弱教師あり領域分割のための一貫性に基づく学習画像の領域分割容易性推定

下田　和[1,a)]　柳井　啓司[1]

1. 導入

深層学習においては，高精度な認識を実現するために膨大な教師付き画像が必要であり，認識対象の拡張などの面で大きな障害となっている。特に，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。

近年, 弱教師あり学習による領域分割結果を教師情報として領域分割用のネットワークを学習することでより高い精度で領域分割が可能であることが明らかになっている [21]。これはノイズを含む教師情報について CNN を活用した EM アルゴリズムの一種であると捉えることができる。この領域分割モデルの繰り返し学習手法は, 初期の教師情報に含まれる誤りの傾向に一貫性があれば, その誤りの傾向も学習してしまうという欠点がある。

本手法は推定結果の一貫性を活用することに着目し, 初期の教師情報の誤りに頑強な領域の再学習を行う。本研究の動機は, 人間が答えの不明瞭な場合に判断をくだすときに, 一貫性を用いて判断の根拠とすることに基づいている。Amazon Mechanical Turk による教師情報の付与において複数のワーカーに仕事を割り振り, ワーカーの選択が重複している場合に決を取る場合があるが, これは結果の一貫性を用いて教師情報を付与しているといえる。また, アンケートによる世論調査なども複数の意見の一貫性を調べていると考えることができる。一貫性による教師情報の妥当性の検証は実世界で広く使われていることがわかる。CNN の学習においても, 教師情報が曖昧である場合に一貫性を用いることで, 結果の改善が可能であることが期待できる。そこで, 本研究においては識別器の推定結果の一貫性を活用することで弱教師における領域分割の精度向上について検証した。以下に本研究の Contribution を示す。

- Backward における認識結果の可視化と領域の再学習を組み合わせることで高精度な弱教師あり領域分割を行った。
- 識別器の推定結果の一貫性から領域分割結果の精度を推定した。
- 領域分割結果精度の推定結果から結果のよいものを用いて画像の水増しを行い精度を向上させた。

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

2. 関連研究

2.1 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。Zeiler ら [22] 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ことを示した。Simonyan ら [17] 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。派生手法に Guided Backpropagation [19] がある。

2.2 CNN の Activation を活用した弱教師あり領域分割

Oquab ら [10], Pathak ら [13] 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。また，Chen ら [11]，Pathak ら [12] 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。一方で, [16] 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 [17] 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。

2.3 領域分割結果の再学習手法

Wei ら [21] 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。Kolesnikov ら [7] は，Global Pooling [23] 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。Saleh ら [15] 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 [20] 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。

3. 提案手法

近年, 領域の再学習により弱教師あり領域分割精度が向上している [21]。領域分割の再学習においては、以下の手順により領域分割のモデルを学習する。

- 既存の弱教師あり領域分割手法を用いて、学習画像について領域分割を行う。
- 学習画像の領域分割結果を領域の教師情報として、領域分割の識別モデルを学習する。

本手法においては、領域分割結果を再学習するうえで、より高い精度で再学習を行う方法として、学習画像を領域分割した際の精度を推定し、領域分割の容易な画像を活用するというアプローチを行った。

3.1 領域分割結果の精度の推定

本研究においては, 可視化結果が物体の領域を反映しているかどうかを, 可視化結果の一貫性から推定した。

3.1.1 差分による変化の一貫性

一般に, 可視化結果は物体の顕著性マップのようになり, 直接領域の推定を行うことは難しかったが, 逆伝搬値について差分をとることによりクラス応答を鮮明にすることが可能である [16]。本研究では, この逆伝搬値について差分をとった場合の変化に着目した。差分をとった際に大きく結果が変化する場合には以下のような原因が考えられる。

- クラス分類に失敗し, クラス応答の勾配が消失した
- 対象のクラスの識別とは別に, 顕著性のある物体が存在する

つまり, 差分をとった場合に変化が小さい場合には, クラス分類が容易であり, 顕著性マップと領域分割結果が一致していることが期待できる。そこで, 本研究では, 差分をとった際の変化から, 領域分割結果の精度を推定した。画像 I における, 差分なしによる領域分割結果を $V_o(x)$ 差分ありによる可視化領域分割結果を $V_w(x)$ としたとき, その領域分割結果の信頼度 $R_{sub}(x)$ を以下の式で定義する。

