教師情報に含まれるノイズに堅牢な弱教師あり領域分割手法

下田　和[1,a)]　柳井　啓司[1]

1. 概要

領域分割においては，学習画像における物体のカテゴリごとにピクセル単位の領域の教師情報が必要であり，教師情報の付与は大きなコスト，時間を要する。一般に，高度な教師情報を必要とする手法を完全教師あり学習領域分割，画像における物体のカテゴリ情報のみから学習する手法を弱教師あり学習領域分割と呼ぶ。弱教師あり学習による領域分割が可能となれば，大幅な学習データを収集するためのコストの削減が可能である。本研究においては，弱教師あり領域分割の精度向上を目指す。近年の弱教師あり領域分割においては，ノイズを含む教師情報を用いて領域分割モデルの学習を行うのが主流になっている。そこで，本研究ではノイズに対して堅牢な弱教師あり領域分割手法の提案を行う。

2. 導入

近年，深層学習の進歩とともに画像の領域分割制度は飛躍的に向上している。しかしながら，領域分割の学習に必要な教師情報のアノテーションコストの高さなどから実用面において課題が残っている。領域分割の教師情報のアノテーションコストを下げる手段の一つとして，弱教師あり学習による領域分割が注目を集めている。領域分割における弱教師あり学習では，教師情報として画像のクラスラベルのみを用いる。クラスラベルのアノテーションは領域のアノテーションと比較して大幅に安価であり，弱教師あり学習による領域分割が可能であれば，大きなコスト削減が期待できる。また，Web 画像における検索クエリをクラスラベルとみなせば，クラスラベル付きの画像は自動で収集することも可能である。

近年の弱教師あり領域分割は二つの段階にわけて行われている。まず，学習画像の大まかな領域分割を行い，その後にその大まかな領域分割結果を活用しより詳細な領域分割ネットワークを学習する。弱教師あり領域分割において，大まかな領域を推定する方法としては，画像の識別結果の可視化や前景領域の抽出があるが，近年はこれら両方を用いる手法が広く使われている。特に，近年はクラス分類の可視化方法として Class Activation Mapping(CAM) が高い精度と汎用性から広く用いられている。Global Average Pooling(GAP) を用いてクラス分類ネットワークを学習することで，物体の大まかな可視化を行うことができる。詳細な領域分割ネットワークの学習においては，完全教師あり領域分割において高い精度を達成しているネットワークが用いられる。また，さらなる精度の改善方法として，CRF による色特徴を用いた改善方法，領域の類似度を学習する改善方法などがある。

現在の弱教師あり領域分割と完全教師あり領域分割の違いには，教師情報にノイズが含まれているという点がある。弱教師あり領域分割の精度改善においては，教師情報におけるノイズの減少，ノイズに堅牢な学習方法の 2 点が重要となる。本手法ではノイズに堅牢な学習に着目した。ノイズに堅牢な学習手法としてはアンサンブルが有効であることが広く知られている。本手法においては，異なる学習データのモデルの出力によるアンサンブルによりノイズに堅牢な学習を目指した。これにより学習画像の不足が生じるが，Web 画像を用いることでこれを補った。また，Web 画像と対象のデータセットのドメインの違いから生じる変化に対応するために，広域の認識結果と局所領域の認識結果との一貫性からの学習を行った。以下に本研究のコントリビューションを示す。

- 弱教師あり領域分割における異なる学習データのモデルの出力によるアンサンブル手法の提案。
- Web 画像を用いた弱教師あり領域分割の学習データ不足の改善。
- 広域の認識結果と局所領域の認識結果との一貫性からの学習手法の提案。

3. 関連研究

3.1 クラス分類器の可視化手法

認識結果の可視化においては，画像におけるクラス分類に寄与した領域を推定する。クラス分類に寄与した領域と領域分割における対象領域との間には相関があり，認識結果の可視化は弱教師あり領域分割の手法として活用できる。Zeiler ら [16] は畳み込み演算と同じ学習パラメータによる逆畳み込み演算と逆 Pooling により，出力を入力空間に戻した際に，物体の位置に対応するピクセルが強く応答する

[1] 電気通信大学
[a)] shimoda-k@mm.inf.uec.ac.jp

ことを示した。Simonyan ら [12] は，Zeilar らと類似した手法で特定のクラスについての信号を逆伝搬させることで，CNN の認識結果に対するクラス応答を可視化させた。CNN の可視化結果から弱教師あり領域分割が可能であることを示した。派生手法に Guided Backpropagation [13] がある。

