The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

解釈可能性をもたらすヒートマップの変化に着目した過剰学習が CNN にもたらす影響の検討

Investigation of the influence of over training on CNN focusing on change of heat map which brings interpretability

古澤 嘉久 *1 FURUSAWA Yoshihisa
田和辻 可昌 *2*3 TAWATSUJI Yoshimasa
松居 辰則 *3 MATSUI Tatsunori

*1早稲田大学人間科学部 School of Human Sciences, Waseda University
*2早稲田大学 大学院人間科学研究科 Graduate School of Human Scineces, Waseda University
*3早稲田大学 人間科学学術院 Faculty of Human Sciences, Waseda University

Although Convolutional Neural Network (CNN) is used in many studies, the interpretability of CNN have been considered problematic and various methods have already been proposed from related research. All of these methods have been verified with selected models based on Early Stopping and are evaluated with only one model. However, in the classification problem, it is reported that the accuracy increases due to over training, so it is not known whether model selection by Early Stopping is appropriate. Therefore, in this research, the influence of this over training is considered from the viewpoint of the explanability of CNN. As a result, for each learning period, we found a dataset in which the value of the correct class and entropy of the output oscillate in the learning process. And we found that it is possible to create a heatmap with different similarity from the heatmap created by Early Stopping.

1. はじめに

近年の機械学習に関する研究の成果として, 特に画像分類に関しては, Convolutional Neural Network(CNN) が高い精度が出るようになり, 注目が集められている. しかし高い精度が出る反面, CNN は「どのようにしてその出力に至ったか」という人間にとっての解釈可能性に問題が残るモデルである.

この問題に関しては, すでに多くの既往研究により研究が進められており [Zhang 18], 入出力の関係を調査する際には, 最終層の出力の特定のクラスに関してのみ逆伝播を行い, 適宜制約を加えることで入力に寄与した画像箇所をヒートマップとして表現することによって, この CNN の解釈可能性を補う手法がいくつか提案されている.

そして, これらの解釈可能性を補う手法は Early Stopping などの指標に基づき学習を止めた学習済みモデルで評価を行なっている. しかしながら, Implicit Regularization と呼ばれる適当な条件下であれば, 学習回数を増やすほど対数的に精度が上昇する正則化現象が生じている [Soudry 18] ことから, 精度に関して言えば, Early Stopping によるモデル選択が最良なモデルを選択できるとは限らない可能性が示唆されている.

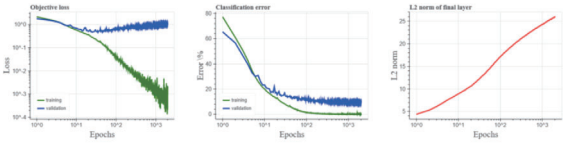
そこで本研究では, 既存の手法が Implicit Regularization が生じるような Early Stopping 以降の学習 (過剰学習) 過程における解釈可能性を補う手法が作成するヒートマップの変化を調査した. そして,Implicit Regularization による精度上昇がモデルにどのような影響を与えるかを実験的に評価した. また, 本研究の図中のプロットの横軸は特別な記載がない限り, 全て Epoch 数とする.

連絡先: 古澤 嘉久, 早稲田大学, 〒 359-1165 埼玉県所沢市堀之内 135-1 フロンティア リサーチセンター 213 実験室, f.y_1996_w-skk@akane.waseda.jp

2. 先行研究

2.1 Implicit Regularization

Implicit Regularization[Soudry 18] とは, 重みの初期値の値に依存することなく, 学習データでの損失関数の値が十分に小さくなった場合に, 適当な条件を満たしていれば, 学習データから学習した境界超平面付近のデータセットが摂動として作用し, 学習が進むことによって成立する重みの正則化現象である. これは例えテストデータにおける損失関数の値が上昇したとしても図 1 のように, テストデータに対するモデルの精度は非常に遅い速度ではあるが, 上昇するというものである. また [Soudry 18] によると, 適当条件にはいくつかあり, まず最適化手法として最急降下法を使用していること, そしてデータセットが Low Noise な線形分離可能なものであること. また, 損失関数の形がクロスエントロピー誤差のような指数型の損失関数であったり, 損失関数を各パラメータで偏微分したものが常に正の値をとるなどが挙げられる.



