視覚言語事前学習と推論で共通の DNN を用いる一般化 Few-Shot 物体検出

長野 紘士朗[1,a)]　佐藤 文彬[1,b)]　八馬 遼[1,c)]　関井 大気[1,d)]

概要

本稿では，一般化 Few-Shot 物体検出の従来研究において応用上拡張性が制限される 2 つの課題，検出対象物体の少数サンプルを Few-Shot 学習するために必要となる特別な学習コスト，少数サンプルのバリエーションの不足，を同時に解決する手法を提案する．具体的には，視覚言語事前学習が適用された DNN の汎化性を活用し，Few-Shot 学習時に DNN の更新なしに事前学習時と共通の DNN を用いて，対象物体の見えに関する特徴量をモデル化する枠組みを提案する．実験では，比較的物体の見えの異なる 2 つの公開データセットを事前学習と Few-Shot 学習に用いて，従来法と提案法の検出精度を比較することで，先述の課題に対する提案法の有効性を検証する．

1. はじめに

物体検出 [20,21] は，動画像から物体を検出するとともに画像上の位置とサイズを求める技術であり，Deep Learning 技術 [11] の導入以降飛躍的に性能が向上し，自動運転 [27] やロボティクス [9]，医用画像処理 [23] など様々な分野に応用されている．近年では，画像に関する文章を自然言語処理技術 [25] を用いて変換した特徴量（以降，文章特徴量）を活用する視覚言語事前学習 [19] が導入され，ユーザーが文章入力で説明した，事前学習時に未知であった物体クラス（以降，未知クラス）を，追加の少数の学習サンプル（以降，少数サンプル）を用いることなしに Zero-Shot で検出することが可能となっている．

これに対して，事前学習に用いるデータセットで網羅できていない希少な特徴を持つクラスに対しては，少数サンプルを活用する Few-Shot 学習が未だ有効である一方，次節で述べるように，Zero-Shot 学習手法に比べ，応用において学習コストや検出の頑健性に関する複数の課題が残っている．本稿では物体検出の Few-Shot 学習に着目し，応用における課題解決に取り組む．

[1] コニカミノルタ株式会社
[a)] koshiro.nagano@konicaminolta.com
[b)] fumiaki.sato1@konicaminolta.com
[c)] ryo.hachiuma@konicaminolta.com
[d)] taiki.sekii@konicaminolta.com

1.1 関連研究と課題

物体検出のための Few-Shot 学習の従来研究は，メタ学習ベースの手法と転移学習ベースの手法に大別できる．前者の従来研究の多くは，未知クラスに関する特徴量の条件付けに少数サンプルを用いる [5,7,8,10,17,18,26,29,30]．近年では，このような条件付けに注意機構が数多く用いられており，少数サンプルに対する注意機構の出力を，検出対象の入力画像（以降，入力画像）全体から検出した物体領域候補の特徴量の重み付け [30] や修正 [5] に用いる．これに対して，転移学習ベース手法は，未知クラスを含まない大規模なデータセットと未知クラスの少数サンプル両方に対して DNN を最適化する．近年では，少数サンプルに対する汎化性を獲得するため，Fine-Tuning 時の正則化 [1] や少数サンプルの拡張 [28] が試みられている．以上に述べた従来研究では，少数サンプルの導入に際して，破滅的忘却 [15,16] と呼ばれる，未知クラス以外のクラス（事前学習した物体クラスなどの既知クラス）に対して検出精度が低下する現象が避けられない．

この問題に対して，いずれのクラスに対しても汎化性を維持することを目的とした，一般化 Few-Shot 物体検出（Generalized Few-Shot Object Detection，G-FSOD）と呼ばれる手法が提案されている．Retentive R-CNN [4] は，事前学習済みの DNN と，少数サンプルで Fine-Tuning された DNN を結合することで，また CFA [6] は，少数サンプルに対する過学習を抑制するための条件付き最適化を導入することで，それぞれ破滅的忘却を抑制している．一方，視覚言語事前学習に基づく Zero-Shot 学習手法 [31] は，破滅的忘却なしにユーザーが文章で説明した未知クラスを検出できる一方，文章で説明できるクラスに検出対象が制限される．