3.2 CNN の Activation を活用した弱教師あり領域分割

Oquab ら [5]， Pathak ら [8] は FCN の最終層に Global Pooling(GP) を用いることで，出力マップをクラス分類の CNN と同じ次元に変換し，画像ラベルのみを用いて FCN を学習させた。GP により学習させた FCN は大まかな物体の位置を推定することが可能であり，逆伝搬による可視化とは異なる形で弱教師あり領域分割を実現した。その後，Pinheiro ら [9] や Zhou ら [17] により，Global Pooling の派生手法についても研究がなされている。また，Chen ら [6]，Pathak ら [7] は Global Pooling を用いずに，弱教師ありで FCN の出力を直接学習させた。FCN の出力の各ピクセルが画像ラベルに含まれていないクラスを出力している場合には修正を加えるような形で，FCN の出力と画像ラベルから動的に領域の教師情報を生成し， これを用いて FCN の学習を行った。一方で, [11] の手法では FCN を GP で学習した認識結果を可視化した。また，Simonyan ら [12] による Backward の可視化手法をマルチクラスの画像に適応させた。においては，シングルクラスの場合に有効であり，マルチクラスの場合に極端に精度が下がるという問題があったが，各クラスの逆伝搬値について差分をとることでマルチクラスにも応用可能であることを示した。類似研究として，Jianming ら [3] による研究がある。

3.3 領域分割結果の再学習手法

Wei ら [15] は低次特徴量による物体顕著性マップを用いて学習画像の領域分割を行い，その領域分割結果を領域の教師情報として再学習を行った。Wei らの手法は単純ながら既存の弱教師あり領域分割の精度を大きく上回った。Wei らの手法は単純ながら既存の弱教師あり領域分割の精度を大きく上回り，弱教師あり領域分割の研究に大きな変化が生まれた。特に，その後の弱教師あり領域分割において，学習画像の領域分割手法（カテゴリは既知である画像の領域分割手法），ノイズを含む領域の教師情報について堅牢な学習手法の重要性が増した。Chen ら [6]，Pathak ら [7] は画像ラベルを活用して教師情報を動的に生成していたが，その方向性が強くなったという見方もできる。Kolesnikov ら [4] は，Global Pooling [17] による学習画像の大まかな位置推定結果について再学習を行った。また，Kolesnikov らは多くの領域の評価をスキップすることによるノイズに堅牢な学習，KL-Divergence による CRF による平滑化結果との近似などを用いて高精度な弱教師あり領域分割を達成した。Saleh ら [10] は特徴マップについて CRF を適用した結果から再学習を行い高精度を達成した。Tokmakov ら [14] は動画から得られるモーションの Segmentaion と Gaussian Mixture Model における動画のフレームの Segmentation 結果を用いて領域分割結果の再学習を行った。本手法においては, 物体の領域と相関のある情報の活用ではなく, 一貫性による教師なしのアプローチに取り組んだ。

4. 提案手法

4.1 Class Activation Mapping

本手法においては，クラス分類の可視化方法として Class Activation Mapping(CAM) 用いる。Class Activation Mapping においては，Global Average Pooling(GAP) を用いてクラス分類ネットワークを学習することで，物体の大まかな可視化を行うことができる。CAM においては，クラス分類モデルの出力マップを GAP によりベクトルに変換し，これを全結合層により分類する識別器を学習する。テスト時に，識別器の出力に対して GAP を行う前に全結合層の重みをかけることで，出力マップにおける各クラスの分布マップを得ることができる。本手法では，この出力マップにおける平均の 2 倍値を閾値として，各クラスの出力マスクとした。

4.2 異なる学習データのモデルの出力によるアンサンブル

セクション 4.1 で得られた大まかな物体の領域を用いて，領域分割モデルを学習する。しかし，この際に用いる教師情報は多くのノイズを含んでおり，そのままこれを教師情報として学習しても十分な精度を得ることは難しい。そこで，本研究では異なる学習データのモデルの出力によるアンサンブルによりノイズに堅牢な学習を目指す。図 1 に提案手法の概要を示す。全体の学習データを S とする。このとき，この学習データを n 個にわけた場合におけるサブセットを $S = (S_0, S_1, ..., S_n)$ とする。この各サブセットにつき，それぞれ領域分割用の子モデルを用意する。各サブセットを教師情報として学習するそれぞれの領域分割モデルのパラメータを $(\theta_0, \theta_1, ..., \theta_n)$，また，それぞれのモデルから得られる出力から生成したマスクを $(m_0, m_1, ..., m_n)$ とする。本手法においてはこの異なるパラメータから得られたマスクについてアンサンブルを行う。データ全体に対しての親モデルのパラメータを θ，出力を $p(\theta)$ とする。各マスク m_i についての segmentaton loss を $\mathrm{s\_loss}_i = \mathrm{seg\_loss}(p(\theta), m_i)$ とする。ただし，クラスラベルのセットを C, y は確率分布 p におけるクラスである。このとき，生成マスクのアンサンブルにより得られるロスは以下の式になる。

