The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

入力データ非貢献部抽出によるCNNの出力の説明

Highlighting Non-contributing Pixels for Visual Explanation of CNNs

池野 光一 *1　原 聡 *1　鷲尾 隆 *1
Kouichi Ikeno　Satoshi Hara　Takashi Washio

*1大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

Explaining the output of Convolutional Neural Networks (CNNs) is a challenging topic. A typical explanation is to identify which pixels are contributing to the output of CNN. In this paper, we propose a new approach for explaining the output of CNNs by finding pixels that are *not* contributing to the output. To highlight non-contirbuting pixels, we propose optimizing a noise level so that additive noise to the input image does not change the CNN output. The experimental results on MNIST show that the proposed method can idntify non-contributing pixels adequately.

1. 概要

近年の機械学習技術の研究の成果では，特に認識・予測の精度の高さに注目が集まっている．しかし，実応用の場面においては，単に精度だけでなくあわせてモデルの説明性が必要とされることがある．例えば医療現場における病気の診断では，単に患者が特定の疾患に罹患しているかを高い精度で判断するだけでは不十分である．医療サービス提供側には，判断結果だけでなくなぜそのような疾患であると判断したのか，その理由を患者に説明する責任が発生する．このような機械学習の実応用における要請を背景に，機械学習モデルの説明性のニーズが高まっている [1].

本研究では，画像認識のための深層学習モデルである Convolutional Neural Network (CNN)[2] を対象に，その説明性を高める方法を提案する．CNN の説明性を高める方法として，モデルが入力画像のどの部分を根拠に認識を行っているかを特定してハイライトする方法が研究されている [3, 4]．このような説明法を用いることで，モデルの認識根拠を人間が視覚的に検証できるようになる．

提案法では，モデル説明のための新しいハイライト法として「モデル出力に寄与しない入力画像の非貢献部をハイライトする」方法を提案する．従来の手法では，モデル出力と関連が強い入力画像領域を貢献部としてハイライトする．これに対し，本研究では非貢献部に着目することで，これら従来の手法とは逆のアプローチを取る．提案法は従来法のようにモデル線形化のような粗い近似を必要としないという点において，モデルのより本質的な情報に基づいたハイライトを生成できると考えている．

2. 既存手法

本節では,CNN の判断根拠の説明のための代表的なハイライト法を紹介する．

2.1 感度マップ

感度マップでは，入力画像中のピクセルのうち, その値の微小変化が出力に大きく影響するピクセルを重要なピクセルだと考えハイライトする手法である．これは説明対象の CNN モデルを線形近似し，入力画像のピクセルごとに出力の勾配を調べることに相当する．

今，入力画像を x とし，p をピクセルのインデックスとする．これを全 C クラスのうちのある 1 つのクラスに分類するようなモデルを考える．入力 x を受けそれぞれのクラス $c \in C$ に対しスコア $S_c(x)$ が計算され，最終的にこのスコアが最も大きいクラス c に分類される．これは以下の式で表される．

$$class(x) = \mathrm{argmax}_{c \in C} S_c(x)$$

これに対し, 感度マップは以下のようにスコアの勾配として定義できる．

$$(M_c(x))_p = \frac{\partial S_c(x)}{\partial x_p}$$

ただし，$(M_c(x))_p$ は $M_c(x)$ の第 p 成分である．$(M_c(x))_p$ は x_p の微小変化に対してスコア $S_c(x)$ がどれほど変化するかを表したものであるから，$(M_c(x))_p$ の値が大きいピクセル x_p がモデルの出力に大きく寄与していると考えられる．図 1 に感度マップの例を示す．

2.2 SmoothGrad [3]

SmoothGrad は，感度マップを改良した手法である．図 1 に見られるように，感度マップでは認識対象でない背景部分もハイライトされている．SmoothGrad は平均化によりこのような背景のノイズを低減する．

元入力 x にガウシアンノイズ $\mathcal{N}(0, \sigma^2)$ を乗せたものを n 個用意し，それぞれを新たな入力として感度マップ $M_{c,k}(x)(k = 1, ..., n)$ を生成する．それらの平均 $\hat{M}_c(x)$ をとり，モデルの説明とする．

$$\hat{M}_c(x) = \frac{1}{n} \sum_{k=1}^{n} M_c(x + \epsilon_k), \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2)$$

