H-008

物体領域に着目した画像分類に関する研究
A Method of Image Classfication Focusing on Object Area

鷲田 武晃† 大野 将樹‡ 獅々堀 正幹‡
Takeaki Washida Masaki Oono Masami Shishibori

1.はじめに

近年、Deep Learning を用いた様々な物体の分類システムが普及しており、物体間の特徴差が明確なものであれば 90%を超える高精度な分類に成功している。Deep Learning の今後の課題として、Fine-grained 画像分類[1]という物体間の特徴差が小さい物体の分類精度の向上が挙げられる。

Fine-grained 画像分類のような特徴差が小さい物体を畳み込みニューラルネットワーク(CNN)を用いて分類する場合、分類対象となる物体領域ではなく、背景領域の特徴差に依存して分類結果を出力し、誤分類を招くケースが多い。

本研究では Grad-CAM[2]と Semantic Segmentation[3]を組み合わせて CNN の注目領域を絞り込むことで、分類精度を向上させる手法を提案した。

2. 提案手法

本研究では、Fine-grained 画像分類において、Grad-CAM と Semantic Segmentation を組み合わせることによって精度向上を図る。

まず、CNN の学習を行う Fine-grained な特徴を持つ画像データセットとして 200 種類の鳥画像 11,788 枚のデータセット「CUB-200-2011」を用いて、Fine-tuning を行う CNN モデルは VGG16[4]を用いて実験を行った。図 1 に提案手法の概要を示す。

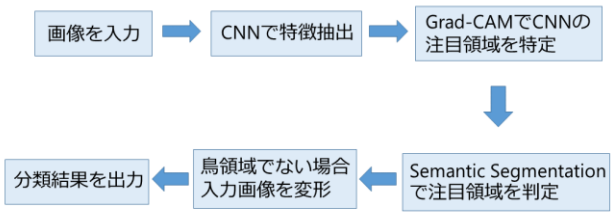

図 1: 提案手法の概要

Grad-CAM と Semantic Segmentation の結果から、図 2.1 から図 2.4 において、CNN の注目領域に背景領域が一定ピクセル以上含まれている場合(図 2.1)及び注目領域が全て背景領域の場合(図 2.3)提案手法を適用する。

具体的には、Grad-CAM 結果画像の RGB 値情報から注目領域のピクセル情報を配列に格納する。同様に Semantic Segmentation 結果画像からも鳥領域と背景領域のピクセル情報を配列に格納し、CNN の注目領域のピクセル情報が鳥領域か背景領域か判定し背景領域を多く含んでいる場合、または画像全体においてこれら二つのピクセル情報が全く一致しない、つまり全て背景領域に注目していた場合、鳥領域のピクセル情報を元に、図 2.5 から図 2.6 のように入力画像を自動で領域削減する。領域削減を行う際、鳥領域が切り出されないように設定した。

図 2.1:Grad-CAM 結果

図 2.2:Semantic Segmentation 結果

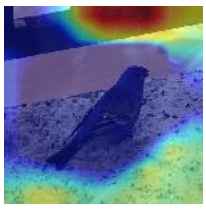

図 2.3:Grad-CAM 結果

図 2.4:Semantic Segmentation 結果

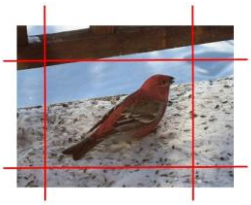

図 2.5:入力画像の領域削減

図 2.6:提案手法適用後の画像

3. 評価実験

画像データセット「CUB-200-2011」より、鳥画像 200 クラス全 1,600 枚を実験データとして用いた。実験結果を比較するため、ここで従来手法と提案手法についてまとめる。

従来手法は Fine-tuning した VGG16 モデルを用いて入力画像 1,600 枚を分類、結果を出力する。提案手法 1 は CNN の注目領域に背景領域のピクセル情報が多く含まれている場合に領域削減を行い、削減後の画像を再び CNN に入力し分類、結果を出力する。提案手法 2 は CNN の注目領域が背景領域のみの場合に領域削減を行い削減後の画像を再び CNN に入力し分類、結果を出力する。提案手法 1 と 2 の差は鳥領域および背景領域に注目している(提案手法 1)か、背景領域のみに注目している(提案手法 2)かどうかである。

実験を行った画像 1,600 枚には同一の画像は含まれておらず、CNN モデルを再学習する際に使用した画像も存在しないものとする。

† 徳島大学大学院先端技術科学教育部, Graduate School of Science and Technology, Tokushima University
‡ 徳島大学大学院社会産業理工学研究部, Graduate School of Technology, Industrial and Social Science, Tokushima

