# Attention Mining Branchを用いたアテンションマップの最適化と高精度化

岩吉 孝明† 足立 浩規† 平川 翼† 山下 隆義† 藤吉 弘亘†
†中部大学
E-mail: iwayoshi@mprg.cs.chubu.ac.jp

## 1 背景・目的

画像認識の分野において，畳み込みニューラルネットワーク (CNN)[1] は高い認識性能を実現している一方，複雑なネットワーク構造であることから推論時における判断根拠を人が解釈することは困難である．この問題に対して，判断根拠をアテンションマップとして可視化する視覚的説明 [2, 3, 4, 5, 6, 11, 12, 13, 14, 15, 16, 17] の研究が行われている．中でも，Class Activation Mapping (CAM) [5]，Gradient-weighted Class Activation Mapping (Grad-CAM) [6]，Attention Branch Network (ABN) [2] は視覚的説明の代表的な手法である．これらの手法はCNN の推論時における注視領域をアテンションマップとして可視化することができる．ABN は，アテンションマップとして可視化するだけでなく，認識処理に活用することで高精度な画像認識を実現した．しかしながら，ABN のアテンションマップは認識対象のみを注目するばかりではなく，対象物体以外を注目することや，もはや対象物体すら注目しないことがある．このような不要な領域が含まれるアテンションマップは，ABN の学習を困難にし，その結果，認識精度の低下を誘発するという問題がある．

Mitsuhara らは人の知見により修正したアテンションマップを用いて ABN を再学習することによって，この問題に対処した [3]．この手法は，ABN で誤認識が発生した際のアテンションマップを人の知見を介してすべて修正し，ネットワークを再学習するが，アテンションマップの修正コストが非常に高い．一方，鳥の種類のようなサブカテゴリを分類する詳細画像分類では，クラス間の相違が小さいため，より識別に有効な領域を注視して分類する必要がある．

本研究では，識別に有効な領域を捉えるアテンションマップを自動で獲得するために，1. 認識に有効な領域の探索，2. 特徴空間の改善，3. 出力空間の改善の 3 つのアプローチを提案する．具体的には，ABN に有効な注視領域を探索する Attention mining branch (AMB)，特徴空間を改善する Prototype conformity loss (PC Loss) と出力空間を改善する Complement objective training (COT) を導入する．評価実験により，提案手法により探索したアテンションマップが識別精度に貢献することを示す．

## 2 関連研究

本章では，視覚的説明に関する手法について述べる．

### 2.1 視覚的説明

視覚的説明モデルは，ネットワークの推論時における注視領域を可視化することができる．視覚的説明モデルの代表的な手法として Class Activation Mapping (CAM) [5]，Gradient-weighted Class Activation Mapping (Grad-CAM) [6]，Attention Branch Network (ABN) [2] がある．

CAM は 畳み込み層により獲得した特徴マップと各チャネルに対応する全結合層の結合重みを用いることでアテンションマップを獲得する手法である．CAM は畳み込み層と全結合層の間に Global Average Pooling[7] を行い，特徴マップを 1 つの値に縮約する．その後，全結合層により重みづけ付けして各クラスの出力を求める．この重みを縮約前の特徴マップに乗算し，足し合わせることによりあるクラスに対するアテンションマップを獲得している．

Grad-CAM は，特定クラスの出力層のユニットから勾配を算出して，アテンションマップを獲得する手法である．入力画像に対し，あるクラスのみの勾配を逆伝播する．その後，対象とする畳み込み層に対する各チャンネルの勾配の平均を特徴マップに重みづけし，足し合わせることによりアテンションマップを獲得している．

ABN は，Feature extractor，Attention branch，Perception branch から構成される．Feature extractor は入力画像から特徴マップを抽出する．これを Attention branch に与え，アテンションマップを獲得する．アテンションマップを特徴マップに乗算し，Perception branch に与えて認識結果を出力する．

### 2.2 人の知見により修正した ABN の再学習法

ABN を応用した研究として，人の知見により修正した ABN の再学習法がある．ABN により獲得した注視領域は，認識する物体以外に発生する場合や，認識対象の物体に発生しない場合がある．このような注視領域は，誤認識を誘発することがある．そこで，本手法で

