

活性固定化による深層学習モデルの視覚的説明の鮮明化

Enhancing Visual Explanation of Deep Neural Networks by Activation Freezing

原 聰

Satoshi Hara

大阪大学 産業科学研究所

The Institute of Scientific and Industrial Research, Osaka University

The “black box” nature of deep neural networks hinder us from inspecting the logic hidden inside the networks. To alleviate this difficulty, for image recognition tasks, several visual explanation methods are proposed. The aim of visual explanation is to highlight a segment of the image as a clue of the network’s recognition. In this paper, we propose a method for enhancing the highlights of the existing visual explanation methods. In the proposed method, we *freeze* the activations of the some of the network nodes in middle layers. With the proposed method, we can obtain clear highlights with less noises than the existing methods.

1. はじめに

深層学習モデルは画像認識など多くの問題で高い認識性能を誇る。しかしながら、深層学習モデルのその複雑な内部のネットワーク構造がゆえに、モデルがどのような認識機構に基づいて画像認識を行っているかを人間がうかがい知ることはできない。このような深層学習モデルのブラックボックス性は、深層学習モデルを実用する上で大きな障害となる[1]。例えば、病気の診断やローンの審査などの人間の生活に深く関わる問題では、医師や銀行は患者や顧客に最終的な判断結果だけでなくその結果に至った理由を説明する責任を有する。深層学習モデルはそのブラックボックス性がゆえに判断結果についての説明能力を持たないため、このような問題へと安易に適用することはできない。

深層学習モデルのブラックボックス性を緩和して説明性を高めるために、特に画像認識の分野において視覚的な説明性を向上させる方法の研究が進められている[2, 3, 4, 5, 6]。これら視覚的な説明では、深層学習モデルが画像中のどの領域を根拠に認識を行なっているか、その認識対象を特定してハイライトする。図1にハイライトの一例を示す。このような説明法を用いることで、モデルが画像中の対象を適切に認識しているか、それとも誤った対象を根拠に誤認識しているか、を人間が事後に検証できるようになる。

既存の視覚的説明法にはハイライトに多くのノイズが載るという欠点がある。ハイライトがあまりにも多くのノイズを含む場合、モデルが適切に対象を捉えているか否かを人間が視覚的に検証することが困難となる。そのため、ノイズを含まない鮮明なハイライトこそが良い視覚的説明であると言える。

本研究では鮮明なハイライトを得るために前処理方法として活性固定化を提案する。既存の視覚的説明法ではハイライトを生成する際に、深層学習モデルのネットワーク全体の構造を利用している。これに対し、活性固定化ではまず前処理として深層学習モデルの認識に寄与している主要な構造だけを残して、残りの非主要部を枝刈りする。その後で既存の視覚的説明法でハイライトを生成する。このように前処理として活性固定化によりモデルの非主要部を枝刈りすることで、非主要部に由来するノイズを除去した鮮明なハイライトを生成できる。なお、活性固定化はモデルへの前処理であるため、任意の視覚的説明

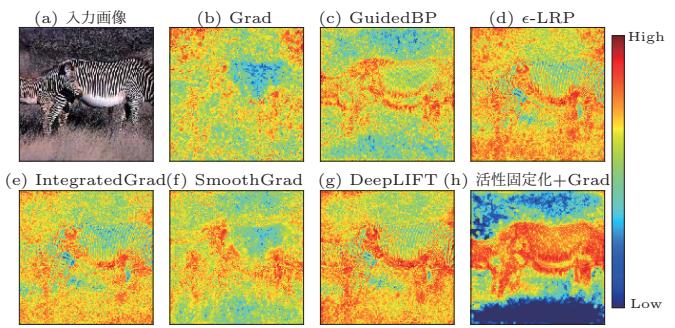


図1: 深層学習モデルの視覚的説明：入力画像(a)を深層学習モデルがシマウマと認識した根拠を既存の説明手法(b)–(g)及び提案法(h)でハイライトした結果。詳細は4.節を参照。