$$R_{sub}(x) = \sum_{c \in C} IoU(V_o^c(x), V_w^c(x)) \tag{1}$$

このとき, $IoU(.,.)$ は各領域の Intersection over union(IoU) を返す関数であるとする。

3.1.2 入力サイズの変化における一貫性

識別結果の可視化においては, 画像の入力サイズを大きくした場合には局所的な認識の可視化結果, 小さくした場合には大局的な認識の可視化結果となる傾向がある。もし, 入力画像のサイズを変化させて得られる可視化結果に一貫性があれば, 領域分割が容易であり, 高い精度で領域分割を行えていると期待できる。本研究ではこれを二つ目の領域分割の精度推定の評価指標とした。$s_n = 320, 416, 512 (n = 0, 1, 2)$, を入力サイズ, それぞれのサイズにおいてクラス応答の差分により得られる領域分割結果を $V_{s_n}(x)$ とする。このとき, 入力サイズの変化における一貫性 $R(I)_{size}$ を以下の式で定義する。

$$R_{size}(x) = \frac{1}{C} \sum_{c \in C} IoU(V_{s_0}^c(x), V_{s_1}^c(x), V_{s_2}^c(x)) \tag{2}$$

このとき, $IoU(.,.,.)$ は 3 つの領域の Intersection over Union(IoU) を返す関数であるとする。最終的な領域分割の信頼度は以下の式で計算した。

$$score = \lambda_1 \cdot R_{sub}(x) + \lambda_2 \cdot R_{size}(x) \tag{3}$$

本研究においては, 単純に $\lambda_1 = 0.5, \lambda_2 = 0.5$ とした。図 1 に本手法における領域分割が容易であると推測された例, 容易でないと推測された例を示した。

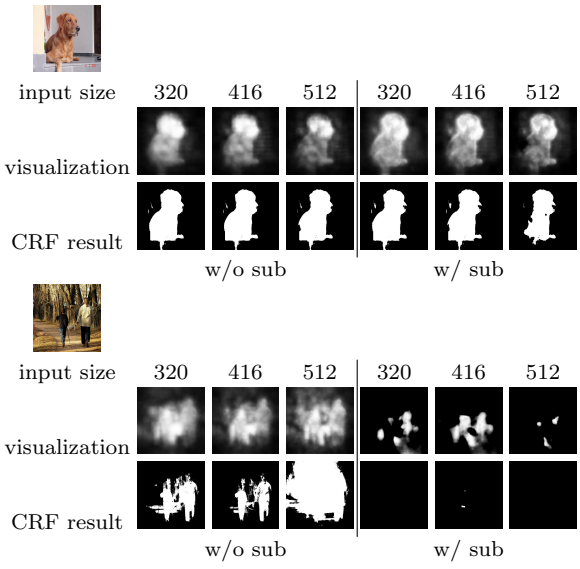

図 1 可視化結果の一貫性による領域分割精度の推定例

3.2 領域の推定結果の一貫性を活用した領域評価

3.1 節においては画像の領域分割の容易性を評価したが, 本節においては, 画像の各領域の信頼性評価する。Kolesnikov ら [7] は領域の再学習を行う際に GAP のアテンションにおける強く反応している極一部分のみの領域を教師情報として学習を行うことで高精度を達成した。本手法においては, これを応用し 3.1 節における領域分割共通部分, 一貫している領域を用いた。$V_o(x), V_w(x), V_{s_n}(x)$ に

おける共通領域のみを学習中に評価し, その他の領域については評価をスキップした。また, 本研究においては, 背景領域の影響が強く出る傾向があったために, 背景領域は, 2 クラス識別器における可視化結果との共通領域のみを評価するというパターンについて検証を行った。図 2 に生成したマスクの例を示す。

図 2 それぞれ, (1 列目) 入力画像, (2 列目) マルチクラス識別器による可視化結果, (3 列目)2 クラス識別器によるマスク (3), (4 列目)(2)(3) の統合結果 (4) である。

3.3 容易な画像を活用した教師情報の水増し

入力画像を小さな変化を与えることによる教師情報の水増しが深層学習において有効であることは現在広く知られている。しかし, 物体検出において, より大きな変化を与えることで精度が大幅に向上していることなどが報告されている [8](論文におけるポスター発表)。本セクションでは, これを弱教師あり領域分割に流用する。特に, 3.1 節で推定した領域分割の容易な画像を用いて, 3.2 節で求めたマスク画像と学習画像のペアについて教師情報の水増しを行う。教師情報の水増しに用いる画像は, 領域分割の容易さの推定結果が閾値以上の物を用いた。