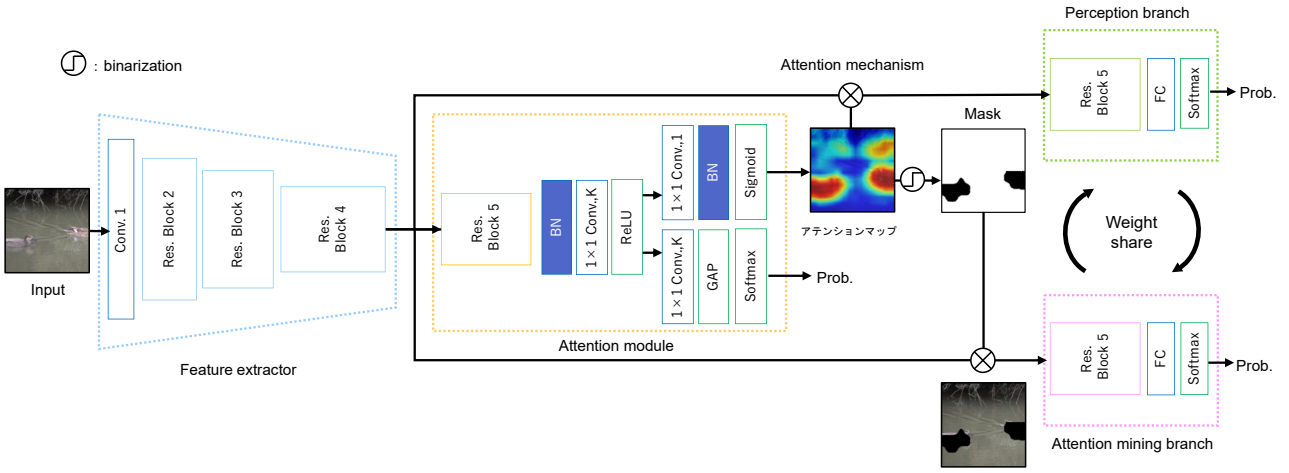



図 1　提案手法の構造

は，ABN で誤認識が発生した際のアテンションマップを人の知見を介してすべて修正する．その後，修正したアテンションマップとネットワークから出力されるアテンションマップとの誤差関数を ABN の誤差関数に追加し，再学習を行うことにより認識精度が向上する．また，本論文ではこの手法を人の知見として表記する．

### 2.3　アテンションのマイニング

アテンションを学習中にマイニングする手法として，Guided Attention Inference Network (GAIN)[4] がある．GAIN は，アテンションマップを用いた弱教師ありセマンティックセグメンテーション手法であり，高精度なセマンティックセグメンテーションを実現している．GAIN は Grad-CAM で獲得したアテンションマップから Mask を作成し，入力画像に適用することで注視領域を隠した画像を作成する．作成した画像をネットワークに入力し，正解クラスの確率を出力する．これが小さいほど Mask は認識対象の物体を隠せているといえる．よって，正解クラスの出力の総和を新たな損失とし，最小になるよう学習することで，対象物のみを注視するネットワークとなる．

## 3　提案手法

本研究では学習によって詳細画像分類に有効な注視領域を自動で獲得することを目的とする．提案手法は，ABN に 3 つのアプローチを導入し，詳細画像分類に有効な領域の獲得を目指す．1 つ目に，ABN に Attention mining branch (AMB) を導入し，認識に有効な領域のみに注視するよう学習する．2 つ目に，ABN に Prototype conformity loss (PC Loss) を導入して特徴空間を改善する．PC Loss は，特徴空間において同じクラスの特徴量を近づけ，異なるクラスの特徴量を離すように学習することから，クラス特有の注視領域の獲得を可能とする．更に，注視領域の改善により，認識精度の向上も見込めると考える．3 つ目に，AMB に不正解クラスを平坦化する Complement objective training (COT) を導入し，不正解クラス確率を抑えることで，認識対象のみの注視領域の獲得を目指す．

