Vol.2020-MPS-129 No.8
2020/7/27

受容野における最適刺激の可視化によるResNetの解釈の検討

小林 源太[1,a)]　庄野 逸[1,b)]

概要：画像認識で利用される手法の一つに，Deep Convolutional Neural Network（DCNN）がある．DCNNは，CNNの隠れ層を深くすることで特徴の表現力を大幅に向上させたモデルである．CNNのアーキテクチャは，哺乳類の視覚皮質のモデルに基づいている．一方，生物学的な観点でなく，学習手法の観点から発展したスキップ接続のあるResidual Network（ResNet）と呼ばれるモデルが提案されている．本研究では，ImageNetの分類タスクにおけるResNetの受容野を調査し，ResNetには方向選択ニューロンと二重反対色ニューロンがあることを確認した．さらに，我々はResNetの最初の層に不活性ニューロンが存在することと，それらが分類に寄与していることを示した．

キーワード：深層畳込みニューラルネットワーク，Residual Network，視覚野，受容野

Interpretation of ResNet by Visualization of Preferred Stimulus in Receptive Fields

Genta Kobayashi[1,a)]　Hayaru Shouno[1,b)]

Abstract: The Deep Convolutional Neural Network (DCNN) is the *de facto* standard in the image classification task. By deepening the hidden layers of Convolutional Neural Network (CNN), the DCNN greatly improves its expressive power of features. The architecture of the CNN is inspired from a model of the primary visual cortex of mammals. On the other hand, there is a model called Residual Network (ResNet) that has a skip connection, which comes from the viewpoint of efficient learning. Thus, ResNet is an advanced model in terms of the learning method, however, it has not been interpreted from a biological viewpoint. In this research, we investigate the preferred stimulus in receptive fields of ResNet neurons, which is designed for on the classification task with ImageNet dataset. We find that ResNet has orientation selective neurons and double opponent color neurons. In addition, we suggest that some inactive neurons in the first layer of ResNet affect the classification task.

Keywords: Deep Convolutional Neural Network, Residual Network, Visual Cortex, Receptive Field

1. 背景

近年，Deep Convolutional Neural Network (DCNN) は画像や音声といった信号処理の幅広い分野で利用されている．特に画像分類問題において，DCNNは既存の手法よりも高い性能を示している．DCNNはConvolutional Neural Network（CNN）の隠れ層を多層にすることによって，表現能力が高くなったモデルの一つである．CNNの特徴は畳込み層とプーリング層を積層していることであり，その層の構造は哺乳類の初期視覚野にある単純型細胞と複雑型細胞に基づいている [2]．Krizhevskyら [5] は勾配学習と大量のデータで学習したDCNNを用いて，自然画像の分類問題におけるDCNNの有効性を示し，DCNNは幅広い分野で使われるようになった．

DCNNの成功は，DCNNを理解する研究を加速させた．

[1] 電気通信大学
The University of Electro-Communications
[a)] genta-kobyashi@uec.ac.jp
[b)] shouno@uec.ac.jp

工学的な観点から，その研究の主流は勾配の逆伝播を用いた DCNN の内部表現の可視化に基づいている [6]．また，DCNN の基礎構造は生物学的な観点 [2] から触発されているが，ここ十年に提案されている DCNN モデルは非生物学的な発展をしている．例えば，Residual Network (ResNet) [3] は勾配学習をうまくために発展したモデルである．

本研究では，我々は受容野における最適刺激の観点に着目し，ResNet の解釈を検討する．受容野は視覚システムの基本的な概念であり，細胞が反応しうる視覚入力領域の一部分を意味する．その最適刺激とは，ある細胞が強く反応する刺激のことである．我々は ResNet の性質を明らかにするために，この考えを用いる．

2. 手法

2.1 Residual Network

He ら [3] は Residual Network（ResNet）の概念を提案し，ResNet34 といったいくつかのモデルを示した．ResNet の特徴はスキップ接続と呼ばれる特徴的な構造を含んでいることである．スキップ接続のアイデアは層による写像を非線形なものと線形なものに分割することである．入力ベクトルを $\boldsymbol{x}$，出力ベクトルを $\boldsymbol{y}$，非線形な写像部分を $F(\cdot)$ とすると，多くのスキップ接続は式 (1) のように表せる．

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{F}(\boldsymbol{x}) \tag{1}$$

$\boldsymbol{x}$ と $\boldsymbol{F}(\boldsymbol{x})$ の次元が異なる場合，$\boldsymbol{x}$ は線形写像により変換される．He らの ResNet では，この線形写像は非線形な活性化関数を持たない畳込み層で構成される．我々はスキップ接続を除外した ResNet のモデルを PlainNet として，ResNet の比較に用いる．

