食事画像に対する Few/Zero-shot Segmentation

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，単語埋め込み用い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 [9] がある．しかし，無数に存在する食事カテゴリに対して，深層学習モデルの学習データには不十分であるといえる．近年研究されている Few-shot Segmentation では，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 では，学習データと検証データ間に共通カテゴリは存在しないため，推論の際に学習カテゴリの領域を誤って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た 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．また，同じ食材を使用しているカテゴリは，見た目が類似するという着想を基に，食材の共起関係をモデルに組み込むことで新たな 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．

2.2 Few-shot Segmentation

Tian らの提案した Prior Guided Feature Enrichment Network (PFENet) [10] では，サポートとクエリの高レベル特徴量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，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む 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 Masked Average Pooling (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 [9] 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と，Pascal-5^i を使用した実験において PFENet は One-shot Segmentation タスクで最高精度を達成しているためである．FEM のアーキテクチャの詳細は図 1 に示す．

[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@cs.uec.ac.jp

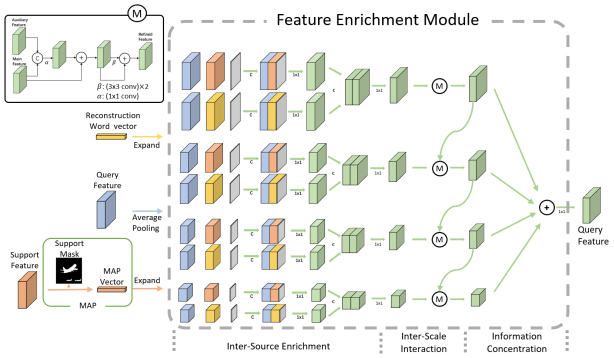

図 1 単語処理を加えた Feature Enrichment Module のアーキテクチャ.

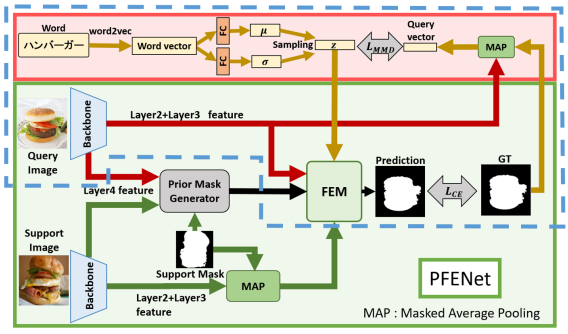

図 2 One-shot タスクに対して本研究で提案するネットワーク図. 青い点線は Zero-shot タスクに対応する部分.

2.3 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. Bucher らは [2] ノイズを連結した埋め込み単語ベクトルに対して，全結合層と Leaky ReLU で構成されている Generative Moment Matching Networks (GMMN) [7] で変換したものを Maximum Mean Discrepancies (MMD) を用いた再構成損失を使用して視覚的特徴量への変換を促す手法を取っている. Kato ら [5] は，Variational Autoencoder (VAE) [6] で潜在的空間の特徴として変換し視覚特徴特徴と連結し，畳み込むことで視覚的特徴量として扱っている. 本研究でもこれらと同様の変換手法と損失関数を使用する.

3. 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.

3.1 単語埋め込み手法

単語埋め込みモデルとして word2vec [8] 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 [9] 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. 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と考えたため，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，これを Few-shot と Zero-shot に組み込む手法を提案する.

3.2 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 [10] 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 [1] で事前学習された ResNet50 [4] 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，MAP を行ったもの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 [8] で埋め込まれたベクトルは VAE [6] を用いて潜在変数として単語ベクトルへと変換され，MMD [7] 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.

