ドメインデータを用いて，

[1] 電気通信大学 大学院情報理工学研究科 情報学専攻
[a)] honbu-y@mm.inf.uec.ac.jp
[b)] yanai@mm.inf.uec.ac.jp

© 2021 Information Processing Society of Japan

未知クラスに対してカテゴリや物体の識別知識を学習することを目的としている．Few-shot タスクのモデルの学習には，主にメタ学習，メトリック学習，Data Augmentation 手法の 3 つの学習方法が存在する．メタ学習では，未知の少数学習データで fine-tune し，その重みを用いてモデルのパラメータを学習し，未知のカテゴリに対処する学習法である．メトリック学習は，既存学習データの特徴空間を学習し，未知の新規データが与えられた際に，学習した特徴空間の距離や類似度をもとにオブジェクトの識別知識を学習する手法である．Data Augmentation 手法としては少数教師データを様々な手法で増強し，少数データをよりよく扱う手法である．

メトリック学習の関連研究例として Relation Network [2] は，特徴抽出をする Embedding Module と特徴同士の類似度を計算する Relation module が存在し，その中でクエリ画像とサポート画像の特徴をチャンネル方向に連結し，畳み込むことで画像間の特徴空間の深層的な距離関係を抽出できることを示した．Li ら [3] は Relation Network を，Few-shot Segmentation において，Unet [4] に追加することで，その有効性を証明している．本研究の学習手法としてメトリック学習を使用しており，サポート画像とクエリ画像の融合方法は Relation Network を参考に使用している．

3.1 Few-shot Segmentation

Few-shot Segmentation とは Few-shot learning をピクセルレベルの分類に拡張したタスクである．近年の Few-shot Segmentation の学習の多くはメトリック学習が使用されている．Shaban ら [5] は，メトリック学習である Siamese Net [6] のアーキテクチャをもとに, サポートブランチとして Conditioning Branch，クエリブランチとして Segmentation Branch の 2 分岐ブランチ構造を用いたネットワークを実装し，最初に One-shot Segmentation を実現した．サポートブランチでは，ターゲットと同カテゴリの画像とマスク画像を使用クエリブランチに対して条件付き重み付けをし，クエリのマスクを予測する情報を伝播する役割を持っている．本研究では，2 分岐ブランチに単語処理ブランチを加えた 3 分岐ブランチ構造を使用している．

Sg-One [7] では，サポート画像のターゲット部分の特徴に対してサポートマスクでマスキングしその領域で平均を取る Masked Average Pooling(MAP) の処理を行うことで画像のターゲット部分の大域的な特徴ベクトルを抽出し，クエリ画像の特徴マップとのコサイン類似度を用いることによって対象領域を効果的にを活性化している．CANet [8] では，サポート特徴とクエリ特徴にコサイン類似度を用いるのではなく，MAP を施したサポートベクトルをクエリ特徴の全領域に連結し，畳み込みを行うことでクエリ画像の空間的位置を，サポート画像の大域的な特徴ベクトルと比較することを可能とし，クエリとサポートのターゲット間の位置不一致問題に対処している．また，Few-shot Segmentation では未学習のカテゴリに適応する必要があり，backbone の高次元の特徴を扱うと学習クラスにオーバーフィットしてしまう問題がある．そのため，Zhang [8] らは，学習中でも backbone を固定することや，ResNet-50 において，backbone の 4 つの Layer の内，オブジェクト部分を構成する 1024 次元特徴と 512 次元特徴の出力を持つ Layer3 との Layer2 の中間特徴を組み合わせたものを使用することによってオーバーフィットを回避し精度を向上させている．

Tian らの提案した Prior Guided Feature Enrichment Network (PFENet) [9] では，高次元特徴量には，意味的情報が多く含まれていることを利用して，重みを固定した backbone から抽出したサポート画像の Masked Average Pooling (MAP) ベクトルとクエリ画像の高次元特徴量同士の類似度で対象領域を推定する学習に依存しない Prior Mask を生成し，特徴マップに連結することで対象物の空間的な位置情報を与えている．また，空間的解像度を増強する手法として，局所特徴から大域的な特徴マップを順に畳み込ことによってスケール間の階層関係を考慮した畳み込みを実現した Feature Enrichment Module (FEM) を提案している．また，Zero-shot モデルも提案しており，FEM で使用されるサポート画像の対象領域で平均を取った MAP ベクトルを 1×1 convolution と Relu で変換した単語ベクトルに置換している．

本研究では，PFENet をベースラインとして使用する．理由としては，使用する UEC-FoodPix Complete [1] の画像には複数の食品が存在する場合があり, 対象領域の空間的位置が重要であると考えられることや，Pascal-5^i, MSCOCO-25^i を使用した実験において PFENet は Few-shot Segmentation タスクで最高精度を達成しているためである．従来手法の比較は表 1，表 2，FEM のアーキテクチャの詳細は図 1 に示す．