法と組み合わせができる。

図1(h)に活性固定化を用いた視覚的説明の一例を示す。既存の視覚的説明法、例えばGudedBP[2]や ϵ -LRP[3], IntegratedGrad[4], SmoothGrad[5], DeepLIFT[6]といった説明法ではノイズを多く含むハイライトが生成されるのに対し、これらの方法に活性固定化を前処理として導入することで鮮明なハイライトが生成できる。

2. 深層学習モデルとその視覚的説明法

本節では本研究で扱う深層学習モデルについて述べる。また深層学習モデルの視覚的説明法の代表的な方法として勾配に基づく方法を紹介する。

2.1 本研究で扱う深層学習モデル

本研究では入力 $x \in \mathbb{R}^p$ を受け取り出力 $y(x) \in \mathbb{R}^q$ を返す $L+1$ 層の q クラス分類のための深層学習モデルを考える。ここで、モデルの第 ℓ 層への入力とその出力をそれぞれ $z^{(\ell-1)}(x), z^{(\ell)}(x)$ とする。なお、モデルの出力 $y(x)$ や中間層の入出力 $z^{(\ell-1)}(x), z^{(\ell)}(x)$ は入力 x に依存する関数であることを明示するために引数に x を持つ関数として記述することとする。このとき、入力 x を受け取る第1層において $z^{(0)}(x) = x$ 、また出力 $y(x)$ を返す第 L 層において $z^{(L)}(x) = y(x)$ である。本研究では特にネットワークの各層で活性化関数として

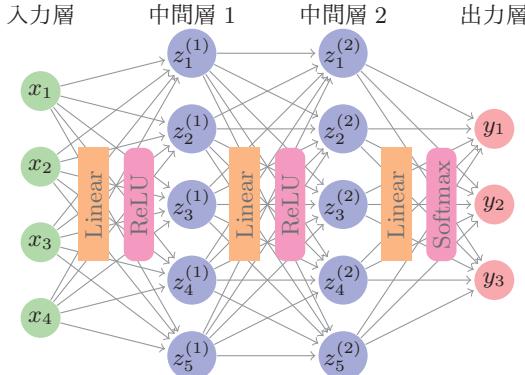


図 2: 4 層の深層学習モデルの例: Linear, ReLU/Softmax はそれぞれ各層の線形変換及び活性化関数である。図中の有向辺は出口側の変数が入口側の変数に依存する関数であることを示している。

Rectified Linear Unit (ReLU) を用いる以下のようなモデルを考える。

$$z^{(\ell-1/2)}(x) = W^{(\ell)} z^{(\ell-1)}(x) + b^{(\ell)}, \quad (1)$$

$$z_i^{(\ell)}(x) = f_{\text{ReLU}}(z_i^{(\ell-1/2)}(x)) = \max\{0, z_i^{(\ell-1/2)}(x)\}, \forall i. \quad (2)$$

ここで式 (1) の $W^{(\ell)}, b^{(\ell)}$ は第 ℓ 層での線形変換のパラメータであり、また式 (2) は $z^{(\ell-1/2)}(x)$ の各要素 $z_i^{(\ell-1/2)}(x)$ についてゼロ以下の値のものをゼロに切り上げる操作である。ただし、最終層 $\ell = L$ では式 (2) の ReLU の代わりに活性化関数として以下の Softmax 関数を用いる。

$$z_i^{(L)}(x) = \frac{\exp(z_i^{(L-1/2)}(x))}{\sum_{j=1}^q \exp(z_j^{(L-1/2)}(x))}. \quad (3)$$

図 2 に 4 層の深層学習モデルの例を示す。

2.2 深層学習モデルの視覚的説明法

既存の視覚的説明法の多くはモデルの勾配を元にハイライトを生成する。出力 $y(x)$ の j 番目の要素 $y_j(x)$ に寄与している入力 x の要素をハイライトする最も単純な方法は $y_j(x)$ に対して x の各要素 x_i に関する勾配を計算することである。これは、入力の要素 x_i を微小変化させたときに出力 $y_j(x)$ が大きく変化する要素 x_i こそがモデルの出力に大きく貢献している要素、つまり認識対象としてハイライトすべき対象であるとするアイディアである。具体的には、入力要素 x_i の出力 $y_j(x)$ への貢献度（ハイライトの大きさ）を以下で定義する。

$$\text{Highlight}(x_i) = \left| \frac{\partial y_j(x)}{\partial x_i} \right|. \quad (4)$$