### 3.1　提案手法のネットワーク構造

提案手法は ABN に Attention mining branch を導入することでアテンションマップを自動で最適化する．Attention mining branch はセマンティックセグメンテーション手法である GAIN のアイデアを導入しており，end-to-end に学習することでアテンションマップの改善と認識精度の向上を実現する．提案手法の構造を図 1 に示す．図 1 に示すように，提案手法は，Feature extractor，Attention module，Perception branch，Attention mining branch で構成されている．Feature extractor は，入力画像に複数の畳み込み処理を施して特徴マップを獲得する．Attention module は，獲得した特徴マップをチャネル数が 1 になるよう畳み込むことでアテンションマップを獲得する．Perception branch は，Feature extractor の出力をアテンションマップに重みづけした特徴マップを用いて，最終的なクラス確率を出力する．Attention mining branch は，Perception branch 同様の構造をしており，注視領域を隠した特徴マップからクラス確率を出力する．また，アテンションマップを特徴マップに重みづけする Attention 機構にはバイパスを用いない．これにより，アテンションマップを学習に反映させやすくする．

### 3.2　Attention mining branch によるアテンションマイニング

Attention mining branch は，認識に有効な領域を獲得するように学習を行う．まず，Attention module で獲得したアテンションマップを用いて Mask を作成する．Mask は，アテンションマップに閾値を設け，2 値化することで生成する．これを，Feature extractor から出力された特徴マップに乗算することで，注視領域を

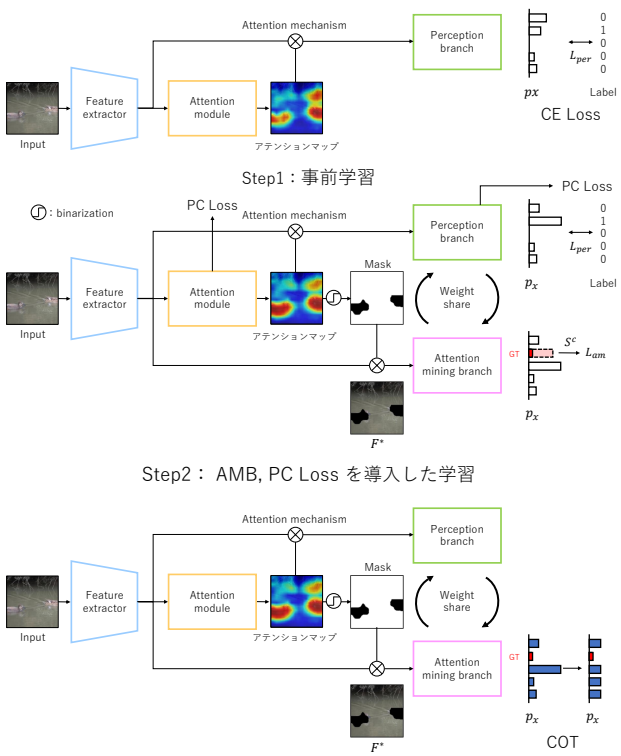



図 2　提案手法における学習の流れ

隠した画像を生成する．マスク処理した特徴マップ $F^*$ は，特徴マップを $F$，アテンションマップを $A$，Mask を $T$ とすると式 (1) のように表される．

$$F^* = F - (T(A) \odot F) \tag{1}$$

Attention mining branch は，Perception branch と同様の構造であり，マスク処理した特徴マップを入力し，クラス確率を出力する．このとき，対象クラスのクラス確率が低いほど，Mask が認識対象の物体領域を隠しているといえる．そこで，対象クラスのクラス確率を $L_{am}$ として，最小化するように学習することで，認識対象の物体のみを注視するようにアテンションマップを最適化する．$L_{am}$ は，正解クラスのインデックスを $g$，クラス確率を $S_i$，サンプル数を $n$ とすると式 (2) のように表される．

$$L_{am} = \sum_{i=1}^{n} S_{ig}(F^*) \tag{2}$$

また，Attention mining branch は，Perception branch と重みを共有する．これにより，認識対象の物体のみを注視するように学習した Attention mining branch の重みを最終的な認識結果を出力する Perception branch に反映することができる．