3.2 Zero-shot Segmentation

Zero-shot タスクでは埋め込んだ単語を視覚的特徴量に変換する手法が存在する．Kato ら [10] の研究では，入力画像の特徴を抽出するブランチと，埋め込み単語を VAE (Variational Auto Encoder) [11] で処理するブランチからなるネットワークを提案している．単語処理ブランチでは，埋め込み単語の意味空間から視覚空間への変換に VAE を用いて，変分写像を行うことによって潜在変数を視覚特徴ベクトル

© 2021 Information Processing Society of Japan

Vol.2021-NL-248 No.21
Vol.2021-CVIM-226 No.21
2021/5/21

表 1 Pascal5^i データセットにおける One-shot Segmentation タスク手法の比較

	VGG-16 backbone				
	Fold-0	Fold-1	Fold-2	Fold-3	Mean
OSLSM [5]	0.336	0.553	0.409	0.335	0.408
SG-One [7]	0.402	0.584	0.484	0.384	0.464
PFENet [9]	**0.569**	**0.682**	**0.544**	**0.524**	**0.580**
	ResNet-50 backbone				
CANet [8]	0.525	0.659	0.513	0.519	0.554
PFENet [9]	**0.617**	**0.695**	**0.554**	**0.563**	**0.607**

表 2 Pascal5^i データセットにおける Zero-shot Segmentation タスク手法の比較

	VGG-16 backbone				
	Fold-0	Fold-1	Fold-2	Fold-3	Mean
Kato ら [10]	0.396	0.526	0.410	0.356	0.422
PFENet [9]	**0.500**	**0.685**	**0.517**	**0.466**	**0.542**

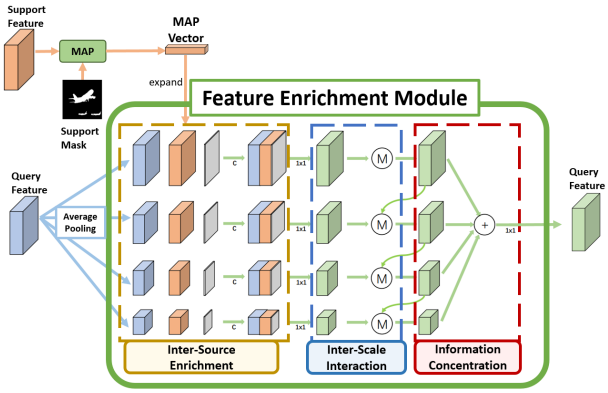

図 1 Feature Enrichment Module のアーキテクチャ．([9] から引用)

として扱っている．VAE では，埋め込まれた単語ベクトルを平均ベクトルと分散ベクトルに線形変換し，正規分布からサンプリングして，視覚空間の意味的特徴量としている．また，Kato らの実験では，変換した特徴と入力画像の視覚特徴の融合方法の比較実験を行っており，変換した単語特徴ベクトルを入力画像の視覚特徴量のチャンネル方向に連結し，畳み込むモデルが一番良い精度を実現している．本研究でも同様に，埋め込みベクトルを視覚特徴に変換したものをクエリの特徴に連結する手法及び，VAE を用いた単語ベクトル再生成手法を実験に使用した．

Bucher らの提案するネットワークの ZS3Net [12] では，単語埋め込みで表現される名前固有のベクトルが，多くの意味的属性を保持していることを利用している．つまり，同じドメインの名前の単語は意味空間内で近くに留まる可能性が高くなり，新規カテゴリの単語埋め込みベクトルからそのカテゴリの視覚特徴量を生成することを目的としたネットワークを提案している．

ZS3Net のアーキテクチャでは，埋め込まれた単語は，

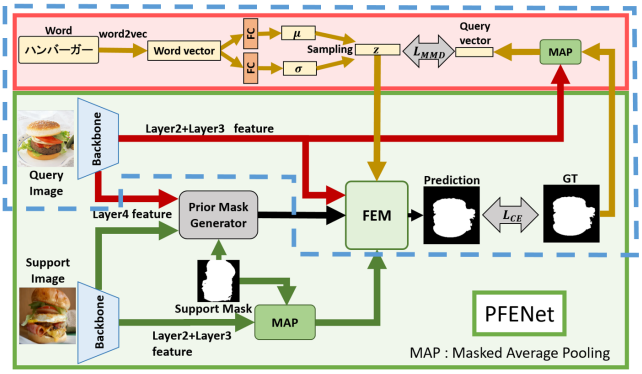

図 2 Few-shot タスクに対して本研究で提案するネットワーク図．青い点線は Zero-shot タスクに対応する部分．

ターゲットカテゴリの単語ベクトルはガウス分布に従うノイズベクトルが連結されることで，多様な表現を促すベクトルとして扱われる．これを Generator によってピクセル単位の視覚特徴を生成する． Generator は，Generative Moment Matching Network(GMMN) [13] に基づいて構成されている．GMMN は，1 つの隠れ層を持つ多層パーセプトロンであり，Leaky Relu，Dropout 層を持ち，入力には単語ベクトルに同サイズのノイズベクトルが連結されたベクトルが使用される．この Generator によって未知のカテゴリの視覚特徴を生成し既存クラスの特徴と未知クラスの生成特徴を分類することによって領域分割を可能にしている．また，損失関数には，Maximum Mean Discrepancy(MMD) を使用した，ターゲットの特徴ベクトル分布と単語の再生成ベクトルの分布間の違いを小さくする損失関数が用いられていることによって，視覚的特徴への変換を促している．本研究では，MMD による損失関数と単語ベクトル再生成手法の GMMN を実験に使用する．