Liu[8] らは元画像とは大きく異なるスケールで Crop しデータの水増しを行うことで精度向上が可能であることを示した。しかし, 弱教師ありにおいてスケールを大きく変化させて Crop を行った場合, その画像内に対象の物体が含まれているか確かではない。そこで, 本研究では Crop された画像を可視化において用いたマルチクラス識別器で識別し, これが画像と対応する 3.2 節で生成した領域の教師情報と一致していた場合に学習に活用した。これを各画像について適用し, 最大で 10 枚の画像を水増しした。また, ランダムにパディングすることにより, 精度が向上したという報告もある。そこで, 本手法においてもこれを取り入れ, ランダムなスケール比でパディングを行い, 各学習画像について 10 枚の画像を水増しした。これらの教師情報の水増し例を図 3 に示す。

4. 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 [4] 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

図 3 Random crop, Random padding により生成した画像の例と対応する教師情報。

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 [5] 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。評価指標は Mean IoU を用い, ピクセルの一致度を評価する。ただし, 評価は一般に公開されている Pascal のサーバーを用いて行った。また、可視化を行う際には [16] と同様に VGG16 モデル [18] を用いた。再学習の際に用いる領域分割の識別モデルとしては VGG16 モデル [18] の派生である Deeplab モデル [3] を用いた。3.1 節における領域分割精度の推定結果を評価する。まず, 高い領域分割精度であると推定された画像で学習した識別器と, ランダムに選んだ画像で学習した識別器の領域分割精度を比較した。ただし, 選ぶ枚数は 3.1 節において計算した信頼度の閾値から決定した。表 1 に結果を示す。同じ枚数における結果では, 本手法による画像で推定した結果が一定して高い精度を達成していることがわかる。また, 表 2 に容易な画像を用いた教師情報の水増しによる結果を示した。単純な教師情報の水増しにより, 2 %程度精度が向上していることがわかる。表 3 は, 学習画像における枚数と水増しに用いる画像枚数を変化させた場合の結果である。全ての画像を使う場合より, 容易な画像を選択した場合のほうが精度が向上していることが確認できる。また, これらの結果は validation set における評価結果である。

表 1 領域分割精度推定の評価 (SI: Selected Image, RI: Random Image)

Image Number	SI	RI
730 (threshold=0.8)	44.7	42.6
2105 (threshold=0.7)	48.7	46.0
4235 (threshold=0.6)	48.9	46.1
6310 (threshold=0.5)	48.7	48.1
7841 (threshold=0.4)	49.2	48.5
8760 (threshold=0.3)	49.3	47.7
9337 (threshold=0.2)	49.0	48.6
10582 (all images)	49.4	49.4

表 2 教師情報の水増しによる精度変化

approach	mIoU
simple augmentation	49.4
+ random crop	51.1
++ random padding	51.3

表 3 学習に用いる画像枚数を変化させた場合の領域分割精度

Base image num	Augment image num	mIoU
8760 (th=0.3)	730 (th=0.8)	50.1
10582 (all)	730 (th=0.8)	48.9
8760 (th=0.3)	2105 (th=0.7)	**51.3**
10582 (all)	2105 (th=0.7)	49.9
8760 (th=0.3)	8760 (th=0.3)	49.7
10582 (all)	10582 (all)	48.8

表 4 に他の弱教師あり領域分割手法との比較を示した。本手法は, 画像ラベルのみを用いる手法において, もっと

表 4 Results on PASCAL VOC 2012 *test set*.