このように,x に微小なノイズを加えてもモデルの判断はほとんど変わらないが，異なるノイズが乗ることで感度マップによってハイライトされる部分はそれぞれ少しづつ異なる．SmoothGrad では, それらの平均を取ることで判断の際に頻繁に重視される部分が浮かび上がり，ノイズに反応してしまった部分はハイライトされづらくなる．

連絡先: 池野 光一，大阪大学 産業科学研究所，
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
k1keno@ar.sanken.osaka-u.ac.jp

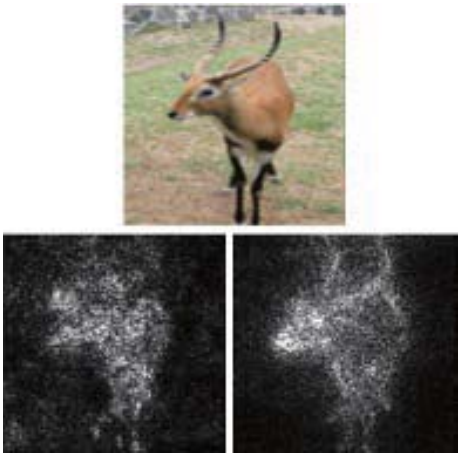

図 1: 上が入力画像，左下が感度マップ，右下が SmoothGrad である．認識対象のぼやけが改善されている．(図は [3] の Fig.3 より引用)

2.3 deep Taylor Decomposition [4]

deep Taylor Decomposition は, 各入力ピクセルの貢献量を出力から計算する手法である．そのために,deep Taylor Decomposition では各層の各ユニットでの貢献量を考える．なお, この際に各層の貢献量の総和は等しいとする以下のような保存則を仮定する．いま, 第 a 層とその次の第 $a+1$ 層について、各層のユニットのインデックスをそれぞれ i, j, 各ユニットの貢献量をそれぞれ $R_i^{(a)}, R_j^{(a+1)}$ とする．このとき, 保存則とはそれぞれの層の貢献量の総和が等しいことを指す：$\sum_i R_i^{(a)} = \sum_j R_j^{(a+1)}$．deep Taylor Decomposition では, 保存則と一次のテイラー展開に基づいて各層の貢献量を計算する．一次のテイラー展開より,$R_j^{(a+1)}(\{\tilde{x}_i\}) = 0$ となる $\{\tilde{x}_i\}$ を用いて,

$$\begin{aligned}\sum_j R_j^{(a+1)} &= \left(\left.\frac{\partial\left(\sum_j R_j^{(a+1)}\right)}{\partial\{x_i\}}\right|_{\{\tilde{x}_i\}}\right)^T \cdot (\{x_i\} - \{\tilde{x}_i\}) + \epsilon \\ &= \sum_i \sum_j \left.\frac{\partial R_j^{(a+1)}}{\partial x_i}\right|_{\{\tilde{x}_i\}} \cdot (x_i - \tilde{x}_i) + \epsilon\end{aligned}$$

と表せる．ϵ_j は十分に小さいものとして無視できるとすると,

$$\begin{aligned}\sum_i R_i^{(a)} &= \sum_j R_j^{(a+1)} \\ &= \sum_i \sum_j \left.\frac{\partial R_j^{(a+1)}}{\partial x_i}\right|_{\{\tilde{x}_i\}} \cdot (x_i - \tilde{x}_i) \\ \therefore R_i^{(a)} &= \sum_j \left.\frac{\partial R_j^{(a+1)}}{\partial x_i}\right|_{\{\tilde{x}_i\}} \cdot (x_i - \tilde{x}_i)\end{aligned}$$

という更新則が導かれる．この貢献量の逆伝搬則を出力層から入力層まで順次適用することで, 入力画像のピクセルごとの貢献量 $\{R_p\}$ を求める．図 2 に deep Taylor Decomposition の一例を示す.