4. 手法

学習カテゴリと検証カテゴリ間の分布シフトは，通常の一般的な物体ではカテゴリ間の大きく，見た目や形状の変化が大きなものが多い．そのため数枚のサポートセットのみでは，正確な領域分割を予測することは困難であることが知られている．そのため本研究では，一般物体とは異なる特性を持ち，見た目や形状が近いものが多い食事画像を使用することで，データを食事ドメインに限定し，カテゴリ間の分布シフトに対処する．また，サポートセットに加えて，単語埋め込みの特徴も加えることによって検証カテゴリに対する情報を強化させることで，対象領域と無関係な領域がより識別的に扱える新たな Few-shot モデルを提案する．また，単語埋め込みを用いた Zero-shot モデルを提案する．提案ネットワークの詳細図は図 2 に示す．

© 2021 Information Processing Society of Japan

4.1 単語埋め込み手法

単語埋め込みモデルとして word2vec [14] を使用する．日本語 wiki で学習された word2vec モデルでは，今回使用するデータセットである岡本 [1] らの作成した UEC-FoodPix Complete には存在しない単語があるため使用はできない．そのため，NII で公開されているクックパッドデータセットに含まれる約 16 万のレシピテキストを利用して，料理単語の word2vec を学習した．レシピテキストを利用した理由としては料理の食材単語を使用することで同じ食材を使う料理が表現空間上の近くに埋め込まれる．料理の見た目は調味料や食材の色に左右されることが多く，食材単語の埋め込みを生成することは，学習データとテストデータに同一のドメイン部分が存在する場合に効果的であると考えたためである．これを Few-shot と Zero-shot モデルに組み込む手法を提案する．

4.2 Few-shot 手法

Few-shot の提案手法のアーキテクチャの詳細は，図 2 の通りである．本研究では, PFENet [9] のネットワークをベースラインとして使用している．Backbone は Food-101 [15] で事前学習された ResNet50 [16] を特徴抽出器として用いる．ここでは通常の ResNet50 とは異なり dilation, padding, stride を調節し layer2,3,4 の出力サイズは元の 1/8 である．PFENet と同様に layer4 で出力される高次元特徴を用いて Prior Mask を生成する．クエリの特徴は，layer2,3 の特徴を連結させ，1 × 1 Convolution と Relu 関数で 256 次元に圧縮したものをクエリ特徴とする．同じくサポートの特徴を圧縮し，Masked Average Pooling (MAP) を行い得られた MAP ベクトルをサポートベクトルとする．これら Prior Mask，サポートベクトル，クエリ特徴を FEM で使用する．

PFENet とは異なる点として，図 2 の赤色で囲う部分の単語処理ブランチを追加している点である．word2vec [14] で埋め込まれたベクトルは VAE [11] を用いて潜在変数として単語ベクトルへと変換され, MMD [13] を用いた損失によってクエリの MAP ベクトルを再構成するように学習される．この単語ベクトルを，FEM の Inter-Source Enrichment の部分で使用され，プーリングされたクエリ特徴に MAP ベクトルと Prior Mask を連結したものと，クエリ特徴に単語ベクトルと Prior Mask を連結したものを別々に畳み込む構成になっている．

4.3 Zero-shot 手法

Few-shot の手法に加えて，サポートセットを使用しない手法の Zero-shot のモデルも提案する．アーキテクチャの詳細は図 2 の青い点線で囲まれている部分である．Few-shot で提案した手法と異なる点としてはサポートセットを使用しない点である．また，Prior Mask も生成されないため，FEM の Inter-Source Enrichment では，クエリの特徴量と生成された単語ベクトルは，チャンネル方向に連結し，処理される構造となっている．Tian らの PFENet [9] で提案されていた Zero-shot モデルとの相違点は，単語ベクトルを 1 × 1 convolution+Relu で変換し，FEM でクエリ特徴に直接連結する手法である点である．

4.4 損失関数

本研究の損失関数では ZS3Net [12] で用いられていた Maximum Mean Discrepancy(MMD) [13] を使用し，サンプリングした単語ベクトル x_G が Masked Average Pooling したクエリ特徴ベクトル y_Q を生成するように学習させる．MMD では 2 つの分布の差異を 2 つの分布 x, y のカーネル平均 m_x, m_y の平均 2 乗誤差 $\left|m_x - m_y\right|_H^2$ を利用して，分布間の違いを定量している．

