CF-014

敵対的標識：標識識別器に対する脆弱性評価と敵対的訓練による対策
Adversarial Sign: Assessing Vulnerability and Countermeasures in Sign Recognition

林 海斗 *1
Kaito Hayashi
内田 真人 *1
Masato Uchida

1 序論

近年，機械学習モデルの社会への実装が急速に広まっている．その 1 つである「自動車運転支援システム」は，道路標識を検知して識別し，その結果を運転者に知らせることで，運転の補助や違反している場合には，警告を行うことを可能としている．しかし，標識識別器の精度は完全ではない．街中に掲示されている一部の飲食店のロゴが実在する標識と誤認識される例が報告されている．国内自動車メーカである本田技研工業株式会社のウェブページでは，図 1 に示すように，運転支援システムが企業の看板や街中に掲げられるのぼり旗を標識と誤認識する現象を公表している [1].

運転支援システムに搭載される標識識別器が誤認識を引き起こすと，誤った表示や警告によって，運転者が混乱し，周囲へ危険を及ぼす可能性がある．したがって，誤認識を引き起こすロゴやシンボルを特定し，その性質を明らかにした上で誤認識を防ぐための対策が求められる．本研究では，標識識別器が標識でないものを誤認識する現象に着目し，既存のロゴやシンボルの誤認識に関する実態調査と誤認識を防ぐ対策に焦点を当てる．

標識ではない掲示物が標識であると誤認識される現象については，次の 2 つの点で未解明の課題がある．1 つ目は，標識でないものが誤認識される現象が現実的にどの程度の割合で発生し得るのかが不明なことである．そのため，既存のロゴやシンボルがどの程度誤認識されるかを調べる必要がある．また，誤認識を誘発するように意図的に作成した架空のロゴを標識識別器に入力する攻撃を行った場合，対策が不十分な標識識別器では容易に誤認識を引き起こすと予想される．したがって，既存，架空に関わらずロゴやシンボルに対する標識への誤認識リスクを定量的に評価する必要がある．

2 つ目は，人間の知覚と機械学習モデルの認識に関する乖離である．標識識別器により誤認識されたロゴ等が人間によって視認されたときに，ロゴのデザインが実在する標識と構造的に類似していたとしても，人間が誤認または問題視することは少ないと考えられる．実在する標識と視覚的な特徴が類似しているものの看板やのぼりなどに名称と共にロゴが掲示されている場合，人間が運転中に標識だと誤認するケースは稀であると考えられる．一方で，機械学習モデルではその類似性に過敏に反応し誤認識を引き起こすと考えられ，人間とモデルの認識の乖離が脅威になっている．また，日本国内の道路交通法では第 76 条に「1. 何人も、信号機若しくは道路標識等又はこれらに類似する工作物若しくは物件をみだりに設置してはならない。2. 何人も、信号機又は道路標識等の効用を妨げるような工作物又は物件を設置してはならない。」という規定がある．一般的に，企業ロゴはこの令に違反していない．しかし，本田技研工業株式会社のウェブページで紹介されているように，そのような企業ロゴであっても，自動車運転支援システムに搭載された

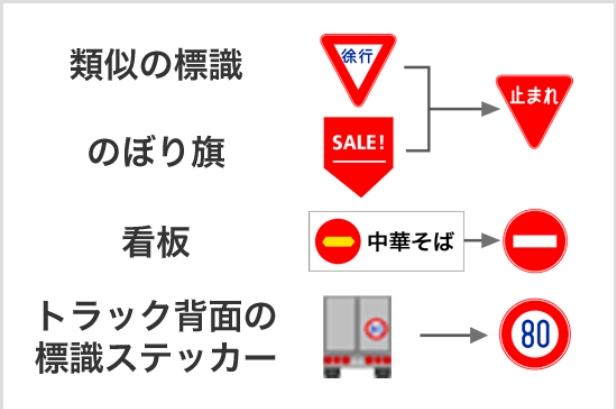


図 1: 運転支援システムの誤認識例
本田技研工業株式会社，アクティブセーフティー標識認識機能，https://www.honda.co.jp/tech/auto/safety/traffic-sign-recognition.html.[1] より引用