3. 提案法

本節では，モデル説明のための新しいハイライト法として「モデル出力に寄与しない入力画像の非貢献部をハイライトする」方法を提案する．提案法は感度マップや SmoothGrad, deep Taylor Decomposition のようにモデル線形化のような粗い近似を使わないので，モデルのより本質的な情報に基づいたハイライトを生成できると期待できる．

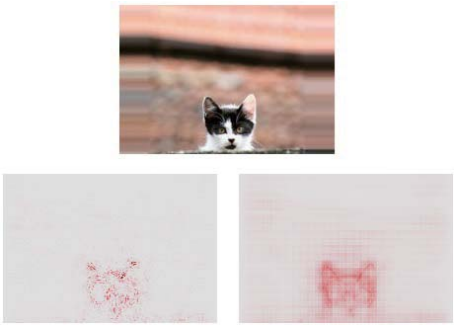

図 2: 上が入力画像，左下が感度マップ，右下が deep Taylor Decomposition である．認識対象のぼやけが改善されている．(図は [3] の Fig.3 より引用)

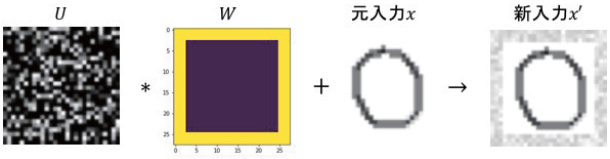

図 3: ノイズを付加する入力側の処理．黄色の部分に大きな値が，紫色の部分に大きな値がある．

3.1 基本アイディア

一般に，画像には認識対象の他に，背景部分など認識とは無関係なものも写っている．このような画像から深層学習モデルが対象を認識して分類するにあたり，認識対象は出力に貢献する貢献部だと言える．一方，認識対象以外の背景部分などは出力に貢献しない非貢献部だと言える．

提案法では，非貢献部をハイライトするため，図 3 のように説明したい入力にノイズを付加した新たな入力を考える．新たな入力 x' に対するモデルの出力 y' は，説明したい入力 x の本来の出力 y から変動するが，この変動はノイズのかかる領域が非貢献部であるか否かに大きく影響を受けると推測できる．例えば，非貢献部だけにノイズが付加されたときは認識対象自体に変化はないため，モデルの出力はほぼ変動しないと考えられる．逆に，認識対象にノイズが付加されると出力は大きく変動すると考えられる．これらの考察に基づいて，本研究では出力がほぼ変動しないノイズを見つけることで，入力画像 x の非貢献部を特定する．

提案法では，視覚的にわかりやすい説明を得るために，説明したいモデルの内部ネットワークに前処理を施す．そのために，出力の変動を考えるにあたって，x を CNN に入力したときに”特徴が見られない”として，あるしきい値より小さな値が伝搬していた部分に注目する．このような”特徴が見られない”部分は出力に寄与しないので，説明のためには不要である．そこで，この部分には必ず 0 が伝搬するようにする．本研究では, この処理を行うユニットをマスクと名付ける.

3.2 学習

提案法では，出力が変動しないようなノイズを，ノイズの大きさパラメータを求めることによって非貢献部をハイライト

する．入力に付加するノイズは以下のように生成する：U を $[-1,1]$ 上の一様分布に従う縦横サイズ L の乱数行列，w を縦横サイズ L のノイズの大きさを決めるパラメータ行列とし，ノイズを $U*w$ により生成する．つまり，新しい入力 x' はピクセル単位で $x'_{ij}=U_{ij}w_{ij}+x_{ij}$ である．このとき，w_{ij} の値が大きくても出力 y' が変動しない，つまりピクセル x_{ij} に大きなノイズを乗せても出力に影響がない場合には，x_{ij} は出力に貢献していないと判断できる．

提案法では，最適なノイズパラメータ w を求めるために以下の目的関数 $E(w)$ を定義し，その最小化により最適な w を求める．

$$E(w)=\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\max(|\,y_k-{y'}_k\,|-a,0)-\frac{1}{L^2}\sum_{i,j=0}^{L-1}w_{ij}$$

$$(0\leq w_{ij}\leq 1)$$

提案法では最適化アルゴリズムとして Adam[5] を使う．ここで，N は最適化の各ステップで生成する乱数行列 U の数，K は分類クラス数，また，y_k, y'_k はそれぞれ出力 y, y' の k 番目の要素である．右辺第 1 項は出力の変動に対する制約で，変動が許容幅 a を超えた時に 0 より大きな値をとる．このとき a を y_k の大小関係が変わらない程度の値に設定することで，認識対象の特徴を損なわないように w を学習させることができる．右辺第 2 項は多くの w_{ij} ができるだけ大きな値をとるように学習を進め，w をマップ化した際に鮮明なイメージとする．最後の制約は，ノイズの大きさを現実的な値に制限するものである．