MMD の推定値はガウスカーネル $k(x,x^{'}) = exp(-\frac{1}{2\sigma}|x - x^{'}|^2)$ を用いて (1) 式のように示すことができる．Li [13] らは，L_{MMD^2} の平方根を損失関数に使用することで，値が小さくなった際に効率よく 0 に近づける学習ができることを示している．本研究でも同様に L_{MMD^2} の平方根を用いた．

$$\begin{aligned} L_{MMD^2} &= |\sum_{x_G} x_G - \sum_{y_Q} y_Q|^2 \\ &= \sum_{x_G, x_G^{'}} {x_G}^T x_G + \sum_{y_Q, y_Q} {y_Q}^T y_Q - 2\sum_{x_G}\sum_{y_Q} {x_G}^T y_Q \\ &= \sum_{x_G, x_G^{'}} k(x_G, x_G^{'}) + \sum_{y_Q, y_Q^{'}} k(y_Q, y_Q^{'}) - 2\sum_{x_G}\sum_{y_Q^{'}} k(x_G, y_Q^{'}) \end{aligned} \tag{1}$$

$$L_{MMD} = \sqrt{L_{MMD^2}} \tag{2}$$

σ はカーネル平滑化パラメータであり, Yujia [13] らは複数のバンド幅を用いるとよりよい推定ができると主張している．そのため，ZS3Net [12] と同様の値 $\sigma_q = \{2, 5, 10, 20, 40, 60\}$ を使用した．複数使用する場合のガウスカーネルは (3) 式になる．

$$k(x,x^{'}) = \sum_{q=1}^{K} k_{\sigma_q}(x,x^{'}) \tag{3}$$

PFENet [9] では，L_{main} と補助損失 L_{aux}^i に Cross Entropy Loss(CE) が使用される．CE は以下の式で表される．

$$CE(p,q) = -\sum_x p(x) log(q(x)) \tag{4}$$

補助損失 L_{aux} は，FEM の Inter-Source Enrichment で出力された，X_{ISE}^i の特徴に対して異なる 4 つの分類出力層 ($Cls^i, i \in \{1,2,3,4\}$) で出力された予測マスク Y_{aux}^i の教師マ

© 2021 Information Processing Society of Japan

表 3 UECFoodPix-25^i を用いた実験結果

	UECFoodPix-25^i						
	Five-shot		One-shot		Zero-shot		
	wPFE (Ours)	PFE [9]	wPFE (Ours)	PFE [9]	zPFE (Ours)	PFE [9]	Kato [10]
Fold0	**0.851**	0.839	**0.847**	0.832	**0.808**	0.781	0.738
Fold1	**0.867**	0.860	**0.865**	0.855	**0.857**	0.822	0.767
Fold2	**0.817**	0.814	**0.818**	0.807	**0.788**	0.766	0.715
Fold3	**0.844**	0.840	**0.842**	0.832	**0.811**	0.776	0.722
Mean	**0.845**	0.838	**0.843**	0.832	**0.816**	0.786	0.736

スク Y_{true} との損失である．L_{main} は FEM で出力される最終予測マスク Y_{main} の Y_{true} との損失である．式 5 は補助損失 L_{aux}，式 7 は，本研究で使用する全体の損失関数である．

$$\begin{aligned} Y^i_{aux} &= Cls^i(X^i_{ISE}) \ , \ i \in \{1,2,3,4\} \\ L_{aux} &= \frac{1}{4}\sum_{i=1}^{4} CE(Y_{true}, Y^i_{aux}) \end{aligned} \tag{5}$$

$$L_{main} = CE(Y_{true}, Y_{main}) \tag{6}$$

$$L_{total} = L_{main} + L_{aux} + L_{MMD} \tag{7}$$

5. 実験

5.1 データセット

食事データセットとして岡本らが作成した UEC-FoodPix Complete [1] を使用した．UEC-FoodPix Complete は UECFood-100 を拡張して作成され，102 カテゴリ，計 1 万枚がアノテーションされており，各アノテーションにはクラスラベルが付与されている．実験では 102 カテゴリから others と beverage のカテゴリを除いた，100 カテゴリを使用した．また，食事画像においてメインで扱っているデータのみを扱うために，UECFood100 のカテゴリに基づいた UEC-FoodPix Complete のカテゴリを使用している．

Amirreza [5] らは，Few-shot Segmentation のデータセットとして，全 20 カテゴリが存在する PASCAL VOC 2012 データセットを 4 等分し，各 5 カテゴリに分けた，Pascal-5^i データセットを作成した．この構成が Few-shot Segmentation では主流となっている．本研究も同様に UEC-FoodPix Complete を 1 つの Fold に 25 カテゴリを割り当て，表 4 に示すように 4 分割したものを UECFoodPix-25^i データセットと定義した．また，一般的な物体を用いた実験には PASCAL-5^i データセットを使用した．Zero-shot タスクの実験では，Kato ら [10] が Pascal-5^i データセットを用いた実験をしていたため，本論文の提案手法との比較を行うために同様に Zero-shot タスクにおいて Pascal-5^i および UECFoodPix-25^i データセットを使用した．