標識識別器が誤認識してしまう可能性がある．そのため，このような誤認識を引き起こさないための対策が必要である．

以上を踏まえ、本研究では、企業ロゴやシンボルが標識識別器によって誤認識されるリスクを評価し、その対策を示す。リスク評価においては、標識か否かを判断する二値分類器を作成し、入力画像が標識であると識別される確信度を定量化することで、実在する企業ロゴ等が標識と誤認識される程度を明らかにする。また、誤認識を引き起こした企業ロゴ等の特徴を意図的に再現する架空のロゴを「Adversarial Sign（敵対的標識）」として複数生成し、その誤認識率を評価する。対策としては、敵対的標識を訓練データに取り込んで再学習（Adversarial Training）を行うことで、未知のロゴや敵対的標識に対して誤認識率を低下させられることを示す.

2 関連研究

本章では，画像分類器に対して誤識別を意図的に誘発させる攻撃手法について述べる．また，関連する防御手法や関連研究と本研究との違いについても述べる.

2.1 Adversarial Example

Adversarial Example (AE)[2] とは，人間は知覚できないほどの微小な摂動を加え，機械学習モデルに対し誤識別を誘発させる目的で生成された入力のことを指す．摂動の作成方法によって様々な攻撃手法が提案されている．代表例として Fast Gradient Sign Method (FGSM)[3] が知られている．FGSM では式 (1) によって AE $\boldsymbol{x}_{\text{adv}}$ を生成する.

$$\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\theta, \boldsymbol{x}, y)) \quad (1)$$

*1 早稲田大学
Waseda University, Tokyo, Japan.

ただし，θ はモデルパラメータ，$\boldsymbol{x}$ は入力データ，y は正解ラベル，$J(\cdot)$ はモデルの損失関数である．Goodfellow らは式 (1) に基づいて生成した AE の攻撃成功率，すなわち画像識別モデルを訓練した後にテストデータで推論させたデータの誤識別率を調べた．

また，誤識別率を下げるための対策として Adversarial Training (AT) が提案されている．AT は，標識識別器の訓練段階で AE を生成し教師データに含める手法である．AT によって，AE が別のクラスに誤識別される割合が下がることが実験的に示されている．

本研究では，道路標識への誤認識を誘発するロゴやシンボルに着目しており，標識画像に対して微小な摂動を加えることはないため攻撃手法としては異なる．しかし，誤認識リスクの高いロゴ等を事前に訓練データに効果的に取り込むことで誤認識リスクを下げることができるという点で AT の手法を簡易的に用いている．

2.2 Robust Physical Perturbation

Robust Physical Perturbation (RP2)[4] とは道路標識にステッカーを物理的に添付することで，道路標識識別器の予測を誤識別させる攻撃を指す．Eykholt らは「一時停止」の標識に小さな白黒のステッカーなどを貼ることで画像認識モデルに対し，「時速 45 マイル制限」の標識に誤識別させることに成功した．実験はコンピュータ上ではなく，屋内に実際の標識を設置して行われたため現実的な脅威として問題を提起している．ステッカーなどが添付されている細工済みの標識は運転者からは認識に大きな問題を与えないが，画像認識モデルが誤認識するため，そのギャップが運転者に混乱を生じさせる可能性がある．すなわち，RP2 はモデルを欺くだけでなく，人間とモデルの認識に関する乖離が現実的な脅威へと発展し得ることを示唆している．

先行研究 [4] が実在する標識に対して加工を施しているのに対し，本研究では，ロゴ等を実在する標識であると誤識別する現象を調査・再現するためにロゴに焦点を当てて加工を施している点で異なる．