Methods	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	**mIoU**
Fully Supervised:																						
O2P [2]	85.4	69.7	22.3	45.2	44.4	49.6	66.7	57.8	56.2	13.5	46.1	32.3	41.2	59.1	55.3	51.0	36.2	50.4	27.8	46.9	44.6	47.6
SDS [6]	86.3	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5	51.6
FCN-8s [9]	-	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
Deeplab Large FOV [3]	92.6	83.5	36.6	82.5	62.3	66.5	85.4	78.5	83.7	30.4	72.9	60.4	78.5	75.5	82.1	79.7	58.2	82.0	48.8	73.7	63.3	70.3
Using Additional Supervision:																						
CCNN w/size [12]	-	42.3	24.5	56.0	30.6	39.0	58.8	52.7	54.8	14.6	48.4	34.2	52.7	46.9	61.1	44.8	37.4	48.8	30.6	47.7	41.7	45.1
One point[1]	80.6	50.2	23.9	38.4	33.1	38.5	52.0	50.9	55.4	18.3	38.2	37.7	51.0	46.1	54.7	43.2	35.4	45.1	33.0	49.6	40.0	43.6
F/B prior + CheckMask[15]	87.4	65.7	26.0	64.2	43.7	53.2	72.6	63.6	59.5	17.1	48.0	43.7	61.2	52.0	69.3	54.8	43.0	50.3	34.6	59.2	42.0	52.9
Weakly Supervised:																						
MIL-FCN [13]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.9
EM-Adapt [11]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	16.7	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	29.2	34.3	46.0	39.6
CCNN [12]	-	21.3	17.7	22.8	17.9	38.3	51.3	43.9	51.4	15.6	38.4	17.4	46.5	38.6	53.3	40.6	34.3	36.8	20.1	32.9	38.0	35.5
MIL-ILP-seg [14]	78.7	48.0	21.2	31.1	28.4	35.1	51.4	55.5	52.8	7.8	56.2	19.9	53.8	50.3	40.0	38.6	27.8	51.8	24.7	33.3	46.3	40.6
DCSM w/ CRF [16]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3	34.8	50.3	48.9	69.0	49.7	38.4	57.1	34.0	38.0	40.0	45.1
F/B prior[15]	80.3	57.5	24.1	66.9	31.7	43.0	67.5	48.6	56.7	12.6	50.9	42.6	59.4	52.9	65.0	44.8	41.3	51.1	33.7	44.4	33.2	48.0
STC [21]	85.2	62.7	21.1	58.0	31.4	55.0	68.8	63.9	63.7	14.2	57.6	28.3	63.0	59.8	67.6	61.7	42.9	61.0	23.2	52.4	33.1	51.2
SEC [7]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
Ours	82.7	63.5	28.9	60.2	29.7	60.0	71.6	61.9	62.9	19.6	45.8	47.5	47.0	56.1	70.6	58.3	31.7	62.4	38.7	36.8	38.3	51.2
Ours + Data Augmentation	83.0	67.5	29.7	69.7	28.8	59.7	71.2	66.4	69.8	18.6	49.8	44.7	49.4	60.5	73.5	61.8	32.7	62.7	39.0	34.3	36.5	52.8

も高い精度を達成している。特に、領域の再学習を行っている F/B prior[15]、STC [21]、SEC [7] と比較しても高い精度を達成した。また、本手法は CNN を用いない完全教師あり領域分割手法である O2P [2]、プロポーザルと CNN を組み合わせた完全教師あり領域分割手法である SDS [6] の精度を上回った。実験は Pascal VOC test set におけるものであり, 評価に用いた画像は他の表における結果のものと異なっている。図 4 は本手法の領域分割結果の例である。

5. 結論

本手法においては, 推定結果の一貫性を活用することで, 弱教師あり領域分割結果の精度を推定し, 精度のよいものを学習画像として再学習することで, 教師情報の誤りに頑強な領域の再学習を行った。これにより, 既存の画像ラベルのみを用いた弱教師あり領域分割の精度を上回っている。また、本手法においては領域分割結果の精度を推定したが、領域分割精度が高く推定したもの、低く推定したものを異なる方法で学習して組み合わせるなど、今後も様々な形での拡張が期待できる。

図 4 Pascal VOC validation set における領域分割結果の例。

参考文献

[1] Bearman, A., Russakovsky, O., Ferrari, V. and Fei-Fei, L.: What ' s the Point: Semantic Segmentation with Point Supervision, *ECCV* (2016).

[2] Carreira, J., Caseiro, R., Batista, J. and Sminchisescu, C.: Semantic Segmentation with Second-Order Pooling, *ECCV* (2012).

[3] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. and L., Y. A.: Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, *ICLR* (2015).

[4] 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).

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

[6] Hariharan, B., Arbeláez, P., Girshick, R. and Malik, J.: Simultaneous Detection and Segmentation, *ECCV* (2014).

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

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y. and Berg, A. C.: SSD: Single Shot MultiBox Detector, *ECCV* (2016).

[9] Long, J., Shelhamer, E. and Darrell, T.: Fully Convolutional Networks for Semantic Segmentation, *CVPR* (2015).

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

[11] 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).

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

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

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

[15] 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).

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

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

[18] Simonyan, K., Vedaldi, A. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *ICLR* (2015).

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

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

[21] 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).

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

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