3.3 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が連結し，処理される構造となっている．Tian らの PFENet [10] 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，再構成はせずに FEM でクエリ特徴に直接連結する点と提案手法では再構成ベクトルに再構成損失を追加している点である．

3.4 損失関数

本研究の損失関数では ZS3Net [2] で用いられていた Maximum Mean Discrepancy(MMD) [7] 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 を Gaussian カーネルで再生核ヒルベルト空間に写像した (m_x, m_y) 分布間の平均 2 乗誤差 $L_{MMD^2} = \|m_x - m_y\|_H^2$ を利用して，分布間の違いを定量している．Li [7] 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．全体の損失はクロスエントロピー損失を使用する L_{PFENet} を用いて $L_{total} = L_{PFENet} + L_{MMD}$ と表される．

4. 実験

4.1 データセット

Few-shot データセットとして，UEC-FoodPix Complete [9] を 4 つの Fold に 4 等分したものを UECFoodPix-25^i と定義した．また，一般的な物体データセットとして従来手法で使用されていた Pascal-5^i も使用した．

4.2 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 Few-shot 用の Pascal-5^i と UECFoodPix-25^i で実験を行った．Pascal-5^i を用いた実験では backbone に ImageNet で事前学習したものを使用した．(2) の実験では，UECFoodPix-25^i データセットを使用した．word2vec [8] 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 モデルで行った．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 で行った．比較した手法は，PFENet [10]+Generator(GMMN) [7]+MMD，PFENet+VAE [6]+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．結果は表 2 である．

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．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．実験 (1) の結果は表 1，実験 (2) の結果は表 2，定性分析の結果は図 3，図 4 である．

表 1 定性的評価による実験結果

	UECFoodPix-25^i					Pascal-5^i				
	One-shot		Zero-shot			One-shot		Zero-shot		
	wPFE (Ours)	PFENet [10]	**zPFE (Ours)**	PFENet [10]	Kato [5]	**wPFE (Ours)**	PFENet [10]	**zPFE (Ours)**	PFENet [10]	Kato [5]
Fold0	**0.847**	0.832	**0.808**	0.781	0.738	0.599	**0.617**	**0.524**	0.522	0.420
Fold1	**0.865**	0.855	**0.857**	0.822	0.767	0.681	**0.695**	0.637	**0.690**	0.583
Fold2	**0.818**	0.807	**0.788**	0.766	0.715	0.523	**0.554**	0.465	**0.524**	0.450
Fold3	**0.842**	0.832	**0.811**	0.776	0.722	0.541	**0.563**	0.442	**0.467**	0.364
Mean	**0.843**	0.832	**0.816**	0.786	0.736	0.586	**0.607**	0.517	**0.551**	0.454

表 2 左:学習データによる比較　右:単語再構成手法の比較

	CookPad(10K)	CookPad(160K)	wiki(40M)		+GMMN [7]+MMD	+VAE [6]+MMD	None
Fold0	0.798	**0.808**	0.807	Fold0	0.831	**0.847**	0.828
Fold1	0.823	**0.857**	0.827	Fold1	0.846	**0.865**	0.859
Fold2	0.780	**0.788**	0.772	Fold2	0.813	**0.818**	0.814
Fold3	0.783	**0.811**	0.801	Fold3	**0.846**	0.842	0.826
Mean	0.796	**0.816**	0.802	Mean	0.834	**0.843**	0.832

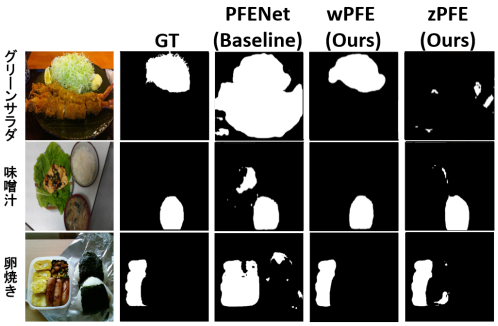

図 3 UECFoodPix-25^i での定性的実験の結果

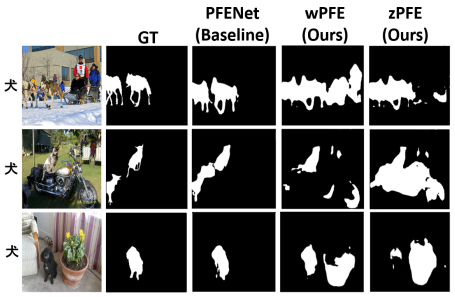

図 4 Pascal-5^i での定性的実験の結果

4.3 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，One-shot 共に精度が低下する結果となった．wiki を用いた埋め込みは，レシピの食材単語で学習した word2vec [8] モデルと比べて，文章で学習したモデルは，図 7 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と考えられる．そのため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，カテゴリ数が少ないために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のも，精度向上が難しい原因となっていると思われる．