4. 実験

本節では，提案法の有効性を確認するための実験とその結果について述べる．

4.1 実験概要

本研究では，図 4 のような CNN を説明対象とし，手書き数字画像データ MNIST[6] を入力画像として実験を行った．このモデルは入力画像を 0〜9 の 10 クラスに分類するもので $(K=10)$，また，説明対象の入力画像のいずれもが正しく分類された．前処理として，図 5 のように，マスクのしきい値は，0.2(第 1 畳み込み層)，0(第 2 畳み込み層・全結合層) とした．学習については，$N=100$ で反復回数を 15k 回とし，Adam の学習率は 1×10^{-4} を用いた．

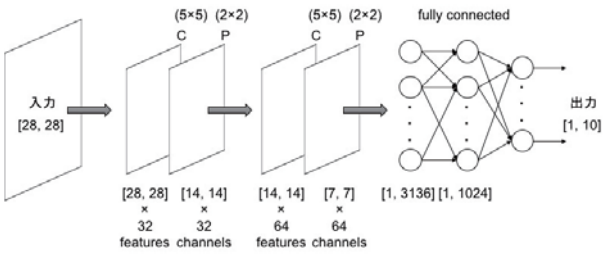

図 4: 説明対象の 3 層 CNN．(図は [3] の Fig.3 より引用)

4.2 実験結果

図 6 のように, 提案法により数字の背景部分が非貢献部としてハイライトされた．逆に，数字及びその近接部は出力に貢献する部分と判断された．これは，モデルが数字とその近接部の

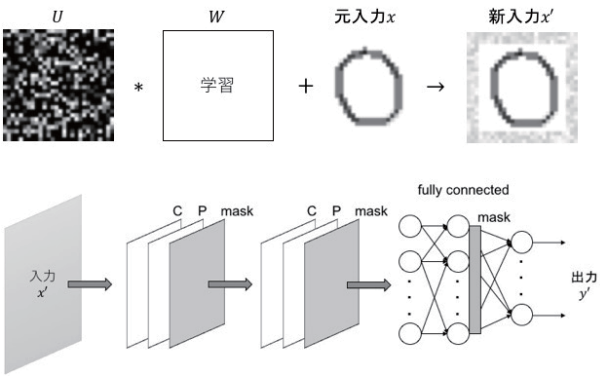

図 5: ノイズパラメータ w を学習するためのネットワークの構造図

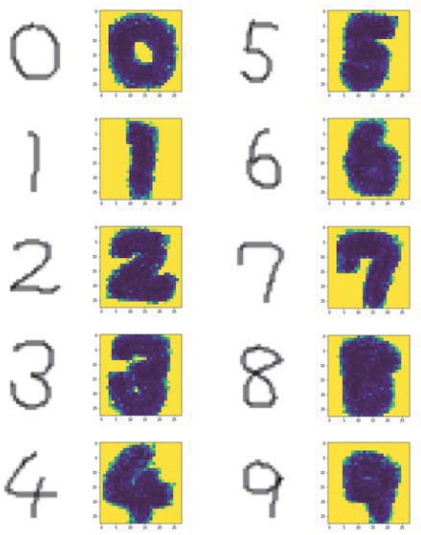

図 6: 左が入力画像，右が提案法による出力の説明である．

両方を用いて数字の輪郭を捉えているためだと考えられ, 直感に即した結果だと言える．

5. まとめ

本研究では，CNN の出力根拠を示すために入力の非貢献部をハイライトする手法を提案した．また，MNIST データを用いた実験により，提案法の有効性を検証した．結果，CNN が適切に数字の輪郭を認識した上で出力を決定していることが確認できた．

参考文献

[1] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, Fosca Giannotti. A Survey Of Methods For Explaining Black Box Models. arXiv:1802.01933v2. (2018)

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1986(11), Pages 2278-2324. (1998)

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

[4] Grégoire Montavon, Sebastian Bach, Alexander Binder, Wojciech Samek, Klaus-Robert Müller. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. Pattern Recognition Volume 65, Pages 211-222. (2017)

[5] D.P.Kingma and J.Ba.Adam A method for stochastic optimization. arXiv:1412.6980. (2014)

[6] Yann LeCun, Corinna Cortes, Christopher J.C. Burges. THE MNIST DATABASE of handwritten digits. http://yann.lecun.com/exdb/mnist/