### 3.3　PC Loss による特徴空間の改善

特徴空間の改善として，PC Loss を Perception branch と Attention module に導入する．PC Loss は同じクラス内の特徴量を近づけ，異なるクラス間の特徴量を離す損失関数である．PC Loss $L_{PC}$ は，サンプル数を $N$, クラス数を $k$，サンプル $i$ の正解クラスを $g$，不正解クラスを $j$，学習可能なクラス重心 $w^c$ と特徴量 $f_i$，$\alpha_i = \|f_i - w_g^c\|_2$，$\beta_i = \|f_i - w_j^c\|_2$，$\gamma_i = \|w_g^c - w_j^c\|_2$，とすると，式 (3) で表現できる．

$$L_{PC} = \sum_{i}^{N} \{\alpha_i - \frac{1}{k-1} \sum_{j \neq g} (\beta_i + \gamma_i)\} \tag{3}$$

これを最小化することで，$\alpha_i$ は特徴量を正解クラスの重心に集め，$(\beta_i + \gamma_i)$ はクラス間を離すように働く．これにより，クラスごとに特有の特徴が獲得され，その結果，より良いアテンションマップを獲得できると考える．

### 3.4　COT による出力空間の改善

AMB の出力空間の改善として，Complement objective training (COT) を導入する．COT は Complement Entropy Loss が最小になるよう重みを更新することで，不正解クラス確率を平坦化する．Complement Entropy Loss $L_{CoE}$ は，$\hat{y}$ を入力 $x_i$ に対するクラス確率，$g$ を正解クラスのインデックスとすると，式 (4) のように表される．

$$L_{CoE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1, j \neq g}^{K} \frac{\hat{y}_{ij}}{1 - \hat{y}_{ig}} \log \left( \frac{\hat{y}_{ij}}{1 - \hat{y}_{ig}} \right) \tag{4}$$

AMB に COT を導入する際，$\hat{y}_{ig} = 0$ とすることで，正解クラス確率が 0，残りの不正解クラス確率が平坦になるよう算出される．これにより，注視領域に悪影響を及ぼす突出した不正解クラス確率を削減する．

### 3.5　学習手順

提案手法における学習の流れを図 2 に示す．提案手法の学習手順を以下に述べる．

**step1**

ABN と同様な構造のモデルを用いて事前学習を行う．事前学習には，提案手法に用いるデータセットと同様のものを使用する．

**step2**

ABN の損失に Attention module，Perception branch から算出した PC Loss $L_{PC}$，AMB の損失 $L_{am}$ を加え，最小になるよう学習する．COT を AMB に導入する場合は，AMB の損失 $L_{am}$ は用いない．損失 $L$ は，Attention module の出力と正解クラスのクロスエントロピー誤差を $L_{att}$，perception branch の出力と正解クラスのクロスエントロピー誤差を $L_{per}$，とすると，式 (5) のように表わされる．

$$L = L_{att} + L_{per} + L_{PC} + L_{am} \tag{5}$$

**step3**

AMB の出力から Complement Entropy Loss $L_{CoE}$ を算出し，$L_{CoE}$ が最小になるよう学習する．

**step4**
Step2,3 をイタレーションごとに繰り返す.

## 4 評価実験

提案手法の有効性を評価するために詳細画像識別タスクにて評価実験を行う.

### 4.1 実験条件

本実験では，Caltech-UCSD Birds 200-2010 (CUB-200-2010) データセット [8] と，Stanford Dogs データセット [9] を用いる．ベースネットワークとして ResNet-50 [10] を使用し，バッチサイズは 16 とする．Mask の閾値は CUB-200-2010 データセットで 0.78，Stanford Dogs データセットで 0.40 とする．$L_{am}$ の重みは 0.0001 に設定する．学習の更新回数は ABN の事前学習，提案手法それぞれで 300epoch とする．また，提案手法との比較として，ABN を用い，CUB200-2010 データセットにおいては人の知見も用いる.

### 4.2 定量的評価

CUB200-2010 データセットにおける認識精度の比較を表 1 に示す．表 1 より，CUB-200-2010 データセットにおいて，AMB は，ABN より分類精度が向上するが，人の知見より低い．PC Loss もしくは COT を導入すると，人の知見よりも分類精度が向上することを確認した．また，PC Loss と COT の両者を AMB に組み合わせることで分類精度が更に向上しており，ABN より Top-1 の分類精度が 14.41 ポイント高い結果となることをことを確認した.