本研究の着目する G-FSOD の従来研究では，以下に述べるように応用面でいずれかの課題があり，利用できるシーンの拡大や性能の改善に際して拡張性に乏しい．

課題 1．少数サンプルの学習コスト

Few-Shot 学習のための処理（未知クラス専用の特徴抽

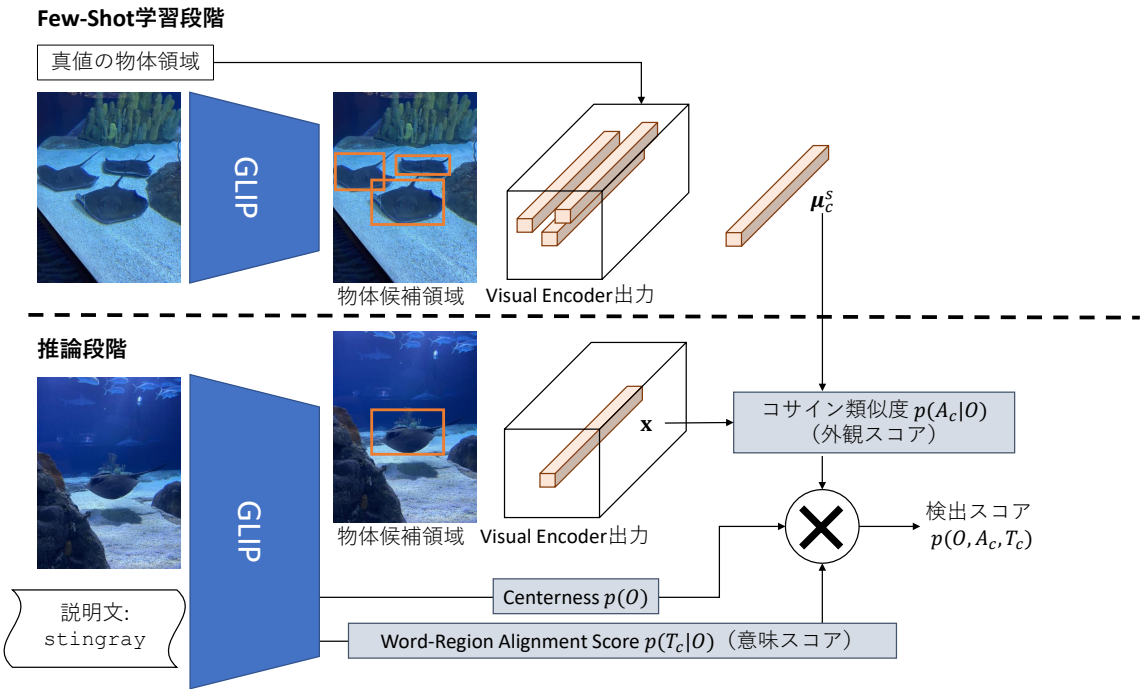

図 1: 提案法の概要.

出器や DNN の更新）に，一定以上の性能のプロセッサや，DNN を更新するシステムが求められ，計算機コストや開発コストが必要となる．

課題 2. 少数サンプルのバリエーションの不足

少数サンプルの数が不足している場合，検出対象物体の見えのパターン数と，既知クラスから学習した特徴表現の影響により，未知クラスの見えに関する情報の不足から検出精度が低下するため，利用できるシーンが制限される．

1.2 本研究の概要と貢献

本研究では，応用における G-FSOD の課題を解決するため，比較的大規模なデータセットを用いた視覚言語事前学習で得られる DNN の汎化性を活用し，Few-Shot 学習時に DNN の更新なしに事前学習時と共通の DNN を用いる問題設定に取り組む．具体的には，Few-Shot 学習時には，未知クラスの物体領域の見えを，少数サンプルから抽出する特徴量をクラスごとに平均した平均ベクトルとしてモデル化 [*1]する．未知クラスごとの平均ベクトルを，推論時に検出される物体候補領域のクラス分類に用いることで，未知クラス専用の特徴抽出器や重みの更新を排し先述の課題 1 に対処する．また，先述の課題 2 に対して，ユーザーが入力した未知クラスの説明文の文章特徴量もまた物体候補領域のクラス分類の条件付けに用いる．これにより，未知クラスに関する情報の不足を補い検出の頑健化を促進する．

実験では，比較的物体の見えの異なる 2 つの公開データセットを事前学習と Few-Shot 学習に用いて，従来法と提案法の検出精度を比較することで，先述の課題に対する提案法の有効性を検証する．