式 (4) の勾配に基づく方法は計算が簡単な一方、ハイライトが多くのノイズを含むことが知られている。そこで、勾配に修正を加えた方法 [2, 3, 6] や勾配の平均を用いてノイズを低減させる方法 [4, 5] など、様々な改善法が提案されている。

3. 提案法：活性固定化

本節では、提案法の活性固定化の基本的なアイディアとその具体的方法について述べる。活性固定化を前処理として用い

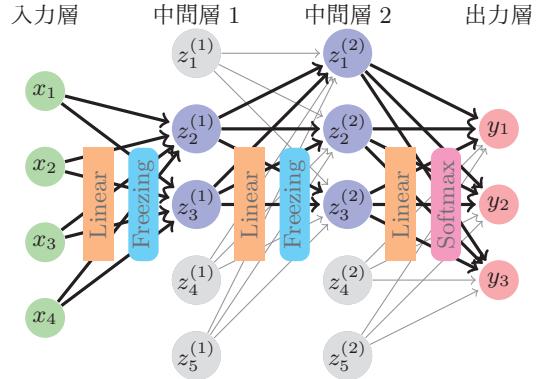


図 3: 活性固定化のイメージ図：活性化関数として活性固定化 (Freezing) を使うことで、入力 x の認識に寄与する主要な構造として太線部が抽出される。 $z_1^{(1)}, z_4^{(1)}, z_5^{(1)}$ 及び $z_2^{(2)}, z_3^{(2)}, z_5^{(2)}$ が活性が固定化されて定数に置き換えられたノードである。

ることで、既存の視覚的説明法により生成されるハイライトからノイズが低減され鮮明なハイライトの生成が可能となる。

以下、本研究では説明対象の深層学習モデルは学習済みであり、その構造及びパラメータは与えられているとする。また、説明対象の入力 x も与えられ固定されているとする。

3.1 基本的なアイディア

提案法の基本的なアイディアは『ハイライト生成時に、認識に寄与するネットワークの主要な構造のみを用いる』というものである。従来の説明法ではハイライトを生成する際に深層学習モデルのネットワーク全体を用いていた。しかし、このような全体を用いるアプローチではネットワーク内で認識にほぼ寄与していない非主要部の影響をもハイライト生成の際に加味してしまう。本研究ではこのような非主要部の影響がハイライトにノイズとして載っているのではないかという仮説を取る。そして、この仮説に基づいて、ハイライト生成の際に非主要部の影響を排除することで鮮明なハイライトを生成することを目指す。

図 3 に提案法の活性固定化のイメージを示す。活性固定化ではネットワーク中の弱い信号しか伝搬していない辺を認識に寄与しない非主要部として枝刈りする。この枝刈りにより、出力 $y(x)$ は強い信号だけを伝搬させるネットワークの主要部を通じてのみ入力 x に依存するようになる。このように弱い信号しか伝搬させない辺を刈ることで、入力 x の微小変化の影響の多くは出力 $y(x)$ までたどり着かなくなる。つまり、入力 x の微小変化の多くは出力 $y(x)$ に影響しなくなる。その結果、強い信号が伝搬している経路に対応する入力 x の要素の微小変化のみが出力 $y(x)$ に影響しそうな要素として、つまりハイライトすべき認識対象として特定できる。

3.2 活性固定化の方法

ここでは活性固定化の具体的な方法について述べる。前節で述べたように、本研究ではネットワークの主要な構造のみをハイライト生成に用いることを考える。活性固定化はこのような主要な構造のみをネットワークから抽出する前処理である。