Stanford Dogs データセットにおける認識精度の比較を表 2 に示す．表 2 より，Stanford Dogs データセットにおいても，提案手法は ABN より，認識精度が向上することが確認できる．ここで，PC Loss のみを AMB に組み合わせることで分類精度が最も向上しており，ABN より Top-1 の分類精度が 1.96 ポイント高い結果となることをことを確認した．また，PC Loss と COT の両者を AMB に組み合わせることで Top-5 の分類精度において最も向上しており，ABN より Top-5 の分類精度が 1.44 ポイント高い結果となることをことを確認した．これらのことから，提案手法が ABN の性能を大幅に向上させることが確認できた.

表 1　認識精度の比較 (CUB-200-2010) [%]

| | AMB | PC Loss | COT | Top-1 acc. | Top-5 acc. |
|---|---|---|---|---|---|
| ABN | - | - | - | 31.68 | 57.01 |
| 人の知見 | - | - | - | 37.42 | 62.08 |
| 提案手法 | ✓ | - | - | 33.33 | 58.56 |
| | ✓ | ✓ | - | 45.10 | **71.68** |
| | ✓ | - | ✓ | 39.76 | 66.57 |
| | ✓ | ✓ | ✓ | **46.09** | 69.24 |

表 2　認識精度の比較 (Stanford Dogs) [%]

| | AMB | PC Loss | COT | Top-1 acc. | Top-5 acc. |
|---|---|---|---|---|---|
| ABN | - | - | - | 71.81 | 93.02 |
| 提案手法 | ✓ | - | - | 71.99 | 92.80 |
| | ✓ | ✓ | - | **73.95** | 94.37 |
| | ✓ | - | ✓ | 73.59 | 93.89 |
| | ✓ | ✓ | ✓ | 73.46 | **94.46** |