$$\mathrm{s\_loss} = \frac{1}{n}\sum_{i}^{n} \mathrm{s\_loss}_i \qquad (1)$$

このアンサンブルにより得られるロスの最適化によりノイズに堅牢な学習が期待できる。

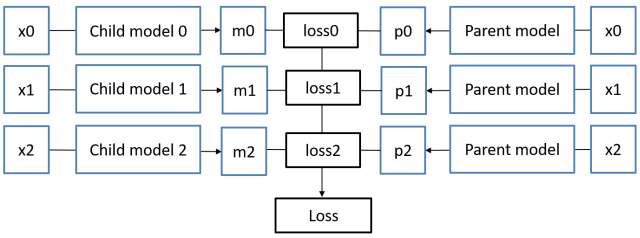

図 1 異なる学習データのモデルの出力によるアンサンブルについての概要。

4.3 Web 画像による学習データの補間

学習データをサブセットに分ける場合，各モデルの学習データ数が $1/n$ になってしまうという問題がある。そこで本研究では Web 画像を用いて全体の学習データを増やした。各クエリで収集した Web 画像についてクエリをラベルとしてこれを弱教師ありデータとして扱った。Bing API，Flickr API を用いて，Web 画像を収集した。本研究では Pascal VOC dataset における 20 クラスの物体についてそれぞれ約 5000 枚ずつ，計 10 万枚の画像を収集した。

4.4 広域領域の認識結果と局所領域の認識結果の一貫性の学習

Web 画像は弱教師あり画像として扱うことができるが，Pascal VOC dataset などの実世界画像におけるデータセットとは異なる部分がある。特に，Web 画像は検索クエリをラベルとして扱っているために，各画像につきラベルは一つである。一方で，Pascal VOC のような実世界画像データのデータセットにおいては複数ラベルも含まれている。また，物体のサイズも大きく異なっている傾向がある。このようなドメインの違いにより生まれる変化により，Web 画像を用いることで直接的に画像枚数の上昇を精度向上につなげることは難しい。そこで，本研究では Web 画像と Pascal VOC データにおけるドメインの違いによる精度の低下を，広域領域の認識結果と局所領域の認識結果の一貫性の学習により抑制した。本手法の概要を図 2 に示した。

各入力画像を x, ランダムにクロップを行いスケールを変化させた画像を x_s とする。本研究ではこれらの入力における出力で共通している領域の出力が同一になるように学習を行う。本手法における広域領域の認識結果と局所領域の認識結果の一貫性の学習の概要を図に示した。x についての出力を $p(x)$, x_s についての出力を $p(x_s)$ とする。$p(x)$ を $p(x_s)$ の出力について対応をとったマスクを $m(x|x_s)$，$p(x_s)$ を $p(x)$ の出力について対応をとったマスクを $m(x_s|x)$ とする。このとき，一貫性のロス c_loss を以下の式で定義する。

$$\begin{aligned} \text{c\_loss} = &\ \text{seg\_loss}(p(x), m(x_s|x)) \\ &+ \text{seg\_loss}(p(x_s), m(x|x_s)) \end{aligned} \tag{2}$$

ただし，対応のない領域については Ignore label として評価を行わない。また，Pascal VOC の画像における物体のサイズが小さいことを考慮し，Web 画像のスケール変化はパディングによる縮小化，Pascal VOC 画像のスケール変化はクロップによる拡大を行う。これにより，物体のスケールの違い，複数ラベル物体の画像などドメインの違いによる過学習を抑制することが期待できる。