2.2 CNN の受容野

CNN の文脈での受容野は入力領域の一部分であり，最適刺激は入力画像パッチの一部分である．**図 1** に受容野の概観図を示す．左の四角形から入力，特徴マップ 1，そして特徴マップ 2 を表している．各特徴マップの人工細胞は二次元格子状の位置を持つ．特徴マップのある細胞を選んだ時，それより低次の接続されている領域を入力まで求めることができる．従って，図 1 に示すように，特徴マップ 2 の赤い細胞の最適刺激を求めることができる．

Zeiler ら [8] は DCNN の受容野における最適刺激のサンプルを示し，各層の特徴を報告した．そのサンプルを示すことは DCNN の内部表現を理解する単純な方法である．我々は人工細胞の性質を調査するために，この受容野における最適刺激を用いる．$\boldsymbol{x}$ を画像とすると，受容野は空間的な位置の集合である．我々は画像 $\boldsymbol{x}$ の受容野 $\boldsymbol{r}$ における受容野画像を $\boldsymbol{x}[\boldsymbol{r}]$ として形式的に書き表す．

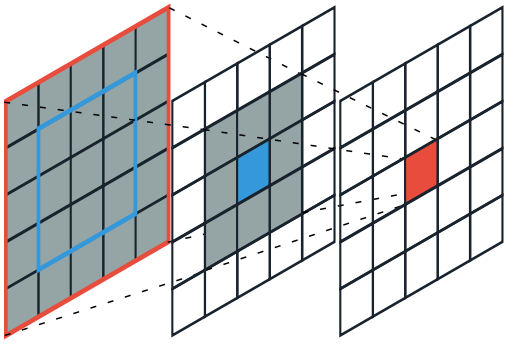


図 1: 受容野の概観．各黒い縁の四角形は細胞であり，入力にある青い枠内の領域は特徴マップ 1 にある青い細胞に対応する受容野である．入力にある赤い枠内の領域は特徴マップ 2 にある赤い細胞に対応する受容野である．

Fig. 1 Overview of receptive field. Each black boder rectangle is a neuron. The area inside the blue border on input is the receptive field and corresponds to the blue neuron in feature map 1. The area inside the red border on input is the receptive field and corresponds to the red neuron in feature map 2.

2.3 勾配を用いた可視化

多くの研究者は DCNN を理解するために勾配に基づいた可視化手法を使っている [6]．最初期の手法として，Erhan ら [1] の活性値最大化法があり，Simonyan ら [6] はその手法を DCNN へ適用した．活性値最大化法は最適化問題として細胞の入力が最大となる入力画像を求めるものである．モデルのパラメータを $\boldsymbol{\theta}$ とし，$f(\boldsymbol{\theta}, \boldsymbol{x})$ を入力 $\boldsymbol{x}$ を与えた時のモデルのニューロンの活性値とする．$\boldsymbol{\theta}$ を固定した時，活性値最大化法は式 (2) で表せる．

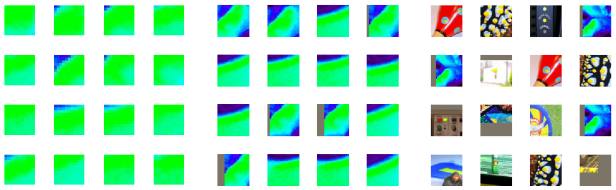
$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \{f(\boldsymbol{\theta}, \boldsymbol{x}) - \lambda \|\boldsymbol{x}\|_2\} \tag{2}$$

入力画像の方向成分について興味があるので，L_2 ノルム制約を入れる．この手法は勾配上昇法によって求められ，微分可能な全てのモデルに適用できる．一方で，最適化方法に依存して局所的な最適解になる可能性がある．

3. 実験と結果

3.1 ResNets の学習

我々は He ら [3] と Szegedy ら [7] の訓練方法を参考に ImageNet データセットで ResNet34 と PlainNet34 を学習させた．前処理として，入力画像を色チャネルについて白色化した．モデルの重みの最適化に，重み減衰が 10^{-4}，初期の学習率を 0.01 で，モーメント 0.9 の確率勾配降下法を適用する．この学習率は 30 エポックごとに 10 で割られ，訓練エポックは 90 エポックで，256 ミニバッチ学習をする．訓練時，データ拡張方法として Szegedy ら [7] の方法を用いた．



(a) 最初のプーリング層. 受容野の大きさは 11 × 11 である. Pooing layer in layer 1. Size of receptive field is 11 × 11.
(b) 3 層目の畳込み層. 受容野の大きさは 27 × 27 である. Conv. layer in layer 3. Size of receptive field is 27 × 27.
(c) 7 層目の畳込み層. 受容野の大きさは 59 × 59 である. Conv. layer in layer 7. Size of receptive field is 59 × 59.