### 4.3 アテンションマップの可視化結果

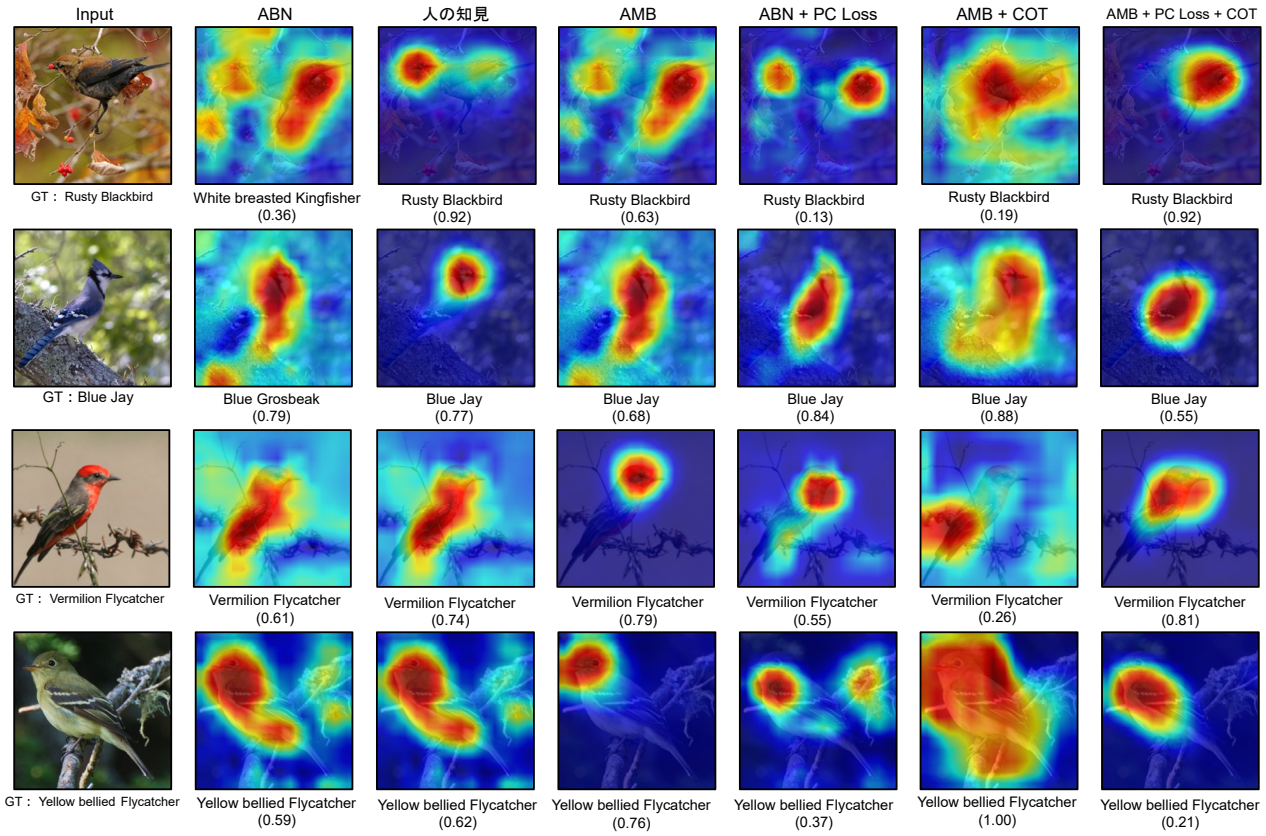
CUB-200-2010 データセットにおける各手法のアテンションマップの比較を図 3 (a) に示す．アテンションマップの下に，モデルが認識したクラスとそのクラス確率を示す．図 3 (a) に示すように，提案手法により ABN で獲得していた認識対象以外の領域，すなわち不要な注視領域が軽減されていることが分かる．また，PC Loss を導入することで，ABN より局所的な領域に着目していることが分かる．一方で，AMB，COT は認識対象を広域に捉えつつ，不要な注視領域を軽減していることが分かる．また，COT は PC Loss より高いクラス確率を獲得していることから，PC Loss では捉えられていない重要な領域を獲得できていると考えられる.

Stanford Dogs データセットにおける 各手法のアテンションマップの比較を図 3 (b) に示す．図 3 (b) に示すように，Stanford Dogs データセットにおいても，提案手法により ABN で獲得していた不要な注視領域が軽減されていることが分かる．また，PC Loss と COT の両者を AMB に組み合わせることで，PC Loss のみを AMB に組み合わせた際の局所的な領域を獲得しつつ，COT のみを AMB に組み合わせた際の広域な領域も注視していることがわかる.