5.2 実験詳細

Few-shot モデル (wPFE) と Zero-shot モデル (zPFE) の提案手法の有効性を検証するために，(1) 各データセットによる定量分析，(2) 埋め込み単語の有効性の検証を行った．(1) では，一般的な物体で構成されている Pascal-5^i と UECFoodPix-25^i データセットを用いて実験を行った．(2) では, UECFoodPix-25^i データセットを用いて検証を行った．

Pascal-5^i を用いた実験では backbone に ImageNet で事前学習したもの，word2vec の学習データには日本語 wiki を使用したモデルを使用し，UECFoodPix-25^i を用いた実験では，事前学習に Food-101 [15] を使用し，word2vec の学習データには CookPad の 16 万レシピを使用した．

(2) の実験では，word2vec [14] に用いる学習データに wiki の 4 千万文を用いる場合と，CookPad のレシピ量 (1 万と 16 万) の違いによる実験を zPFE モデルを用いて行った．日本語 wiki を用いた場合に“がんもどき”などの存在しない単語は，使用されている食材のベクトルの平均を取ることで，その食事カテゴリの埋め込みベクトルとした．さらに，埋め込んだ単語を視覚特徴に変換する再構成手法の比較の実験を Few-shot モデルで行った．比較した手法は，PFENet [9]+Generator(GMMN) [13]+MMD, PFENet+VAE [11]+MMD(提案手法) と PFENet の One-shot モデルに埋め込み単語を再構成せずに直接 FEM で用いる手法 (“None”) を使用した．

Five-shot 手法は，5 枚のサポート画像から抽出した特徴の

表 4 UECFoodPix-25^i データセット

Fold-0	Fold-1	Fold-2	Fold-3
親子丼	ピラフ	鰻丼	白米
チキンライス	寿司	ビーフカレー	かつ丼
トースト	ビビンバ	天丼	チャーハン
菓子パン	レーズンパン	ロールパン	クロワッサン
うどん	サンドイッチ	ピザ	ハンバーガー
チャーシュー麺	ラーメン	そば	天ぷらうどん
お好み焼き	スパゲッティ	焼きそば	天津麺
コロッケ	野菜炒め	グラタン	たこ焼き
味噌汁	野菜の天ぷら	ほうれん草のソテー	焼きナス
オムレツ	おでん	ソーセージ	ポタージュスープ
焼き魚	シチュー	餃子	がんもどき
刺身	鮭のムニエル	焼き鮭	魚のフライ
鰹のたたき	酢豚	すき焼き	秋刀魚の塩焼き
豚カツ	唐揚げ	天ぷら	茶碗蒸し
ハンバーグ	肉じゃが	煮魚	鰺の南蛮漬け
麻婆豆腐	生姜焼き	干物	ステーキ
目玉焼き	卵焼き	ロールキャベツ	焼き鳥
冷やし中華	春巻き	冷奴	納豆
海鮮丼	筑前煮	豚の角煮	青椒肉絲
ローストチキン	エビチリ	鯛焼き	ちらし寿司
ミートスパゲッティ	カツカレー	オムライス	シュウマイ
マカロニサラダ	グリーンサラダ	ポテトサラダ	エビフライ
牛丼	中華スープ	豚汁	けんちん汁
つけ麺	ピザトースト	おにぎり	きんぴらごぼう
ゴーヤチャンプルー	炊き込みご飯	ポテト	ホットドック

© 2021 Information Processing Society of Japan

Vol.2021-NL-248 No.21
Vol.2021-CVIM-226 No.21
2021/5/21

表 5　Pascal-5^i を用いた実験結果

	Pascal-5^i						
	Five-shot		One-shot		Zero-shot		
	wPFE (Ours)	PFE [9]	wPFE (Ours)	PFE [9]	zPFE (Ours)	PFE [9]	Kato [10]
Fold0	0.614	**0.628**	0.599	**0.617**	**0.524**	0.522	0.420
Fold1	0.683	**0.709**	0.681	**0.695**	0.637	**0.690**	0.583
Fold2	0.528	**0.564**	0.523	**0.554**	0.465	**0.524**	0.450
Fold3	0.539	**0.584**	0.541	**0.563**	0.442	**0.467**	0.364
Mean	0.591	**0.621**	0.586	**0.607**	0.517	**0.551**	0.454

平均を取った特徴をネットワークで使用している．評価指標として mIoU(mean Intersection over Union) を使用した．

学習は，3 つの Fold で学習し 1 つの Fold でテストを行った．学習は 200 エポック行い，この期間の中で評価値が最大だったものを本研究で使用するモデルとする．初期学習率は 2.5e-03，Deeplab [17] で使用されている，式 (8) の “poly” に従い学習率は更新される．poly で使用される *power* は 0.9 である．最適化手法として MomentumSGD を使用し，Momentum の値は 0.9 とし，バッチサイズは 4 とした．また，Data Augmentation として Random Scaling，Random Rotate，Random GaussianBlur，Random HorizontalFlip，Random Crop を使用した．Few-shot も Zero-shot も同様に出力は各ピクセルの背景か対象領域かの 2 クラス分類であり，ネットワークの出力に softmax 関数を使用し，ピクセルの前景背景を推論した．学習は 3 つの Fold で行い，評価に残り 1 つの Fold で評価を行った．Few-shot の手法では，サポートセットが 1 組の One-shot Segmentation とサポートセットが 5 組の Five-shot Segmentation を行った．サポートセットはランダムに選択し，5 回の平均スコアを使用した．Zero-shot の手法では，推論時に入力画像とその画像のカテゴリ名がネットワークの入力として渡され，カテゴリ名は埋め込み単語へと学習済み word2vec で変換されていく．Zero-shot では，サポート画像を使用しないため平均スコアは計算しない．実験 (1) の結果は表 3，表 5，実験 (2) の結果は表 6，表 7，定性分析の結果は，定量評価においてベースラインとの差が 0.1 以上改善されたカテゴリを使用し，図 4，図 6 に示した．