本論文の貢献は，前節で述べた Few-Shot 学習の応用上の課題に対して，(1) Few-Shot 学習時に DNN の更新なしに事前学習時と共通の DNN を用いる G-FSOD が実現可能であり，(2) 未知クラスの説明文の文章特徴量が，未知クラスに関する追加の情報として検出の頑健化に有効であることの 2 点を示すことである．

2. 提案法

提案法の概要を図 1 に示す．本研究では，物体検出の基本的な枠組みとして GLIP [12] を採用する．視覚言語事前学習が適用された GLIP を Few-Shot 学習段階から推論段階まで共通に用いる．Few-Shot 学習で得られる，未知クラスの見えを表現する特徴ベクトルは，ユーザーが入力した未知クラスの説明文とともに，推論時に検出される物体領域候補のクラス分類に用いられる．

2.1 Few-Shot 学習段階

Few-Shot 学習段階では，クラスラベル付きの物体領域（バウンディングボックス）の真値がアノテーションされた少数サンプルの画像に GLIP を適用することにより，物体候補領域を検出する．次に，これらの物体候補領域を真値の物体領域と対応付ける．対応付いた候補領域に対して，Visual Encoder から出力された特徴マップ上で該当する位置からチャンネル方向の特徴ベクトルを割り当てる．ただし，真値の物体領域には，IoU（Intersection of Union）が 0.9 以上かつ候補の中で最大となる候補領域を対応付ける．

[*1] 本研究ではこのステップを Few-Shot 学習と位置付ける．

特徴マップの各スケール s において，真値と対応付いた候補領域の特徴ベクトルを未知クラス c ごとに平均することにより，スケール・クラスごとの平均ベクトル μ_c^s を得る．

2.2 推論段階

推論段階では，前節同様に GLIP を用いて入力画像から物体候補領域を求めるとともに，未知クラスごとの平均ベクトルと未知クラスを説明する文章を用いて，各候補領域に対して次式のように未知クラス c に関する検出スコア $p(O, A_c, T_c)$ を計算する．

$$p(O, A_c, T_c) = p(O)p(A_c|O)p(T_c|O), \qquad (1)$$

ただし，$p(O)$ は候補領域の検出スコア（Centerness [24]）である．$p(A_c|O)$ は，候補領域が未知クラス c の見えに合致するスコア（以降，外観スコア）であり，次式のようにコサイン類似度 $\mathrm{Cos}(\cdot,\cdot)$ を用いて近似する．

$$p(A_c|O) \sim \max\left\{0, w_1 \mathrm{Cos}(\mathbf{x}, \mu_c^s)^{1/w_2}\right\}, \qquad (2)$$

ただし，$\mathbf{x}$ は，前節で述べた μ_c^s の計算と同様の方法で Visual Encoder から抽出した候補領域の特徴ベクトルであり，w_1，w_2 はそれぞれ正規化定数，温度パラメータである．$p(T_c|O)$ は，ユーザーが入力した未知クラスの説明に候補領域の物体が合致しているかを表すスコア（以降，意味スコア）である．本論文では，$p(T_c|O)$ を GLIP で導入された Word-Region Alignment Score を用いて近似する．

2.3 提案法の特長

GLIP を構成する DNN は未知クラスを含まないデータセットを用いて事前学習済みであり，Few-Shot 学習段階で μ_c^s を計算する際には，DNN が少数サンプルに合わせて更新されることはなく，破滅的忘却は原理的に起こらないことが保証される．加えて，推論段階でも同様に GLIP が用いられることを踏まえると，1.1 節の課題 1 で述べた Few-Shot 学習のための特別な学習コストは発生しないことがわかる．

また，1.1 節の課題 2 に対して，外観スコアとともに意味スコアを物体候補領域の検出スコアの条件付けに用いることにより，外観スコアの計算に用いる少数サンプルに不足している情報を，意味スコアを通じてユーザーは候補領域の検出スコアに反映できる．

3. 評価実験

3.1 データセット

提案法が用いる GLIP の設定にしたがい，DNN の視覚言語事前学習には比較的大規模なデータセットである Object-365 [22] を，また，Few-Shot 学習の評価には，Object-365 と物体の見えが異なる Aquarium データセット [12] を用いて提案法の有効性を検証する．Object-365 データセットは 365 種類の物体クラスから構成される一方，Aquarium データセットのクラスは 7 種類の海洋生物から成る．

表 1: Aquarium データセットを用いた従来法と提案法の物体検出精度の比較結果.