図 2: ResnNet34 のチャネル 18 の上位 16 の最適刺激のサンプル.

Fig. 2 Samples of the top 16 preferred stimulus images of channel 18 in ResNet34.

3.2 受容野における最適刺激の分析

我々は最適刺激を見つけるために，まず，ImageNet の評価画像セット X を ResNet に入力する．その後，各層において活性値の降順になるように刺激を揃える．$\boldsymbol{r}_i$ を受容野とし，受容野画像（刺激）を与えた時の i 番目の細胞の活性値を $f_i(\boldsymbol{x}[\boldsymbol{r}_i])$ とすると，正値評価画像 X^+ における平均受容野画像を式 (3) のように書き表せる．我々はこれを平均最適刺激として分析に用いる．

$$\bar{\boldsymbol{x}}^i = \frac{1}{N} \sum_{\boldsymbol{x} \in X^+} \boldsymbol{x}[\boldsymbol{r}_i] \tag{3}$$

正値評価画像は，ある細胞が正の値を出力した時の評価画像セット中の受容野画像の集合であり，式 (4) のように表せる．

$$X^+ = \{\boldsymbol{x} \in X \mid f_i(\boldsymbol{x}[\boldsymbol{r}_i]) > 0\} \tag{4}$$

図 2 と 4 に，ResNet34 と PlainNet34 のある細胞の最適刺激の上位 16 個を示し，図 3 と 5 にその細胞に対応する最初の畳込み層の重みと平均最適刺激を示す．最適刺激のサンプルから，両方の DCNN は高次の層になるにつれて多様な刺激に反応することがわかる．図 2(c) と 4(c) は一貫性のないサンプルに見えるが，図 3(d) と 5(d) から中心領域に特徴があることがわかる．PlainNet と比べて，ResNet は異なる層の同じチャネルの特徴は似るという性質をもつことがわかる．平均最適刺激は大まかな傾向を捉えることができるが，細胞の細かな傾向を見つけることは困難であることがわかる．

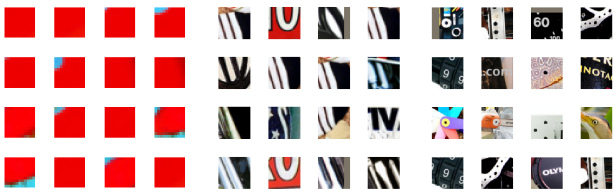
3.3 活性値最大化法を用いた可視化

主流な手法と比較するために，活性値最大化法を ResNet34 に適用し，図 6 にその結果を示す．具体的には，特徴マップの中心に位置する細胞に対して，重み減衰 10^{-6} と学習率 0.1 で Adam [4] と呼ばれる勾配法で最適化

(a) 最初の畳込みフィルタ. First conv. filter.
(b) 最初のプーリング層の平均最適刺激. Mean preferred stimulus in pooling layer in layer 1.
(c) 3 層目の畳込み層の平均最適刺激. Mean preferred stimulus in conv. layer in layer 3.
(d) 7 層目の畳込み層の平均最適刺激. Mean preferred stimulus in conv. layer in layer 7.

図 3: ResNet34 のチャネル 18 の最初の畳込みフィルタと平均最適刺激.

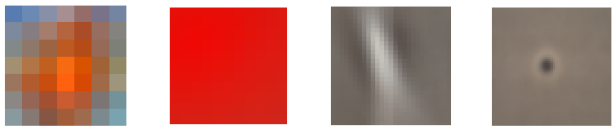
Fig. 3 First convolutional filter and mean preferred stimulus images of channel 18 in ResNet34.



(a) 最初のプーリング層. Pooling layer in layer 1.
(b) 3 層目の畳込み層. Conv. layer in layer 3.
(c) 7 層目の畳込み層. Conv. layer in layer 7.

図 4: PlainNet34 のチャネル 19 の上位 16 の最適刺激のサンプル.

Fig. 4 Samples of the top 16 preferred stimulus images of channel 19 in PlainNet34.

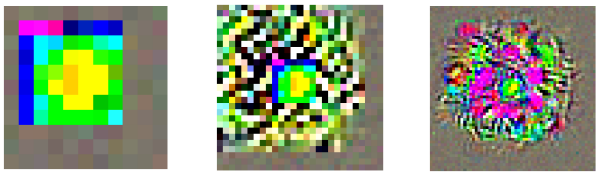


(a) 最初の畳込みフィルタ. First conv. filter.
(b) 最初のプーリング層の平均最適刺激. Mean preferred stimulus in pooling layer in layer 1.
(c) 3 層目の畳込み層の平均最適刺激. Mean preferred stimulus in conv. layer in layer 3.
(d) 7 層目の畳込み層の平均最適刺激. Mean preferred stimulus in conv. layer in layer 7.