本研究ではネットワークの各層において強い信号が伝搬しているノードとそうでないノードとを区別して考える。いま入力 x が与えられ固定されている状況を考えているので、各層におけるノードの関数 $z_i^{(\ell)}(x)$ の具体的な値 $v_i^{(\ell)} = z_i^{(\ell)}(x)$ を計算できる。そこで、この値 $v_i^{(\ell)}$ がある閾値よりも大きい場合に、そのノードは強い信号を伝搬させるネットワークの主要部だと

判定する。逆に、 $v_i^{(\ell)}$ が閾値より小さい場合にはそのノードは弱い信号しか伝搬させないネットワークの非主要部だと判定する。

活性固定化では、活性化関数として ReLU の代わりにネットワークの主要部、非主要部に対応した以下のような Freezing 関数を用いる。

$$\begin{aligned} z_i^{(\ell)}(x) &= f_{\text{Freezing}}(z_i^{(\ell-1/2)}(x); t^{(\ell)}) \\ &= \begin{cases} z_i^{(\ell-1/2)}(x) & \text{if } z_i^{(\ell-1/2)}(x) \geq t^{(\ell)}, \\ v_i^{(\ell)} & \text{otherwise.} \end{cases} \end{aligned} \quad (5)$$

ここで $t^{(\ell)}$ は Freezing 関数のパラメータで、ネットワークの主要部・非主要部を判定するための閾値である。閾値 $t^{(\ell)}$ の具体的な与え方については 4. 節にてその一例を示す。Freezing 関数は閾値 $t^{(\ell)}$ よりも強い信号が伝搬するノードをネットワークの主要部として信号を通し、閾値 $t^{(\ell)}$ よりも弱い信号が伝搬するノードでは活性化関数を定数で置き換える処理を行う。本研究では後者の定数による活性化関数の置き換えを活性の固定化と呼ぶこととする。Freezing 関数の特徴はこの固定化を通じて関数 $z_i^{(\ell)}(x)$ の入力 x への依存性を除去する、つまりネットワークの枝刈りを行う点にある。図 3 に活性固定化の適用例を示す。活性固定化で弱い信号が伝搬する部分の辺を刈ることで、十分大きな信号だけが伝搬しているネットワークの主要な構造を抽出したモデルを構築できる。

4. 実験

本節では活性固定化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 実験設定

本実験では説明対象の深層学習モデルとして Tensorflow のリポジトリで配布されている学習済みの VGG16 [7] を用いた。VGG16 は ImageNet [8] という 1000 クラスの画像を認識・分類するための深層学習モデルである。また、説明対象の入力画像には COCO データセット^{*1} を用いた。COCO データセットは 8 種類の動物の写真を含むデータセットである。

活性固定化で用いる各層の閾値 $t^{(\ell)}$ は以下のように設定した。まず、説明対象の入力 x が与えられたときの第 ℓ 層の各出力の値を $v_i^{(\ell)} = z_i^{(\ell)}(x)$ とする。このとき、これらの値 $v_i^{(\ell)}$ の多くは ReLU による活性化のためにゼロを取る。そこで、正の値を取る出力の集合 $\{v_i \mid v_i > 0\}$ を考え、この集合の $p\%$ 分位点を閾値 $t^{(\ell)}$ とした。本実験では p の値は 80 で固定した。これは、ReLU を通過した信号の経路の上位 20%のみを主要なネットワークとして抽出することに相当する。

実験では視覚的説明法として Grad (式 (4)) , GuidedBP [2], ϵ -LRP [3], IntegratedGrad [4], SmoothGrad [5], そして DeepLIFT [6] の計 6 種類を用いた^{*2}。そして、これら 6 種類の説明法全てについて活性固定化を前処理として用いなかつた場合と用いた場合とを比較した。

4.2 実験結果

図 4-6 に示す通り、全ての結果において活性固定化を用いることでノイズの少ない鮮明なハイライトを生成できた。

^{*1} cs231n.stanford.edu/coco-animals.zip

^{*2} Grad, GuidedBP, SmoothGrad の計算には <https://github.com/PAIR-code/saliency> を変更したものを、 ϵ -LRP, IntegratedGrad, DeepLIFT の計算には <https://github.com/marcoanconca/DeepExplain> をそれぞれ用いた。