Method	AP （%）
GLIP w/ Original Prompts [12]	17.7
GLIP w/ Manually Designed Prompts [12]	18.4
Ours w/o $p(T_c\|O)$	15.5
Ours w/ $p(T_c\|O)$	19.5

Few-Shot 学習段階では，Aquarium データセットに含まれる学習データの画像 448 枚を少数サンプルとして 2.1 節で述べた平均ベクトル μ_c^s を計算し，また，推論段階では，同データセットのテストデータの画像 127 枚を用いる．Few-Shot 学習・推論段階両方において，入力画像はすべてアスペクト比を保ったまま $800 \times 1,066$ [px^2] の解像度にリサイズし用いた．

3.2 実装の詳細

GLIP の DNN 構造には，Visual Encoder として Swin Transformer（Tiny） [14] が，物体候補領域からの画像特徴量・文章特徴量の Encoder としてそれぞれ DyHead [2]・BERT [3] が，また，GLIP の物体検出全体の枠組みには RetinaNet [13] が採用された．DNN の重みは，GLIP の著者らによる実装[*2]により事前学習されたものである．

2.1 節で述べた未知クラス c ごとの平均ベクトル μ_c^s を，Visual Encoder からの 5 スケール（$s \in \{1, \cdots, 5\}$）の特徴マップそれぞれに対して計算した．また，式 2 における w_1，w_2^s（w_2 を s ごとに調整）を検出精度を最大化するよう調整し，それぞれ $w_1 = 1.0$，$(w_2^1, w_2^2, w_2^3, w_2^4, w_2^5) = (1.5, 2.0, 2.0, 1.0, 1.0)$（$s$ はスケール順），と定めた．意味スコアの計算に必要となる未知クラスの説明文に関する設定は，従来法の評価方法 [12] と同様である．

3.3 従来法に対する比較実験

従来法と提案法の物体検出精度（以下，精度）を比較した結果を表 1 に示す．比較対象の従来法として，提案法同様，事前学習から推論にかけて共通の DNN を用いる GLIP [12] を採用した．意味スコアを利用する提案法（Ours w/ $p(T_c|O)$）の精度は，認識対象の説明文として未知クラスの名称（例えば，`stingray`）を用いる従来法（GLIP w/ Original Prompts）を +1.8 [%pt] 上回り，さらに，認識対象の説明文に未知クラスの色や形状などの説明（例えば，`stingray which is flat and round`）が追加された従来法（GLIP w/ Manually Designed Prompts）を +1.1 [%pt] 上回っている．また，意味スコアを利用しない場合（Ours w/o $p(T_c|O)$）であっても，未知クラスの名称

[*2] https://github.com/microsoft/GLIP

を用いる従来法（GLIP w/ Original Prompts）との精度の差は数 [%pt] にとどまる．以上の結果を踏まえると，提案法は，1.1 節の課題 1 で述べた学習コストなしに物体の外観特徴を Few-Shot 学習可能であることがわかる．

また，提案法において物体候補領域の検出スコアに意味スコアを用いる場合（Ours w/ $p(T_c|O)$）の精度は，意味スコアを用いない場合（Ours w/o $p(T_c|O)$）の精度を約 +4 [%pt] 上回っている．これは，未知クラスの説明文が，1.1 節の課題 2 で述べた未知クラスの見えに関する情報の不足を，意味スコアを通じて補ったことによる効果である．

本実験では，未知クラスの海洋生物と同様の特徴表現を持つ物体が，事前学習用の Object-365 データセットに含まれていると考えられる．提案法が効果を発揮するためには，未知クラスの特徴表現を一定以上獲得できる物体が，事前学習する既知クラスに十分含まれている必要がある．

4. まとめ

本稿では，一般化 Few-Shot 物体検出の従来研究において応用上拡張性が制限される 2 つの課題，検出対象物体の少数サンプルを Few-Shot 学習するための特別な計算機コストや開発コスト，少数サンプルのバリエーションの不足，を同時に解決する手法を提案した．具体的には，視覚言語事前学習が適用された DNN の汎化性を活用し，Few-Shot 学習時に DNN の更新なしに事前学習時と共通の DNN を用いて，物体の見えをモデル化する枠組みを提案した．実験では，視覚言語事前学習に用いた Object-365 と物体の見えが異なる Aquarium データセットを用いて，比較対象の従来法である GLIP と提案法の検出精度を比較することにより，先述の課題に対する提案法の有効性を検証した．今後の課題として，Few-Shot 学習時に DNN が更新されない提案法の条件を，事前知識として DNN の正則化や帰納バイアスを通じて事前学習に導入することで，DNN 学習を条件付けし汎化を促進する改良があげられる．

参考文献

[1] Chen, H., Wang, Y., Wang, G. and Qiao, Y.: LSTD: A Low-Shot Transfer Detector for Object Detection, *AAAI* (2018).

