Attention branch の匿名化による Adversarial examples に頑健な ABN

ER17079 山川翔太郎　　指導教授：藤吉弘亘

1. はじめに

Attention branch network (ABN) [1] は，学習時または推論時にネットワークの注視領域を利用することで，高精度な認識かつ判断根拠の視覚的な説明が可能である．しかし，人工的な摂動を付与した画像である Adversarial examples (AEs) をネットワークへ入力すると，高い信頼度で誤認識する．そこで，本研究では ABN の Attention branch を匿名化することにより，AEs に対して誤認識を緩和する手法を提案する．

2. 関連研究

ABN は，入力から特徴を抽出する Feature extractor (FE)，注視領域を出力する Attention branch (AB)，注視領域を反映した特徴マップから最終的な出力をする Perception branch (PB) で構成されている．注視領域を学習時，及び推論時に利用することで，高精度な認識が可能である．また，注視領域は入力画像の認識結果に対する視覚的理解の促進につながる．

Adversarial attack は，人が知覚困難な摂動を画像に付与することで，Convolutional neural network (CNN) の誤認識を誘発する．Fast Gradient Sign Method (FGSM) [2] は，CNN の認識誤差 $\mathcal{L}(x, y, \theta)$ を入力画像 x の各画素に関して微分することで勾配を求める．そして，求めた各画素の勾配を Sign 関数によって符号を抜き出すことで，摂動の単位ベクトルを獲得する．AEs である $\hat{x}$ を作成するには，単位ベクトルに摂動の強度を表す $\epsilon \in [0, 255]$ を乗算して x に加算する．摂動を加算することで，誤差を最大化する AEs を作成することが可能となる．FGSM の一連の流れは，式 (1) によって表現することができる．

$$\hat{x} := x + \epsilon \cdot \mathrm{sign}\left(\nabla_x \mathcal{L}\left(x, y, \theta\right)\right) \tag{1}$$

3. 提案手法

ABN は Adversarial attack を想定した設計でないため，AEs を入力すると著しく性能が劣化する．そこで本研究では，AB を匿名化する．そして，FE と PB を攻撃対象として AEs を作成する．AEs の推論時は AB を追加して最終結果を求める．提案手法は以下に示す 3 ステップから成る．

Step1 ABN を通常の学習データを用いて訓練して，優秀になった AB を切り離して匿名化する．

Step2 PB の出力と教師信号との誤差を入力画像に関して微分することで誤差を最大にする単位ベクトルを求める．

Step3 匿名化した AB を付け足したネットワークに，AEs を入力し，注視領域を求める．そして，注意領域を重み付した特徴マップを PB に入力して推論を行う．

図 1 に Step2 と Step3 の一連の流れを示す．

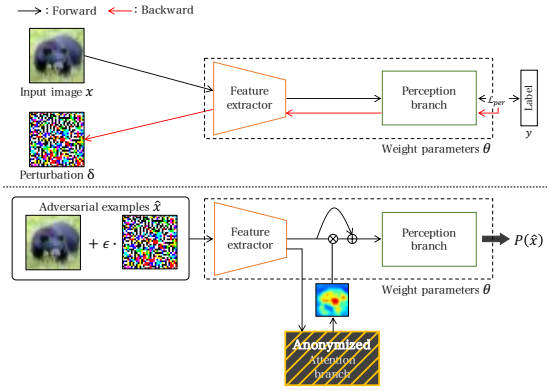


図 1：提案手法の流れ

4. 評価実験

提案手法の有効性を示すために，定量的な認識性能の評価及び定性的な注視領域の評価をする．

4.1. 実験条件

本実験では，データセットとして 100 クラスの自然画像を含む CIFAR-100 を用いる．CIFAR-100 は画像サイズが 32×32 ピクセルの RGB 画像で，5 万枚が学習用，1 万枚が推論用である．ABN のベースネットワークとして 110 層の ResNet を使用する．ミニバッチサイズは 128 として 300 epoch 学習する．摂動の作成手法は，$\epsilon = 8$ の FGSM とする．比較手法は通常の ABN の推論結果及び，ABN を AEs で攻撃した結果とする．

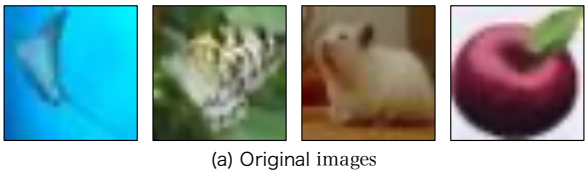
4.2. 実験結果

表 1 に示すように，ABN は AEs の影響により認識性能が著しく低下した．一方，提案手法は，匿名化した AB が出力した注視領域を特徴マップに反映することで，60 ポイント以上の性能向上を確認した．AEs を用いない場合 (ABN) の推論結果と比較すると，提案手法は 11 ポイント程度の劣化に留めることを可能とした．

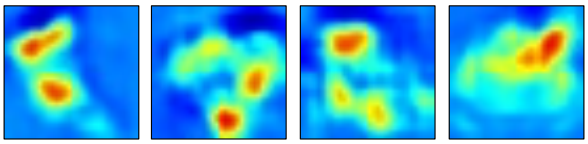
表 1：認識精度の比較 (%)

	ABN	ABN+AEs	Ours
top-1	77.18	1.82	**65.22**
top-5	93.54	9.12	**87.49**

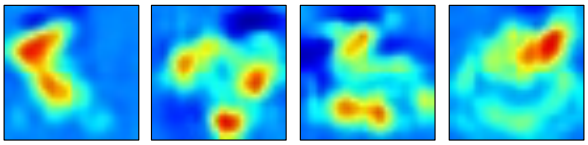
注視領域を図 2 に示す．図 2 より，提案手法は AEs を入力しているにも関わらず，発火する領域は ABN とほぼ同じであることが確認できる．この結果から，AB は AEs によって外部から攻撃されても，入力画像に対して正しい領域を注視することができる．また，AEs により FE が変動した特徴量を計算しても，正しい注視領域を強調することで，性能を維持できることがわかった．



(a) Original images



(b) ABN



(c) Ours

図 2：注視領域の例

5. おわりに

本研究では，人工的な摂動である Adversarial examples(AEs) に頑健な Attention branch network(ABN) を提案した．実験より，AEs が入力されても注視領域の変動が微小であることを確認した．また，認識結果は入力画像に対して正確な領域を強調することで，ある程度維持できることが判明した．今後は，本提案手法の更なる分析をする予定である．

参考文献

[1] H. Fukui, *et al.*, “Attention Branch Network: Learning of Attention Mechanism for Visual Explanation”, CVPR, 2019.

[2] l. Goodfellow, *et al.*, “Explaining and Harnessing Adversarial Examples”, ICLR，2014.