図 4 では、活性固定化を前処理として施すことで、活性固定化なしの場合よりも鮮明なハイライトが生成されていることを確認できる。特に、活性固定化を用いない場合はハイライトが全体的に明るくなり認識対象のシマウマが適切にハイライトできていないのに対し、活性固定化を用いた場合はシマウマの胴体部分だけをハイライトすることに成功した。

図 5 においても、図 4 同様に活性固定化を使うことで認識対象の馬車をより鮮明にハイライトできた。特に大きな差異として、活性固定化を使った場合には地面のハイライトを認識に関係ないものとして十分低く抑えることができた。この点は、ノイズを多く含み全体的に明るいハイライトになりがちな既存の説明法に対する、活性固定化を用いることの明確な優位点であると言える。

図 6 は本来「馬」に分類されるべき画像が誤って「野球選手」として誤認識された画像とそのハイライトである。活性固定化を用いたハイライトでは、馬だけでなく背景にも大きな値が割り振られていることがわかる。このことから、深層学習モデルが背景を野球場と勘違いしたことで画像を「野球選手」と誤認識したと推測できる。これに対し、活性固定化を用いない場合では明確なハイライトが得られず、画像のどの領域が誤認識の原因となったかを推測することは困難である。この結果は鮮明なハイライトが得られる活性固定化の実用上の有用性を示している。

5. まとめ

本研究では、深層学習モデルの視覚的説明をより鮮明にするための方法として、活性固定化というモデルの前処理法を提案した。活性固定化では、入力中の認識対象をハイライトする前に入力と閾値に基づいてモデルのネットワークの枝刈りを行い、強い信号が伝搬する主要な構造だけを抽出する。そして、抽出された主要な構造だけを対象に既存の説明法で認識対象をハイライトする。このような前処理を導入してネットワーク中の認識にほぼ寄与していない部分に由来するノイズを除去することで、ハイライトを鮮明にすることができます。VGG16 を用いた実験では、活性固定化を既存の説明法と組み合わせても非常に鮮明なハイライトを生成できることが確認できた。また、このような鮮明なハイライトがモデルの誤認識の原因を特定するために有用であることも確認できた。

参考文献

- [1] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.
- [2] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv:1412.6806*, 2014.
- [3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS ONE*, 10(7):e0130140, 2015.
- [4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv:1703.01365*, 2017.