Epoch	50	100	200	400	2000	4000
L_2 norm	13.6	16.5	19.6	20.3	25.9	27.54
Train loss	0.1	0.03	0.02	0.002	10^{-4}	$3 \cdot 10^{-5}$
Train error	4%	1.2%	0.6%	0.07%	0%	0%
Validation loss	0.52	0.55	0.77	0.77	1.01	1.18
Validation error	12.4%	10.4%	11.1%	9.1%	8.92%	8.9%

図 1: [Soudry 2018] より結果の引用

2.2 解釈可能性を補う手法

本研究では, 過剰学習期間における解釈可能性を補う手法の影響を調査するために, 以下のような手法を使用した. そして今準備として, 入力画像を $\boldsymbol{x} \in \mathbb{R}^n$ とし, $i \in \{1, 2, ..., n\}$ をピ

クセルのインデックスとする. これを C クラス分類するモデルを考え, 入力画像 $\boldsymbol{x}$ に対して, クラス $c \in C$ の出力を $S_c(\boldsymbol{x})$ と表すことにする.

2.2.1 Saliency Map

[Simonyan 14] は, $S_c(\boldsymbol{x})$ が線形的なモデルの場合は, 入力画像に対する重要さとはそのモデルの重みとして表現されるとして, 下記のように, クラス c へ連結する重み $\boldsymbol{w_c} \in \mathbb{R}^n$ が重要さに該当していると主張する.$b_c \in \mathbb{R}$ はバイアス項.

$$S_c(\boldsymbol{x}) = \boldsymbol{w_c^T x} + b_c \tag{1}$$

また実際のモデル $S_c(\boldsymbol{x})$ は非線形であるため, 入力画像に関してテイラー展開し, 一次近似した際に, 以下のように入力画像での勾配が重み部分に該当することがわかる.

$$S_c(\boldsymbol{x}) \approx \boldsymbol{w^T x} + b, \ \ \boldsymbol{w} = \left.\frac{\partial S_c(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x_0}} \tag{2}$$

そこで [Simonyan 14] は, 最終的なクラス c への出力から逆伝播を行い, 入力画像に逆伝播されてきた勾配を Saliency Map として採用することを提案している.

2.2.2 Deconvolution

[Zeiler 14] は, CNN の誤差逆伝播と同じ計算過程を誤差ではなく入力に対して行い, どの層の出力がどのような画像的特徴を持った箇所に寄与しているのかに関する調査を行なった. 本実験では, [Ramprasaath 17] のように誤差逆伝播の際の制約として, この Deconvolution の制約を使用する. 深層学習は, 誤差逆伝播法と呼ばれる最急降下法の各パラメータの更新に必要な勾配を連鎖律に基づき計算を行うが, この際に活性化関数に ReLU(max(0, x)) を使用する場合は, 入力時にユニットの出力が正をとる箇所のみに誤差が逆伝播していく. つまり, Deconvolution の制約とは, この ReLU の制約ではなく, 順伝播の際の出力にかかわらず勾配が正の値をとる時のみ逆伝播するというものである.

2.2.3 Guided Backpropagation

[Springenberg 15] は, 先ほど紹介した Deconvolution の制約と ReLU の制約を同時に満たす場合のみに逆伝播を行う手法を提案している. つまり, 順伝播の際には正となり, 逆伝播の際にも勾配が正となる箇所のみで逆伝播を行い, その他の部分はその層よりも入力側に存在する層には, 逆伝播をしないようしている. また今回は, 解釈可能性を補う手法としてではなく, 逆伝播される勾配の変化を見るために, Guided Backpropagation Minus として,Guided Backpropagation のように入力時に正となり, 勾配が負になるようなものを手法の一つとして使用する.

2.2.4 Grad-CAM

[Zhou 16] は, k 番目の特徴マップ $\boldsymbol{A^k} \in \mathbb{R}^{u \times v}$ とそれに対応する重み $a_k^c \in \mathbb{R}$ を用いた重み付け和を計算することで, 解釈可能性を補うヒートマップを出力する手法を提案している. つまり, 最終的なヒートマップ $\mathrm{L}^{\boldsymbol{c}}_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$ は下記のようになる.

$$\mathrm{L}^{\boldsymbol{c}}_{\text{Grad-CAM}} = \ \mathrm{ReLU}\left(\sum_k \alpha_k^c \boldsymbol{A^k}\right) \tag{3}$$

$$\alpha_k^c = \frac{1}{Z}\sum_i^u \sum_j^v \frac{\partial S_c(\boldsymbol{x})}{\partial A_{ij}^k} \tag{4}$$

$z \in \mathbb{R}$ は正規化定数であり, 本研究では $u \times v$ を使用. つまり, α_k^c の計算は, 最終的なクラス c の出力である $S_c(\boldsymbol{x})$ から逆伝播を行い, 特徴マップ $\boldsymbol{A^k}$ に返ってきた勾配の幾何平均をとっている.