図 5: PlainNet34 のチャネル 19 の最初の畳込みフィルタと平均最適刺激.

Fig. 5 First convolutional filter and mean preferred stimulus images of channel 19 in PlainNet34.

を行った．また，初期入力は 0 で埋められた画像で，31 回繰り返し行う．



(a) 最初のプーリング層のチャネル 18. Channel 18 in conv. layer in first layer.
(b) 3 層目の畳込み層のチャネル 18. Channel 18 in conv. layer in layer 3.
(c) [7 層目の畳込み層のチャネル 18. Channel 18 in conv. layer in layer 7.

図 6: ResNet34 の活性値最大化法による最適化結果の例.
Fig. 6 Examples of visualizations by activation maximization for the neuron in ResNet34.

図 3 と 6 を比較すると，最適化による結果と平均画像が似ていることがわかる．高い層の可視化結果について，活性値最大化法の結果は中心が黄色に反応し，その周辺が赤色に強く反応するといった細かい傾向がわかる．一方，平均最適刺激は中心が黄色に強く反応するということのみがわかる．

3.4 不活性ニューロン

評価画像において，我々は最初のプーリング層のいくつかのチャネルがいかなる入力を与えても，0 のみを出力していることを発見した．我々はこのチャネルの細胞を不活性ニューロンと呼ぶ．表 1 に ResNet34 と PlaiNet34 の不活性ニューロンの数を数え上げたもの示す．両モデルの不活性数を比較すると，ResNet34 は PlainNet34 の不活性ニューロンの数が多いことがわかる．

不活性細胞が識別に与える影響を調査するために，不活性ニューロンにノイズを与える 2 種類の識別実験を行った．一つは全ての不活性ニューロンにノイズを加えるもので，もう一方は毎ミニバッチごとにランダムに選ばれた不活性ニューロンにノイズを加えるものであり，我々はノイズとして，不活性ニューロンの各空間的な次元に $\mathcal{N}(0,1)$ からサンプルされたガウスノイズ x に ReLU（$\max(x,0)$）を施したものを用いる．ΔL は全てにノイズを加えた時の評価損失からノイズを加えない時の評価損失を引いた値とし，ΔL_{rnd} はランダムにノイズを加えた時の評価損失からノイズを加えない時の評価損失を引いた値とし，表 1 の右側に識別実験の結果を示す．ResNet34 の ΔL と ΔL_{rnd} は PlainNet34 のものより大きく正の値であることから，ResNet34 の不活性ニューロンは識別タスクにおいて影響を与えていることがわかる．

4. 結論

我々は ResNet34 を理解するために，最適刺激を用いた分析と活性値最大化法を ResNet と PlainNet に適用した．

表 1: 不活性ニューロンの数と最初の層に存在する不活性ニューロンが識別に与える影響．評価画像セットにおける ResNet34 と PlaiNet34 の結果である．
Table 1 Count of the inactive neurons and effect of the inactive neuron in first max-pooling layer for validation dataset in ResNet34 and PlainNet34.

モデル	不活性数	ΔL	ΔL_{rnd}
ResNet34	13	$1.26962e+0$	$2.40560e-2$
PlainNet34	2	$-1.66893e-6$	$-8.34465e-7$

両手法は低次の層の傾向をつかむことはできるが，最適刺激の分析は活性値最大化法のように高次の層の傾向をつかむことが困難である．また，我々は識別に影響を与える不活性ニューロンが ResNet34 に存在することを発見した．この現象はスキップ接続によるチャネル共有が原因であると我々は推測している．一つの仮説として，あるチャネルが最初の畳込み層の特徴と似ない特徴を使うために不活性ニューロンが存在することが考えられる．今後の研究として，我々は高次の層の分析に適用できる手法を考える必要と，我々の仮説を支持する根拠を示す必要がある．

参考文献

[1] Erhan, D., Bengio, Y., Courville, A. and Vincent, P.: Visualizing higher-layer features of a deep network, *University of Montreal*, Vol. 1341, No. 3, p. 1 (2009).
[2] Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biological Cybernetics*, Vol. 36, No. 4, pp. 193–202 (online), DOI: 10.1007/BF00344251 (1980).
[3] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (online), DOI: 10.1109/CVPR.2016.90 (2016).
[4] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
[5] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25* (Pereira, F., Burges, C. J. C., Bottou, L. and Weinberger, K. Q., eds.), Curran Associates, Inc., pp. 1097–1105 (2012).
[6] Simonyan, K., Vedaldi, A. and Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (2013).
[7] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going deeper with convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9 (2015).
[8] Zeiler, M. D. and Fergus, R.: Visualizing and understanding convolutional networks, *European conference on computer vision*, Springer, pp. 818–833 (2014).