実験 (2) について，表 2 の結果から，word2vec [8] の学

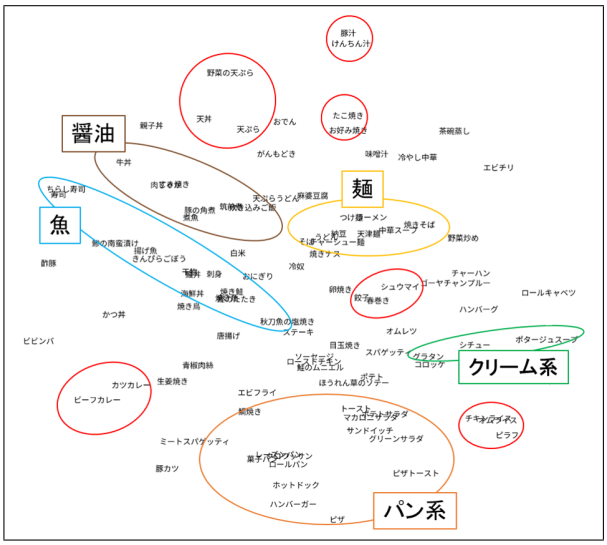

図 5 UECFoodPix-25^i のカテゴリ散布図 (囲っている部分は同じ食材が使用されている)

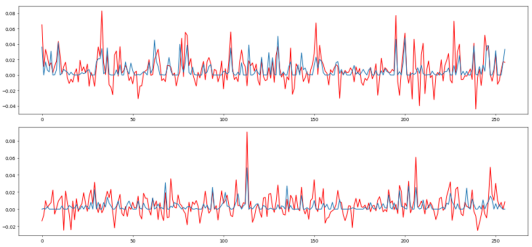

図 6 新規カテゴリの再構成ベクトル（赤色が再構成した特徴量，青色がクエリの対象領域の特徴量で平均を取った MAP ベクトル　上：酢豚，下：けんちん汁）

習には 16 万レシピを使用し，再構成には VAE [6] 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 5 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 6 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 2 の左表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，表 1 や図 3 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．

5. 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は Few-shot，Zero-shot ともに従来手法を上回る性能を達成し，食事データセットにおける提案手法の有効性を示した．

今後の展望としては，一般物体にも適応可能な汎用性の高いモデルの構築や，よりよい単語埋め込み手法として

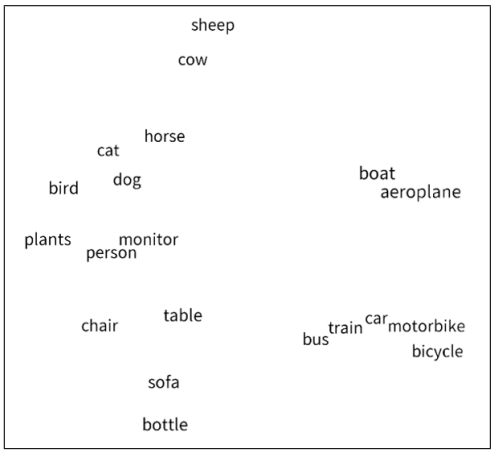

図 7 Pascal-5^i のカテゴリ散布図

BERT [3] 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] 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).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

[6] Kingma, P. and Welling, M.: Auto-Encoding Variational Bayes, *Proc.of International Conference on Machine Learning (ICML)* (2014).

[7] Li, Y., Swersky, K. and Zemel, R.: Generative Moment Matching Networks, *Proc.of Proceedings of International Conference on Machine Learning (ICML)* (2015).

[8] 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).

[9] 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).

[10] 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).