2.2.5 Guided Grad-CAM

Grad-CAM は, 特徴マップの重み付け和を行なっているため, 特徴マップの大きさとヒートマップの大きさが同じになり, 入力画像 x とサイズが異なってしまう. つまり, 入力画像と同じ解像度の画像として見た際に, ピクセル単位でのヒートマップにはならない. そこで, Grad-CAM により作成したヒートマップをバイリニアサンプリングし, 拡大した後に, 入力画像の要素積を計算することでピクセル単位での可視化を可能にしている.

3. データセットの選択

本研究は Early Stopping により学習をやめたモデルから, Implict Regularization により精度が上昇したモデルの変化を比較することを目的とするため, 下記のようなデータセットをテストデータより抜粋した.

3.1 選択基準

下記のような選択基準に基づきデータセットを抜粋した. この指標は, [Wang 16] が Active Learning の分野で使用した Least confidence の考えに基づいている. データの選択には式 (5) を使用し, 上の式は最終層の最大予測クラスと正解クラスが同じ場合に, 正解クラスの予測確率をとり, 異なった場合は, 最大予測確率と正解クラスの予測確率の絶対値をとる.

$$\begin{cases} S_c(\boldsymbol{x}) \left(\arg\max_{i \in C} S_i(\boldsymbol{x}) = c\right) \\ \left|\max_{i \in C} S_i(\boldsymbol{x}) - S_c(\boldsymbol{x})\right| \left(\arg\max_{i \in C} S_i(\boldsymbol{x}) \neq c\right) \end{cases} \tag{5}$$

3.2 データセットの作成方法

データの選択は, 二つのリストから行なっており, 一つ目は式 (5) の上の式についてまとめたリスト A であり, 二つ目は式 (5) の下の式についてまとめたリスト B がある. またこれを各クラスに対して行なった. そして, リスト A, B を共に降順に並べ, リスト A から上位 5 つをデータ A「明らかに正確に分類されているもの」とし, リスト B から上位 5 つをデータ B「明らかに誤分類されるもの」とし, リスト A から下位 3 つとリスト B の下位 3 つを合わせたものをデータ C「分類がどちらにもされうるもの」とした.

4. 対象となるモデルの作成

解釈可能性を補う手法を使用する対象であるモデルを作成した際の詳細と実験的に得られた知見について述べる.

4.1 モデル詳細

モデルとしては [Soudry 18] の実装を元に ResNet18[He 15] を採用した. データセットは CIFAR-10, 活性化関数は ReLU, 最適化手法はモーメンタム確率的最急降下法を使用.

4.2 結果

モデルを準備する上で図 2 のように, Implict Regularization の有無が発生に寄与している可能性が実験的に示唆された. 本研究では, Data Augmentation として, Random Crop と左右反転をしている.

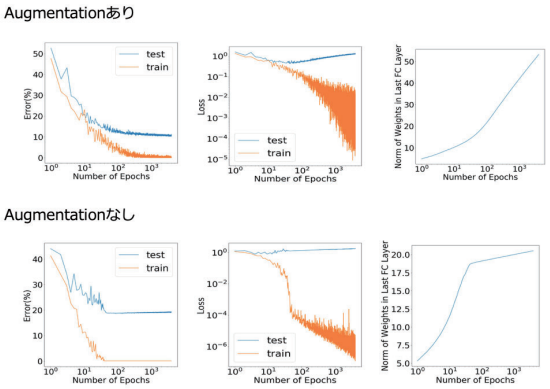


図 2: Augumentation の有無による影響. エラー率 (左図),Loss(中央図), 最終層の重み行列のノルム (右図)

5. 評価

解釈可能性をもたらすヒートマップの変化について考察をするために, まず定性的な変化を先ほどのデータセットに対して行い, 学習過程において, それぞれの手法がどのような変化をするか考察した. 次に定性的に敏感に変化していた手法に関して, SSIM(Structural Similarity), スピアマンの順位相関係数を用いて, Early Stopping を行なった際のヒートマップと各学習過程でのヒートマップの類似度を比較した. また Early Stopping によって選択したモデルが Epoch 数が 41 の時のモデルであるため, ヒートマップの比較に関しては, こちらを比較対象とする.