$$lr_{current} = lr_{base} \times \left(1 - \frac{iter}{max\_iter}\right)^{power} \qquad (8)$$

5.3　実験結果

実験 (1) について，提案手法は，UECFoodPix-25^i データセットを用いた場合，全 Fold においてベースラインよりも有効であることを示す結果となった．一方，Pascal-5^i を用いた実験では Zero-shot，Few-shot 共に精度が低下する結果となった．理由としては，wiki を用いた埋め込みはレシピ

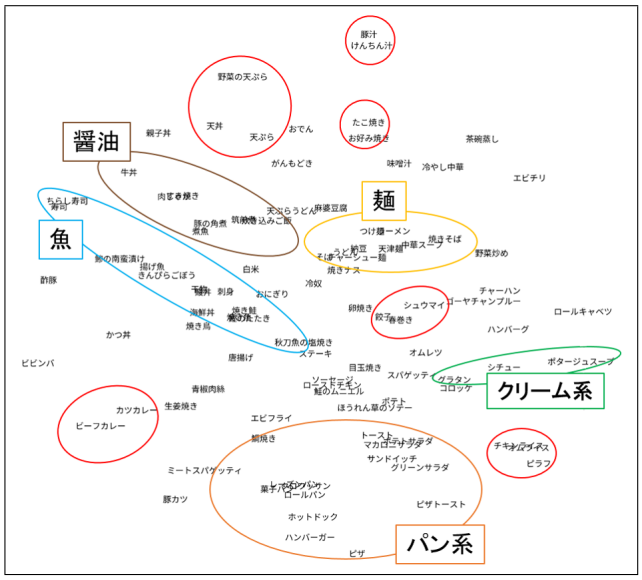

図 3　UECFoodPix-25^i のカテゴリ散布図 (囲っている部分は同じ食材が使用されている)