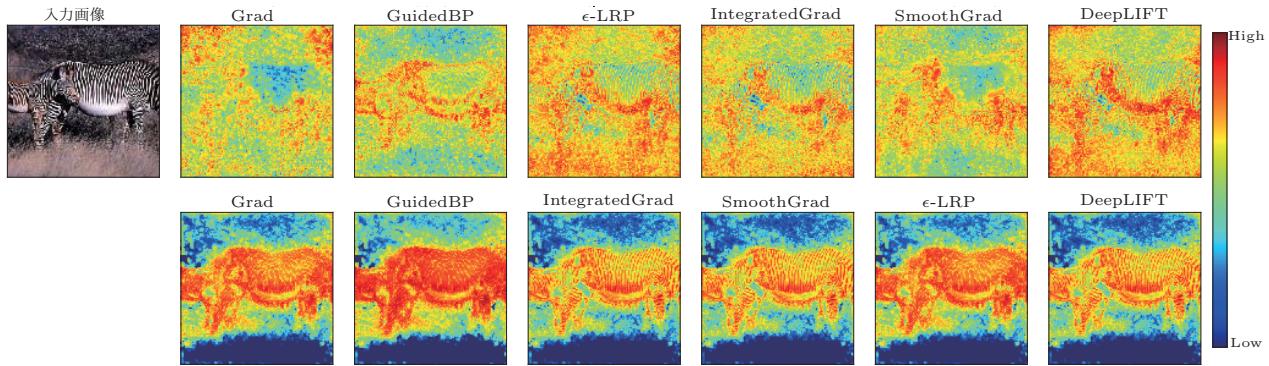


図 4: 活性固定化の適用結果 1 [認識結果 = シマウマ] : (上段) 活性固定化なし; (下段) 活性固定化あり。

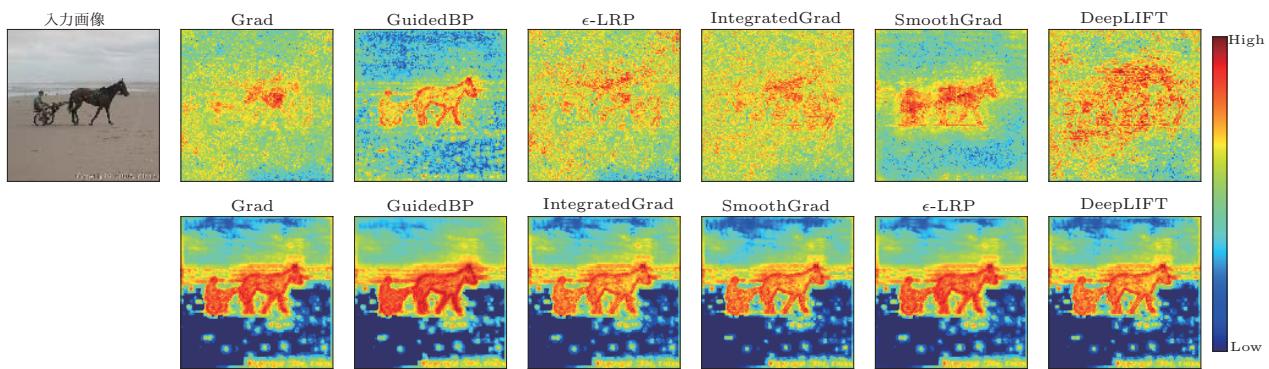


図 5: 活性固定化の適用結果 2 [認識結果 = 馬車] : (上段) 活性固定化なし; (下段) 活性固定化あり。

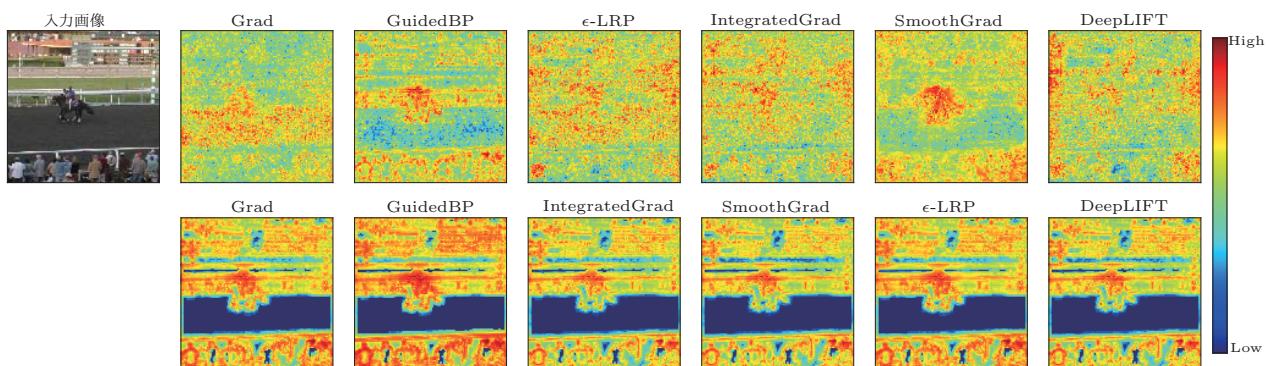


図 6: 活性固定化の適用結果 3 [認識結果 = 野球選手] : (上段) 活性固定化なし; (下段) 活性固定化あり。

- [5] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv:1706.03825*, 2017.
- [6] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of International Conference on Machine Learning*, pages 3145–3153, 2017.
- [7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.
- [8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.