5.1 定性的評価

5.1.1 解釈可能性を補う手法の変化

選択したデータセットに関して, 各手法の変化の過程を定性的に評価した. 結果として, Saliency Map, Grad-CAM 系の手法は, 敏感に変化している様子が見て取れた. 変化の傾向としては, 局所的な特徴に留まるもしくは, 様々な箇所を可視化するようになるなど人間にとっての解釈性に問題が生じるということが実験的に確認された. また Grad-CAM 系に関しては, 図 3 のように,Implict Regularization によっても他のあまり変化しなかった既往研究の手法に関しては, [Nie 18], [Adebayo 18] からモデルの重みに依存しにくい手法と報告されているため, 今回も変化が少なかったと考えられた. また今回調査のために使用した Guided Backpropagation Minus に関しても変化があまり見られず, 逆伝播に単純な制約をつける手法は, モデルの変化に依存しにくい手法となる可能性も示唆されたと考えている. この変化の具合に関しては, 今回分割したデータ群において明らかな違いや傾向が見られなかった.

5.1.2 出力層のエントロピーと正解クラスの出力値の変化

図 4 は, データ C のクラス 0 の場合に, 対する出力層のエントロピーと正解クラスの予測確率と Epoch 数のプロットである. データ A に関しては, 下段中央図のようにエントロピー, 正解クラスの出力値共に 1 に近くで一定であった. そして, データ B, C に関してはエントロピー, 出力値共に振動するように, 不安定になる様子が多く見られた. この結果から,Implict Regularization は精度面では上昇するが, Early Stopping を採用したモデルの時点で「明らかに正確に分類されているもの」以外のデータセットでは, 出力が不安定になる可能性が示唆された.

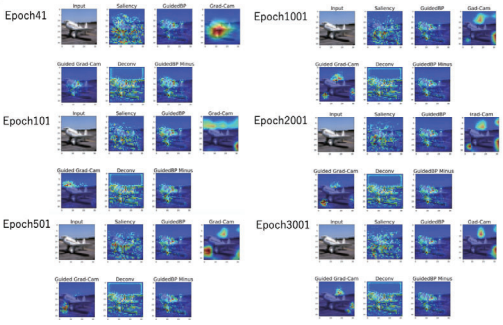


図 3: 学習過程ごとのヒートマップの変化

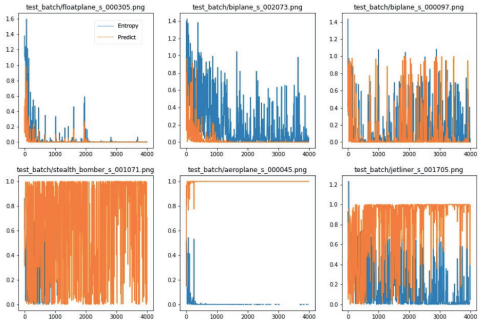


図 4: クラス 0(飛行機) に対する出力層のエントロピーと正解クラスの予測確率の変化

5.2 定量的評価

5.2.1 評価指標

本研究ではヒートマップの変化具合の定量的な評価を行うために, [Nie 18] を参考にし, SSIM とスピアマンの順位相関係数を使用した. SSIM とは, 画像圧縮の分野などで使用される指標であり, 0 から 1 の値をとり, 値が大きい方がより類似度が大きい画像とされている. また, 0.9 以上の値を下回ると画像としての劣化が見られるとされている.SSIM の定義に関しては, 下記のようになる. X,Y はそれぞれ比較する画像であり, window と呼ばれる crop 処理を行い, $\mathbf{x_j} \in \mathbb{R}^n$,$\mathbf{y_j} \in \mathbb{R}^n$ は, それぞれの画像から j 番目に crop してきた画像である.$\mu_x \in \mathbb{R}$ は,x_j 内の標本平均であり, σ_{xy} は不偏共分散, C_1, C_2 は定数項である.M は総 window 数である. そして, 全ての window に関しての SSIM の平均をとることで今回の指標とした.

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{6}$$

$$\mathrm{MSSIM}(\mathbf{X}, \mathbf{Y}) = \frac{1}{M}\sum_{j=1}^{M} \mathrm{SSIM}(\mathbf{x}_j, \mathbf{y}_j) \tag{7}$$

次にスピアマンの順位相関係数について説明する. スピアマンの順位相関係数はノンパラメトリックな指標の一つであり, -1 から 1 の値をとり, 絶対値が大きい方が強い相関を表している. d_i は, 画像 X,Y の同じ位置にある i 番目インデックスを表し, その順位の差である. n は総ピクセル数.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{8}$$

5.2.2 変化に関する評価

定性的に評価した際に, Saliency Map と Grad-CAM 系の手法がモデルに対して敏感な変化をしていたことから,Saliency Map と Grad-CAM に絞り, 上記の指標を用いて定量的に評価を行なった. 結果としては, 過剰学習期間においてのヒートマップを比較した際に, 図 5 のように SSIM に関しては 0.4 以下になり, スピアマンの順位相関係数に関しては 0.4 以下になった. 特に 1000Epoch 以上学習させた場合に関しては, 全てのデータセット・指標において 0.3 以下になった. つまり, どちらも Early Stopping で学習を止めた場合と画像としての類似度に違いがでていることが確認された.