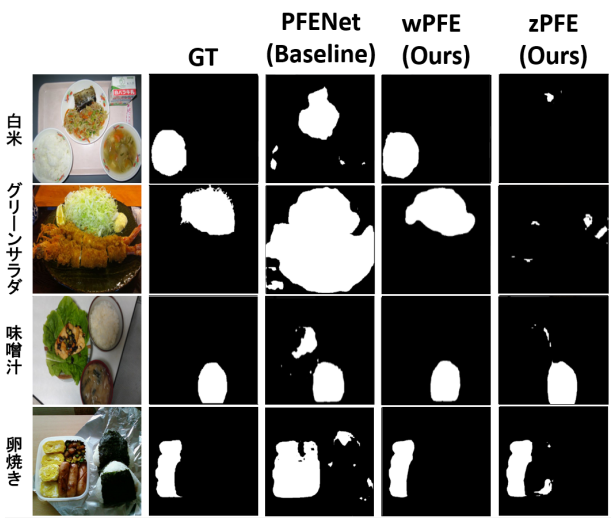

図 4　UECFoodPix-25^i での定性的実験の結果

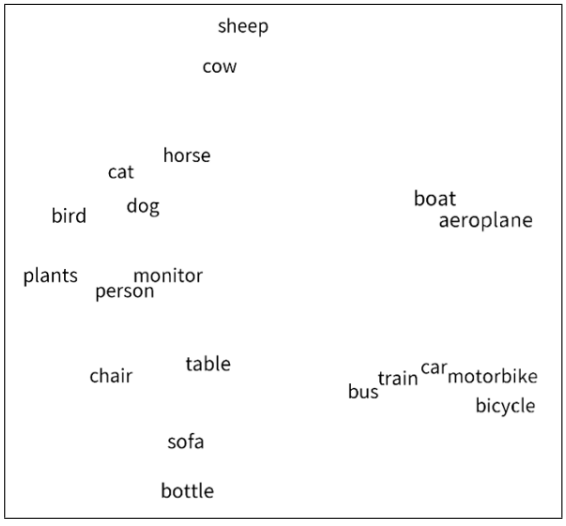

図 5　Pascal-5^i のカテゴリ散布図

© 2021 Information Processing Society of Japan

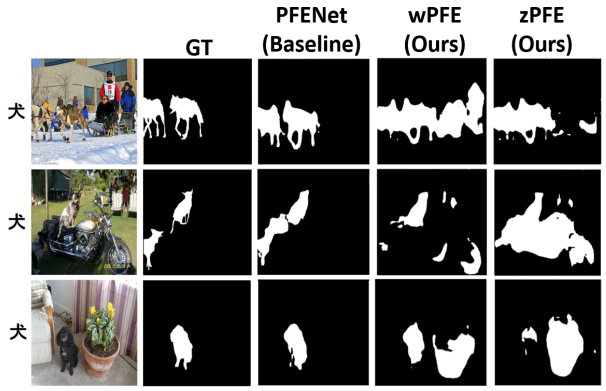

図 6 Pascal-5^i での定性的実験の結果

の食材単語で学習した word2vec [14] モデルと比べて，文章で学習したモデルであるため，図 5 で見られるように潜在空間内に視覚情報を反映していない埋め込みがされることが見てわかる．そのため結果的に再構成したベクトルにも影響を与え，精度が低下したと考えられる．また，図 6 で見られるように，提案手法は犬だけでなく学習カテゴリに含まれる植木鉢や人，バイクの領域までも誤ってセグメンテーションしている．提案手法は，テスト時に学習カテゴリ内の単語ベクトルの中間表現で表現しようとしする．そのため，カテゴリ数が少ない Pascal-5^i データセットを使用した場合，食事データセットに比べて類似カテゴリの数が少なくなっているため，類似している他の単語の特徴が扱えず，精度向上が困難であったと思われる．このことから提案手法は，同一ドメインのデータセットとの相性の良いモデルではないかと考えられる．

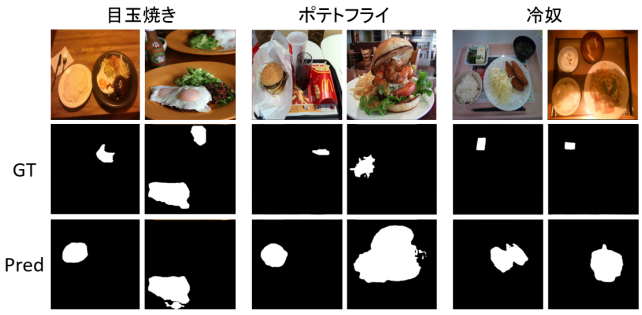

図 7 mIoU:0.6 以下のカテゴリの定性的評価 (wPFE モデル)

図 7 は，wPFE モデルによる評価の低かったカテゴリの推論結果である．特に評価値が低くなるカテゴリの傾向としては，入力画像に占める対象領域の部分が小さな画像である場合に評価値が低い傾向であることが分かった．原因としては，対象領域が小さい分情報量が少なくなることでサポート画像によって対象領域と認識するのが困難になるのではないかと考えられる．また，カテゴリ内に対象領域の小さなものが多く存在する場合，サポート画像に対象領域の小さい画像を選択する可能性が高くなる．結果として，対象物体が大きいサポート画像と比較すると，対象領域の小さい画像のサポート画像の MAP (Masked Average Pooling) ベクトルは対象物体の大域的な特徴のみを捉えるため，誤った領域を分割するのではないかと考えられる．

表 6 単語再構成手法の比較

	+GMMN [13]+MMD	+VAE [11]+MMD	None
Fold0	0.831	**0.847**	0.828
Fold1	0.846	**0.865**	0.859
Fold2	0.813	**0.818**	0.814
Fold3	**0.846**	0.842	0.826
Mean	0.834	**0.843**	0.832

表 7 学習データによる比較

	CookPad(10K)	CookPad(160K)	wiki(40M)
Fold0	0.798	**0.808**	0.807
Fold1	0.823	**0.857**	0.827
Fold2	0.780	**0.788**	0.772
Fold3	0.783	**0.811**	0.801
Mean	0.796	**0.816**	0.802

実験 (2) について，表 7，表 6 の結果から，word2vec [14] の学習には 16 万レシピを使用し，再構成には VAE [11] を用いたモデルが有効であることが分かった．word2vec の学習で用いる単語はレシピから抽出した食材単語であるため，図 3 で見られるように料理の食材的に近いもの同士が同じ位置に分布し，見た目に影響する食材を用いる料理同士の分布も近くなる．そのため，単語ベクトルに視覚的特徴量が反映され，新規カテゴリの単語で，学習カテゴリの対象領域に似た料理の特徴量を組み合わせたベクトルを生成するようになる．このことによって，図 8 で見られるように，再構成手法による対象領域の視覚的特徴量への変換が容易になったと言える．また，表 7 のようにレシピ量を増やすとより多くの視覚的情報を扱うことが可能となるため精度が向上し，図 4 にあるようにベースラインよりも識別的な領域分割ができたと考えられる．

6. おわりに

本研究では，PFENet をベースに単語埋め込みを加えた新しい手法を提案した．実験では，食事データセットで Few-shot で従来手法を上回る性能を達成し，提案手法の有効性を示した．食事画像データセットを用いることによって，学習カテゴリとテストカテゴリ間で見た目や形状の類似性があるため，Few-shot タスクの問題点であるカテゴリ間での大きなドメインシフトが存在しない．そのため，一

