6U-06

# 局所応答正規化を導入した視覚野モデルによる Adversarial Examples へのロバスト性の向上

宮澤 隆太 † 服部 元信 ‡
山梨大学 医工農学総合教育部 † 山梨大学 大学院総合研究部 ‡

## 1 はじめに

現在, 人間の脳の構造を模倣したニューラルネットワークが様々な分野で利用されている. ニューラルネットワークを用いることで分類, 識別タスクにおいて極めて優れた性能を実現することが可能になっている.

一方, ニューラルネットワークにおいては, Adversarial Examples[1] と呼ばれる問題が明らかになっている. Adversarial Examples とは, 図 1 に示すような元のデータに対し微小なノイズを加えただけにも関わらず, 誤識別を起こしてしまうデータである. 人間には 2 枚の画像は同じ画像に見えるが, Adversarial Example にはノイズが加えられており誤識別が起きている.

(a) 通常画像 (車)　(b) Adversarial Example (飛行機)

図 1: Adversarial Examples の例. () 内は識別結果.

これまで,Adversarial Examples に関する研究は多く行われてきたが原因については明らかになっていない. 本研究では, Adversarial Examples の対策として提案されている視覚野モデルに局所応答正規化を導入することで, さらにロバスト性を向上させるモデルを提案する.

## 2 関連研究

ここでは提案手法で用いる視覚野の機能を導入したモデルと局所応答正規化について説明する.

### 2.1 VOneNet[2]

このモデルは脳の視覚野の情報処理を模倣することで, Adversarial Examples への耐性を獲得した手法であり, 学習を行わない VOneBlock とそれに続く通常の畳み込み層から構成される. VOneBlock には次の 3 つの要素が存在する.

#### 2.1.1 ガボールフィルタ

ガボールフィルタとは, 生物の視覚野をモデル化する際に使われるフィルタであり方向線分の特徴に反応するフィルタである. 通常の畳み込みニューラルネットワークではフィルタの重みを学習するが, VOneNet では学習を行わず, 生物学的知見から決められたパラメータのフィルタを畳み込みに用いる. このフィルタには方位, 位相, 周波数などのパラメータが存在する. このフィルタを用いることで, 物体のテクスチャではなく, 形状の特徴を捉えることができる.

#### 2.1.2 2 種類の細胞

視覚野には単純型細胞と複雑型細胞と呼ばれる 2 種類の細胞がある. それぞれの細胞はガボールフィルタを持ち畳み込み演算を行う. 単純型細胞はガボールフィルタと, 方位, 位相, 周波数が等しい入力に対して反応する.

一方で複雑型細胞は方位と周波数が等しい入力に反応する. そのため VOneNet では, 位相が異なるガボールフィルタを持つ 2 つの単純型細胞の出力を合わせることで, 複雑型細胞を実装している.

#### 2.1.3 確率性

VOneBlock では確率的に応答するニューロンを用いている. これにより各ニューロンは, 同じ入力であっても毎回異なる反応を示すことになる. この確率性を導入することで, Adversarial Examples の生成が難しくなりロバスト性が向上する.

### 2.2 局所応答正規化 [3]

局所応答正規化 (Local Response Normalization) とは, 式 (1) で計算される正規化である. 正規化とはデータを扱いやすい範囲に加工することである. 局所応答正規化は脳における近傍ニューロンからの抑制に着想を得た正規化であり, 出力を近傍のチャネルで除算するという処理である. ここで $a$ は正規化前の出力, $b$ は正規化後の出力, $c$ は正規化するチャネル, $x$ と $y$ は畳み込みの

---

「Improved robustness to Adversarial Examples with a visual field model introducing local response normalization」

† Integrated Graduate School of Medicine, Engineering and Agricultural Sciences, University of Yamanashi
‡ Faculty of Interdisciplinary Research, University of Yamanashi

出力の座標, $C$ は全チャネル数, $n$ は正規化の計算に含む近傍チャネル数である. $\alpha, \beta, k$ はあらかじめ設定するパラメータである. 局所応答正規化では, 近傍ニューロンから抑制を行うことで特徴がぼやけることを防ぎ汎化性能を向上させる.

$$b_{c,x,y} = \frac{a_{c,x,y}}{(k + \alpha(\sum_{c_n=max(0,c-\frac{n}{2})}^{min(C-1,c+\frac{n}{2})}(a_{c_n,x,y})^2))^{\beta}} \quad (1)$$

## 3 提案手法

本研究では視覚野のモデルに重みづけ局所応答正規化を導入したモデルを提案する. 視覚野モデルの単純型細胞と複雑型細胞の出力に対して局所応答正規化を行う. 視覚野モデルの畳み込みフィルタはパラメータで設定しているため, 似た特徴に反応するフィルタを持つニューロンの特定が容易である. そのため似た特徴に反応する近傍ニューロンからの抑制が実装できる.