また, Saiency Map は全てのデータセットと学習期間に対して, スピアマンの順位相関係数が 0.1 以下と小さくなっていた. 既往研究よりノイズがひどい手法であることが問題視されている手法であるため [Smilkov 18], 物体の存在しない部分の順位の大小が結果に大きく反映されている可能性が考えられた. そこで, SmoothGrad[Smilkov 18] のような Saliency Map と同じように制約の加えていない入力層での勾配を使用しつつ, ノイズの少ない手法での評価を追加し, 行う必要があることが示唆された.

スピアマンの順位相関係数に関して,Grad-CAM の方が Saliency Map よりも大きくなっている背景としては, Grad-CAM は, 最終出力が 0 以上になるように処理されるため, マイナスの出力は 0 に直されることがあげられた. これは, 物体の存在する可視化を行なっているために, 物体外に存在する箇所に関しては 0 となり, その部分に順位の違いが生じにくいと考えられルためである. この結果から, モデル選択として, Implicit Regularization を考慮し学習を続けることは, Early Stopping により作成したヒートマップとは異なったものを作成するようになるため, いたずらに学習をさせ続ければ良いとはならない可能性が示唆された.

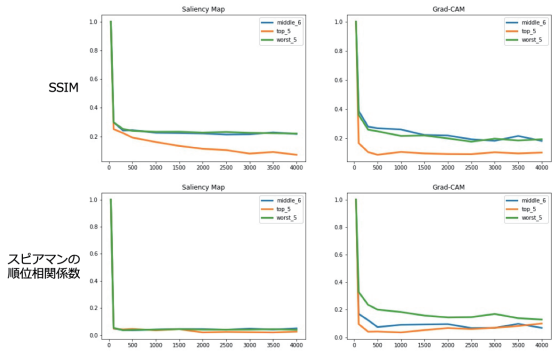


図 5: 学習過程ごとの SSIM(上段) とスピアマンの順位相関係数 (下段) の変化

6. まとめと今後の展望

本研究により, 出力層のエントロピーと出力値が振動するように学習が進む傾向がいくつかのデータセットにおいて見られ, それは主に「明らかに正確に分類されているもの」以外のデータセットで見られた. 解釈可能性を補う手法に関しては, Grad-CAM 系の手法は Implicit Regularization の影響を受けやすく, Grad-CAM によるヒートマップは過剰学習の与える精度の上昇によって, 局所的な特徴に留まるもしくは, 様々な箇所を可視化するようになるなど大きく変化する傾向にあることが定性的評価から確認された. また, 定量的な評価に関しても, 画像としての類似度は低下していることが確認された. 今後の展望としては, 変化したヒートマップが人間にとっての可読性に沿っているかを評価するために, [Ramprasaath 17] のようにアンケート調査を行い, Early Stopping と比べた際に人間にとっての可読性に差が現れるか調査する必要がある. また, その他のデータセット, モデルでの評価を行い, この結果が CIFAR-10 と ResNet 系のモデル特有のものであるかという汎用性に関する調査が必要である.

参考文献

[Adebayo 18] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps, Advances in Neural Information Processing Systems 31, (2018)

[He 15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition, Large Scale Visual Recognition Challenge 2015, (2015)

[Nie 18] A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations: Weili Nie, Yang Zhang, Ankit Patel, International Conference on Learning Representations 2018, (2018)

[Ramprasaath 17] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra: Grad-CAM:Visual Explanations from Deep Networks via Gradient-based Localization, International Conference on Computer Vision 2017, (2017)

[Simonyan 14] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, International Conference on Learning Representations 2014, (2014)

[Soudry 18] Daniel Soudry, Elad Hoffer, Nathan Srebro: The Implicit Bias of Gradient Descent on Separable Data, International Conference on Learning Representations 2018, (2018)

[Springenberg 15] Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller: Striving for Simplicity: The All Convolutional Net, International Conference on Learning Representations Workshop 2015, (2015)

[Smilkov 18] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Vi egas, Martin Wattenberg: SmoothGrad: removing noise by adding noise, International Conference on Machine Learning Workshop 2018, (2018)

[Wang 16] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, Liang Lin: Cost-Effective Active Learning for Deep Image Classification, Transactions on Circuits and Systems for Video Technology 2016 (2016)

[Zeiler 14] Zeiler,M. D.,Fergus, R: Visualizing and understanding convolutional networks, European Conference on Computer Vision 2014, No.818 833 (2014)