© 2021 Information Processing Society of Japan

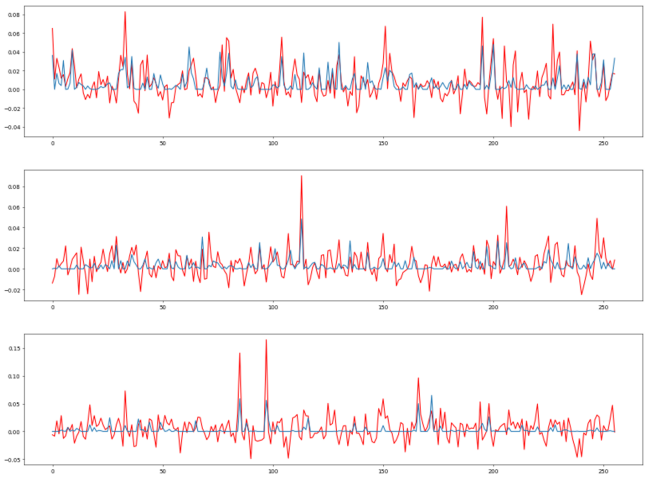

図 8 再構成ベクトル（赤色が再構成した特徴量，青色がクエリの対象領域の特徴量で平均を取った MAP ベクトル　上：酢豚，中：けんちん汁　下：冷やし中華）

般的な物体で構成されているデータセットの Pascal-5^i の実験においてベースラインよりも精度は劣るが，食事画像データセット UECFoodPix-25^i を用いた実験では，改善することができ，有効性を示すことができたといえる．

埋め込み手法の word2vec では，CookPad のレシピデータから食材単語を使用し，生成する潜在空間に視覚的特徴を持たせることによって，従来手法と比べ識別的な領域分割が可能となった．また，MMD を用いた損失関数によって再生成ベクトルをクエリの視覚的特徴に類似させるように学習を促すことによって精度が向上した．

今後の課題としては，Pascal-5^i データセットを用いた実験ではベースラインよりも精度が下がってしまうという結果となってしまった．そのため，一般的な物体で構成されている，80 カテゴリの MSCOCO [18] データセットを用いることで提案手法の有効性が示せるのではないかと期待できる．また，よりよい単語埋め込み手法として BERT [19] の利用が考えられる．

謝辞：本研究では NII 情報学研究データリポジトリで公開されている CookPad データセットを利用している．

参考文献

[1] Okamoto, K. and Yanai, K.: UEC-FoodPix Complete: A Large-scale Food Image Segmentation Dataset, *Proc.of the ICPR Workshop on Multimedia Assisted Dietary Management (MADIMA)* (2021).
[2] Flood, S., Yongxin, Y., Li, Z., Tao, X., Philip, H. and Timothy, H.: Learning to Compare: Relation Network for Few-Shot Learning, *Proc.of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
[3] Li, X., Wei, T., Chen, Y., Tai, Y. and Tang, C.: FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation, *Proc.of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
[4] Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, *Proc.of Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015).
[5] Shaban, A., Bansal, S., Liu, Z., Essa, I. and Boots, B.: One-shotlearning for semantic segmentation., *Proc.of British Machine Vision Conference (BMVC)* (2017).
[6] Koch, G., Zemel, R. and Salakhutdinov, R.: Siamese neural networks for one-shot image recognition, *Proc.of International Conference on Machine Learning (ICML)* (2015).
[7] Zhang, X., Wei, Y., Yang, Y. and Huang, T.: SG-One: Similarity Guidance Network for One-Shot Semantic Segmentation, *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
[8] Zhang, C., Lin, G., Liu, F., Yao, R. and Shen, C.: CANet: Class-Agnostic Segmentation Networks with Iterative Refinement and Attentive Few-Shot Learning, *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).
[9] Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R. and Jia, J.: Prior Guided Feature Enrichment Network for Few-Shot Segmentation, *Proc.of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020).
[10] Kato, N., Yamasaki, T. and Aizawa, K.: Zero-Shot Semantic Segmentation via Variational Mapping, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* (2019).
[11] Kingma, P. and Welling, M.: Auto-Encoding Variational Bayes, *Proc.of International Conference on Machine Learning (ICML)* (2014).
[12] Bucher, M., Vu, T., Cord, M. and Pérez, P.: Zero-Shot Semantic Segmentation, *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).
[13] Li, Y., Swersky, K. and Zemel, R.: Generative Moment Matching Networks, *Proc.of Proceedings of International Conference on Machine Learning (ICML)* (2015).
[14] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *Proc.of International Conference on Learning Representations (ICLR)* (2013).
[15] Bossard, L., Guillaumin, M. and Van Gool, L.: Food-101 Mining Discrimi-native Components with Random Forest, *Proc.of European Conferenceon Computer Vision (ECCV)* (2014).
[16] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
[17] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, *Proc.of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2018).
[18] Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R. Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, L.: Microsoft COCO: Common Objects in Context, *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).
[19] Devlin, J., Chang, M., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc.of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (2019).

© 2021 Information Processing Society of Japan