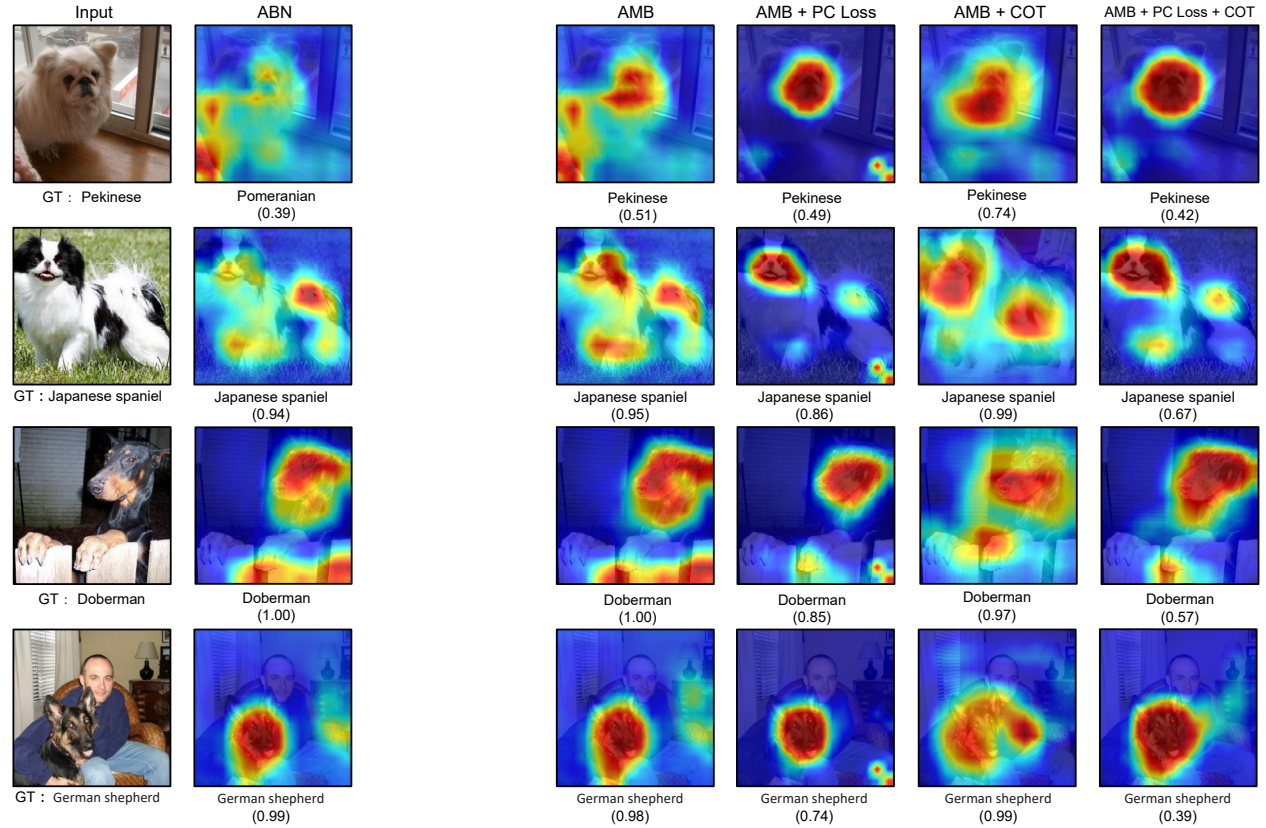
### 4.4 アテンションマップの有効性の評価

提案手法により獲得した Attention の有効性を定量的に評価する．評価方法として Insertion, Deletion を [11] を行う．Insertion，Deletion の例を図 4 に示す．アテンションの閾値ごとに認識精度を確認し，Area under carve (AUC) による評価を行う．Insertion はアテンションが閾値より高い領域を用いて評価を行う．そのため，Insertion は 1 に近いほど良い注視領域といえる．一方，Deletion はアテンションが閾値より低い領域を用いて評価を行う．そのため，Deletion は 0 に近いほど良い注視領域といえる．また，Insertion 結果から Deletion 結果を差し引いたものを最終的なスコアとして評価を行う．表 3 に CUB200-2010 データセットにおける Insertion, Deletion 結果を示す．表 3 に示すように，提案手法により，ABN や人の知見と比べスコアが向上しており，認識に有効な領域を捉えることができた．また，PC Loss と COT を AMB に組み合わせた場合に，スコアが最も高い精度となった.

表 4 に Stanford Dogs データセットにおける Inser-




(a) CUB200-2010




(b) Stanford Dogs

図 3　アテンションマップの可視化例

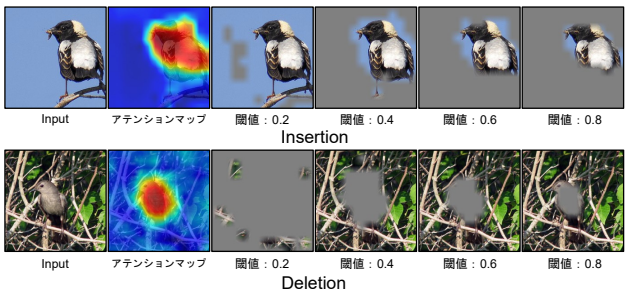



図 4　Insertion, Deletion の例

tion，Deletion 結果を示す．表 4 に示すように，Stanford Dogs データセットにおいても，PC Loss と COT を AMB に組み合わせることでスコアが最も高い．これらのことから，AMB に PC Loss と COT を組み合わせることで，認識に有効な領域を獲得できるため，認識精度の向上に寄与するといえる．