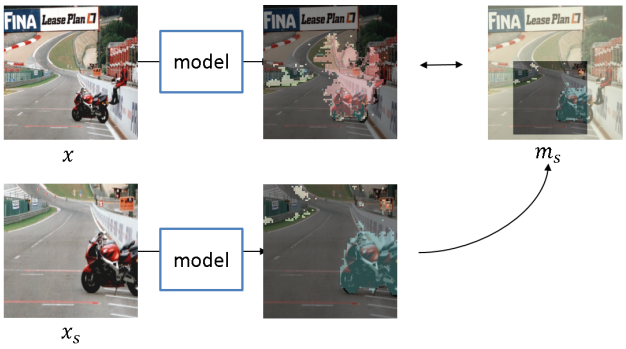

図 2 広域領域の認識結果と局所領域の認識結果の一貫性の学習についての概要。

5. 実験

領域分割のベンチマークのデータセットとして PASCAL VOC 2012 [1] を用いた。PASCAL VOC 2012 における領域分割ベンチマークでは, 20 の異なるクラスと Back Ground クラスを含む 21 のクラスの領域分割の精度を比較する。また, 1464 枚の train 画像, 1449 枚の validation 画像, 1456 枚の test 画像があるが, 近年は [2] により提供された 10582 枚の train_aug 画像を用いるのが一般的となっている。本実験においてもこれを学習画像として用いた。また追加画像として Web 画像 10 万枚を追加し、これも学習画像として用いた。本手法における領域分割の出力結果の例を図 3 に示す。

6. 結論

本手法においては，弱教師あり領域分割における異なる学習データのモデルの出力によるアンサンブル手法の提案を行った。上記の手法により学習データの不足が生じたが，これを Web 画像を用いて補完した。また，広域の認識結果と局所領域の認識結果との一貫性を用いることで Web 画像と Pascal VOC data のドメインの違いから生じる精度低下を抑えた。今後の課題としては異なる学習データのモデルの出力によるアンサンブルを行うためのモデルの省メモリ化，サリエンシーなどを活用した Seed 値の改善などが考えられる。

図 3 Pascal VOC validation set における領域分割結果の例。 1 3 段目については成功例、4 段目については失敗例を示した。

謝辞

本研究は JSPS 科研費 17J10261 の助成を受けたものです。

参考文献

[1] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective, *International Journal of Computer Vision*, Vol. 111, No. 1, pp. 98–136 (2015).

[2] Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S. and J., M.: Semantic contours from inverse detectors, *ICCV* (2011).

[3] Jianming, Z., Zhe, L., Jonathan, B., Xiaouhui, S. and Sclaroff, S.: Top-down Neural Attention by Excitation Backprop, *ECCV* (2016).

[4] Kolesnikov, A. and H.Lampert, C.: Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation, *ECCV* (2016).

[5] Oquab, M., Bottou, L., Laptev, I. and Sivic, J.: Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks, *CVPR* (2014).

[6] Papandreou, G., Chen, L.-C., Murphy, K. and Yuille, A. L.: Weakly-and semi-supervised learning of a dcnn for semantic image segmentation, *ICCV* (2015).

[7] Pathak, D., Krahenbuhl, P. and Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation, *ICCV* (2015).

[8] Pathak, D., Shelhamer, E., Long, J. and Darrell, T.: Fully convolutional multi-class multiple instance learning, *ICLR* (2015).

[9] Pedro, P. and Ronan, C.: From Image-level to Pixel-level Labeling with Convolutional Networks, *CVPR* (2015).

[10] Saleh, F., Akbarian, M., Salzmann, M., Petersson, L., Gould, S. and M.Alvares, J.: Built-in Foreground/Background Prior for Weakly-Supervised Semantic Segmentation, *ECCV* (2016).

[11] Shimoda, W. and Yanai, K.: Distinct Class Saliency Maps for Weakly Supervised Semantic Segmentation, *ECCV* (2016).

[12] Simonyan, K., Vedaldi, A. and Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps, *ICLR WS* (2014).

[13] Springenberg, J. T., Dosovitskiy, A., Brox, T. and Riedmiller, M.: Striving for Simplicity: The All Convolutional Net, *ICLR WS* (2015).

[14] Tokmakov, P., Alahari, K. and Schmid, C.: Weakly-Supervised Semantic Segmentation using Motion Cues, *ECCV* (2016).

[15] Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M., Zhao, Y. and Yan, S.: STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation, *ECCV* (2016).

[16] Zeiler, M. and Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning, *ICCV* (2011).

[17] Zhou, B., Khosla, A., Lapedriza, A. and Oliva, A. Torralba, A.: Learning Deep Features for Discriminative Localization, *CVPR* (2016).