[2] Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L. and Zhang, L.: Dynamic Head: Unifying Object Detection Heads With Attentions, *CVPR* (2021).

[3] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding, *NAACL* (2019).

[4] Fan, Z., Ma, Y., Li, Z. and Sun, J.: Generalized Few-Shot Object Detection Without Forgetting, *CVPR* (2021).

[5] Fan, Z., Yu, J.-G., Liang, Z., Ou, J., Gao, C., Xia, G.-S. and Li, Y.: FGN: Fully Guided Network for Few-Shot Instance Segmentation, *CVPR* (2020).

[6] Guirguis, K., Hendawy, A., Eskandar, G., Abdelsamad, M., Kayser, M. and Beyerer, J.: CFA: Constraint-based Finetuning Approach for Generalized Few-Shot Object Detection, *CVPRW* (2022).

[7] Hsieh, T.-I., Lo, Y.-C., Chen, H.-T. and Liu, T.-L.: One-Shot Object Detection with Co-Attention and Co-Excitation, *NeurIPS* (2019).

[8] Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J. and Darrell, T.: Few-Shot Object Detection via Feature Reweighting, *ICCV* (2019).

[9] Karaoguz, H. and Jensfelt, P.: Object Detection Approach for Robot Grasp Detection, *ICRA* (2019).

[10] Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R. and Bronstein, A. M.: RepMet: Representative-Based Metric Learning for Classification and Few-Shot Object Detection, *CVPR* (2019).

[11] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *NeurIPS* (2012).

[12] Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W. and Gao, J.: Grounded Language-Image Pre-Training, *CVPR* (2022).

[13] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollar, P.: Focal Loss for Dense Object Detection, *ICCV* (2017).

[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, *ICCV* (2021).

[15] Lopez-Paz, D. and Ranzato, M.: Gradient Episodic Memory for Continual Learning, *NeurIPS* (2017).

[16] McCloskey, M. and J. Cohen, N.: Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem, Psychology of Learning and Motivation, Vol. 24, Academic Press, pp. 109–165 (1989).

[17] Michaelis, C., Ustyuzhaninov, I., Matthias, B. and S. Ecker, A.: One-Shot Instance Segmentation, *arXiv:1811.11507* (2018).

[18] Perez-Rua, J.-M., Zhu, X., Hospedales, T. M. and Xiang, T.: Incremental Few-Shot Object Detection, *CVPR* (2020).

[19] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, *ICML* (2021).

[20] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection, *CVPR* (2016).

[21] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *NeurIPS* (2015).

[22] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J. and Sun, J.: Objects365: A Large-Scale, High-Quality Dataset for Object Detection, *ICCV* (2019).

[23] Sheoran, M., Dani, M., Sharma, M. and Vig, L.: DKMA-ULD: Domain Knowledge augmented Multi-head Attention based Robust Universal Lesion Detection, *BMVC* (2021).

[24] Tian, Z., Shen, C., Chen, H. and He, T.: FCOS: Fully Convolutional One-Stage Object Detection, *ICCV* (2019).

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *NeurIPS* (2017).

[26] Wang, Y.-X., Ramanan, D. and Hebert, M.: Meta-Learning to Detect Rare Objects, *ICCV* (2019).

[27] Wei, J., He, J., Zhou, Y., Chen, K., Tang, Z. and Xiong,

Z.: Enhanced Object Detection With Deep Convolutional Neural Networks for Advanced Driving Assistance, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21, No. 4, pp. 1572–1583 (2020).

[28] Wu, J., Liu, S., Huang, D. and Wang, Y.: Multi-Scale Positive Sample Refinement for Few-Shot Object Detection, *ECCV* (2020).

[29] Xiao, Y. and Marlet, R.: Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild, *ECCV* (2020).

[30] Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X. and Lin, L.: Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning, *ICCV* (2019).

[31] Zhang, H., Zhang, P., Hu, X., Chen, Y.-C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.-N. and Gao, J.: GLIPv2: Unifying Localization and Vision-Language Understanding, *NeurIPS* (2022).