3.1 実験結果

評価方法には適合率を用いた。適合率とは、システムが出力した結果に対する正解データの割合である。入力画像全 1,600 枚に対する実験結果を表 1 に示す。

表 1:実験結果

	適合率(%)
従来手法	52%
提案手法 1	53%
提案手法 2	54%

入力画像全 1,600 枚のうち、提案手法を適用した画像の適合率の変化を表 2 に示す。

表 2:提案手法を適用した画像の実験結果

	適合率(%)
従来手法	25%
提案手法 1	30%

	適合率(%)
従来手法	21%
提案手法 2	38%

提案手法 1、提案手法 2 共に従来手法より向上が見られ、領域削減を行い CNN の注目領域を絞る手法の有効性が示された。提案手法による適合率変化が大きい例として図 3.1 から図 3.5 に実験結果の一例を示す。入力画像(図 3.1)に対し、従来手法での分類結果では、正解クラスは 7 位に分類されていた。Grad-CAM と Sematic Segmentation の結果

図 3.1:入力画像

図 3.2:提案手法適用後の画像

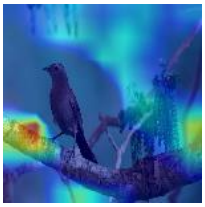

図 3.3:Grad-CAM 結果

図 3.4:Semantic Segmentation 結果

図 3.5:Grad-CAM 結果

(図 3.3, 図 3.4)より、CNN は背景領域にのみ注目しているため、提案手法 2 を適用する。適用後の画像(図 3.2)を再び CNN に入力した分類結果では、正解クラスは 1 位に分類されており、提案手法の有効性が確認できる。提案手法適用後の Grad-CAM 結果(図 3.5)からも、提案手法適用後は物体領域に正しく注目していることが確認できる。

3.1 実験結果に対する考察

提案手法 2 の方が適合率の変化量が大きい理由として、Fine-tuning をして特徴の学習を行った際に鳥領域と背景領域の相関性を学習しているケースが存在するからだと考えられる。例えば、森の中にいる学習画像が多いクラスの鳥などは、周りの木の枝などもその鳥の特徴として学習し、実際に分類を行う際に、この鳥は赤色の羽を持っていて森の中にいるからこの鳥である。と分類するように、背景も分類の手がかりになっている可能性があると考える。よって背景領域と物体領域どちらにも注目している提案手法 1 は結果的に精度向上しているが、入力画像の領域削減を行わない方が良いケースも多いと言える。

対して提案手法 2 が大幅な精度向上に繋がった要因として、CNN は画像間の特徴差が大きい箇所に注目して学習を行うため、本研究のような Fine-grained な特徴を持つ画像を学習する際には、物体領域の特徴差に加えて、提案手法 1 の考察の際にも触れた物体領域と背景領域の相関性も特徴として学習を行い、最後に背景領域の特徴差もそのクラスの特徴として学習すると考えられる。提案手法 1 に関しては物体領域と背景領域の相関性を切り離すような結果となるケースが存在するため精度低下する例が存在するが、提案手法 2 に関しては物体領域に全く注目しておらず全て背景領域のみに CNN が注目してしまっている場合なので、本来学習してほしくない背景領域のみの特徴に注目してしまっているケースを切り離せるため、有効な手法であり大幅な精度向上に繋がったのだと考えられる。

4. まとめ

本研究では Grad-CAM と Semantic Segmentation を適用して CNN の注目領域を絞り込むことで精度向上を図った。

今後の課題は、本研究では CNN が注目領域を誤っている一部の画像にしか提案手法が適用されないが、CNN が抽出する特徴量のうち、物体領域の特徴量のみを用いて分類を行うといった全ての画像に有効な手法を確立することである。

参考文献

[1] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, “ Bilinear CNN Models for Fine-grained Visual Recognition",IEEE International Conference on Computer Vision, pp. 1449-1457, 2015.

[2] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and DhruvBatra, “Visual Explanations from Deep Networks via Gradient-based Localization", The IEEE International Conference on Computer Vision (ICCV),pp. 618-626, 2017.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", Computer Vision ECCV, pp. 833-851, 2018.

[4] Karen Simonyan, and Andrew Zisserman,“Very Deep Convolutional Networks for Large-Scale Image Recognition", Computer Vision and Pattern Recognition, 2014.