### 3.1 重みづけ局所応答正規化

一般的な局所応答正規化は近傍ニューロンから等しく抑制を受ける. 一方で, 私たちの脳の細胞は似た特徴に反応する細胞から強い抑制を受けると考えられている. そこで本研究では似た特徴に反応するニューロンからの抑制が大きくなるように重みづけを導入し, 式 (2) で正規化の計算を行う. $w_{i,j}$ はチャネル $i$ からチャネル $j$ への抑制の重みである. ニューロンはそれぞれ方位, 位相, 周波数のパラメータが設定され, このパラメータにより畳み込みのフィルタが決まる. そのためこのパラメータの値が近いほど似たフィルタになり, 式 (3) で重みを計算することで似た特徴に反応するニューロンからの重み $w$ を大きくすることができ, 抑制の大きさを調整することができる.

$$b_{c,x,y} = \frac{a_{c,x,y}}{(k + \alpha(\sum_{c_n=0}^{min(C-1,n)} w_{c,c_n}((z_c)_{c_n,x,y})^2))^{\beta}} \quad (2)$$

$$w_{i,j} = \frac{1}{\sqrt{(o_i - o_j)^2 + (p_i - p_j)^2 + (f_i - f_j)^2}} \quad (3)$$

$z_c$ は正規化するチャネル $c$ に対して重みの大きい順に並べられた正規化前のニューロンの出力, $o$, $p$, $f$ は方位, 位相, 周波数のパラメータである.

## 4 実験

VOneNet(Vone), 局所応答正規化を導入した VOneNet(Vone+L) 重みづけ局所応答正規化を導入した VOneNet(Vone+WL) で Adversarial Examples へのロバスト性を比較する実験を CIFAR10 データセットを用いて行った. 図 2 は,PGD を用いて Adversarial Examples を生成した際のノイズサイズに対する各手法の精度を示している. この図から分かるように通常の VOneNet と比べ局所応答正規化を導入するだけでは精度の変化がないが, 局所応答正規化に重みづけを導入することでロバスト性が向上している.

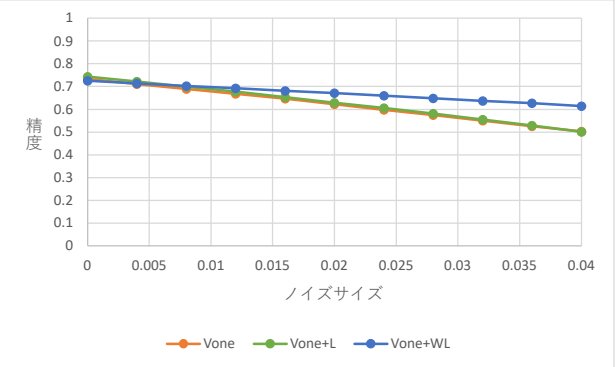


図 2: VOneNet のロバスト性比較 (CIFAR10).

## 5 結論

本研究では, Adversarial Examples へのロバスト性を向上させるために視覚野のモデルに重みづけ局所応答正規化を導入するモデルを提案した. 結果として, 局所応答正規化のないモデルに比べロバスト性を向上させることができた. 通常の局所応答正規化では, 正規化のための近傍を決める際に似た特徴に反応するニューロンであるかを無視するため, 出力が小さくなるだけで精度の変化がないと考えられる. 一方で, 重みづけ局所応答正規化では似た特徴に反応するニューロンから抑制を受けるため, 適切な抑制がかかり, 捉えた特徴を失わずにノイズの影響を抑えることができ, ロバスト性の向上に繋がったと考えられる.

## 参考文献

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. : Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199, 2013.

[2] Dapello, J., Marques, T., M. Schrimpf, M., Geiger, F., D. Cox, D., and DiCarlo, J.J. : Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations, NeurIPS 2020, vol. 33, pp. 13073–13087.

[3] Krizhevsky, A., Sutskever, I. and Hinton, G, E., : Imagenet classification with deep convolutional neural networks, Communications of the ACM, vol 60, no. 6, pp. 84-90, 2017.