表 3　Insertion，Deletion 結果 (CUB-200-2010)

| | AMB | PC Loss | COT | Insertion↑ | Deletion↓ | Score↑ |
|---|---|---|---|---|---|---|
| ABN | - | - | - | 0.21 | 0.09 | 0.12 |
| 人の知見 | - | - | - | 0.17 | 0.16 | 0.01 |
| 提案手法 | ✓ | - | - | 0.22 | **0.08** | 0.14 |
| | ✓ | ✓ | - | 0.29 | 0.13 | 0.17 |
| | ✓ | - | ✓ | 0.28 | 0.12 | 0.16 |
| | ✓ | ✓ | ✓ | **0.33** | 0.13 | **0.21** |

表 4　Insertion，Deletion 結果 (Stanford Dogs)

| | AMB | PC Loss | COT | Insertion↑ | Deletion↓ | Score↑ |
|---|---|---|---|---|---|---|
| ABN | - | - | - | 0.42 | 0.28 | 0.14 |
| 提案手法 | ✓ | - | - | 0.41 | 0.28 | 0.13 |
| | ✓ | ✓ | - | 0.44 | 0.30 | 0.14 |
| | ✓ | - | ✓ | 0.46 | 0.29 | 0.17 |
| | ✓ | ✓ | ✓ | **0.58** | **0.22** | **0.36** |

## 5　おわりに

本研究では，ABN に注視領域が認識に有効か考慮しながら学習する AMB，特徴空間を改善する PC Loss，出力空間を改善する COT の 3 つのアプローチにより ABN の注視領域を改善する手法を提案した．評価実験では，提案手法を用いることにより，不要な注視領域を軽減し，認識精度が向上したことを確認した．今後は，学習方法やモデル構造の検討による更なる注視領域の改善を行う．

## 6　謝辞

本研究は，新エネルギー・産業技術総合開発機構 (NEDO) から委託されたプロジェクト JPNP20006 から得られた結果である．

## 参考文献

[1] Alex, K., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, Neural Information Processing Systems, pp. 1097–1105 (2012).

[2] Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H.: Attention branch network: Learning of attention mechanism for visual explanation, 2019 IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705–10714 (2019).

[3] Mitsuhara, M., Fukui, H., Sakashita,Y., Ogata,T., Hirakawa, T., Yamashita, T., and Fujiyoshi, H.: Embedding human knowledge in deep neural network via attention map, VISAPP, pp. 626–636 (2021).

[4] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu: Tell me where to look: Guided attention inference network, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9215–9223 (2018).

[5] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A.: Learning deep features for discriminative localization, 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016).

[6] Ramprasaath, R. S., Michael, C., Abhishek, D., Ramakrishna, V., Devi, P., and Dhruv, B.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Local-ization, International Conference on Computer Vision, pp. 618–626 (2017).

[7] Lin, M., Chen, Q., and Yan, S.: Network in network, in 2nd International Conference on Learning Representations, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings (2014).

[8] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P.: Caltech-UCSD Birds 200, Technical Report CNS-TR-2010-001, California Institute of Technology (2010).

[9] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li: Novel dataset for fine-grained image categorization: Stanford dogs, in Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Vol. 2, Citeseer (2011).

[10] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, Computer Vision and Pattern Recognition, pp. 770–778 (2016).

[11] Vitali Petsiuk, Abir Das, and Kate Saenko: RISE: Randomized Input Sampling for Explanation of Black-box Models, British Machine Vision Conference (BMVC), (2018).

[12] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg: SmoothGrad: Removing noise by adding noise, arXiv preprint, arXiv:1706.03825 (2017).

[13] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry Jackel, Urs Muller, and Karol Zieba: VisualBackProp: Efficient visualization of CNNs, arXiv preprint, arXiv:1611.05418 (2016).

[14] Zhang, Qinglong and Rao, Lu and Yang, Yubin:Group-cam: Group score-weighted visual explanations for deep convolutional networks,arXiv preprint arXiv:2103.13859 (2021).

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, (2016).

[16] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasub- ramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Com- puter Vision, pp. 839–847, (2018).

[17] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In European Conference on Computer Vision, pp. 818–833, (2014).