2.3 DARTS

Sitawarin らは自動運転車に対し，敵対的な標識によって誤認識させる複数の手法を Deceiving Autonomous caRs with Toxic Sings (DARTS)[5] と位置付けて提案した．DARTS は主に，標識に摂動を付加する手法，ロゴに摂動を付加する方法 (Logo Attacks)，実在する標識の外形をベースにオリジナルの摂動を付加する方法 (Custom Sign Attacks) を提案している．特に，Logo attacks と Custom Sign Attacks は Out-of-distribution (OOD) Attack に分類される．OOD データは，学習時には想定されていなかったドメインのデータで，意図していないモデルの出力を引き起こすことがある．標識識別器に対しては，これらの攻撃に用いられる加工前の元データは OOD データに該当する．出力空間に該当する適切なラベルが存在しないため，モデルの推論結果は複数の誤ったラベルに対して低い確信度が出力される．例えば，街中で見かける飲食チェーン店のロゴは標識識別器に対しては OOD データであり，該当する標識ラベルが存在しないにも関わらず，識別器は標識ラベルの中から尤もらしいものを無理に出力することになる．標識のブランクに関しても，個別の標識を特定するような印がブランクに印字されていれば正しい標識ラベルが付与されるが，ブランクでは意味をなさないため OOD データに該当する．Sitawarin らはさらにこのようなオリジナルのロゴやブランクにした標識に対して摂動を付与し，

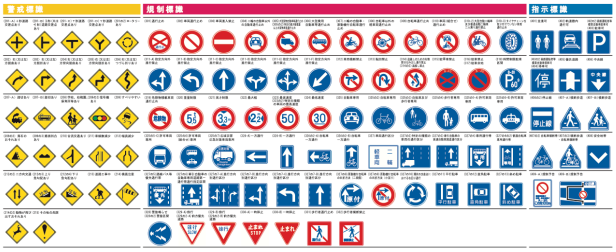


図 2: 道路標識一覧：国土交通省，`https://www.mlit.go.jp/road/sign/sign/douro/ichiran.pdf`.[7] より一部抜粋

「一時停止」のような特定の標識ラベルを意図的に付与させることに成功した．ここでの摂動とはピクセルの物理的なプリントを意味する．

本研究では，Sitawarin らの研究結果から，ロゴ等が標識と誤認識される現象の調査・再現のために 2 つの観点に着目する．1 つ目は，標識識別器の視点では街中のロゴ等は OOD に該当するという点である．OOD データが一度標識識別器に入力されると誤ったラベルが出力されるため，OOD データは未然に除外されるべきである．したがって，本研究では，標識とロゴを二値分類できる標識検出器を作成する．2 つ目は，標識検出器がロゴなどを誤検出するよう誘発する Adversarial Sign の生成にブランクの標識を利用するという点である．標識のブランクは青，赤，黄色のように単色で，かつ丸，三角，四角などの単純な図形から構成されており，その上に独自のロゴや図形を載せることで Adversarial Sign を生成する．したがって，先行研究 [5] が最適化により摂動を生成しブランクに付与するのに対し，本研究ではブランクに複数用意した単純な図形やロゴを重ねることで標識であると誤識別を誘発するシンボルを人工的に再現することを試みる．詳しい手法に関しては 3 章で述べる．

3 提案手法

本章では，ロゴ等が標識へと誤認識されるリスクの調査手法と，誤認識を防止するための対策手法について述べる．

3.1 誤認識リスクの評価

本研究では，標識識別器に入力される画像データに OOD データが含まれないように，識別器に入力する前段階に，標識かロゴ等かを二値判別する「標識検出器」を作成する．標識検出器の段階で OOD データであるロゴ等を取り除くことができれば，標識でないシンボルが標識であると誤認識される現象は防止される．しかし，1 章で述べたように現実の自動車運転支援システムでは飲食店ロゴを標識と誤認識する例が存在している．そこで標識とロゴ等を高精度に判別できる二値分類器を実装し，訓練した二値分類モデルに対し，未知のロゴ等を複数入力することでその誤り率や確信度を定量化し誤認識リスクを評価する．二値分類器の実装には VGG19 Batch Normalization 事前学習モデル [6] に対し転移学習を行う．訓練データの「標識ラベル」データとして，図 2 に示すような国土交通省が公開している道路標識一覧 [7] から主要標識 105 種類を各 1 枚ずつ収集した．ロゴラベルのデータ数と均衡を保つため，標識画像 105 枚から無作為に 30 枚を選んで回転・平行移動する加工を行なった画像も加えた．

訓練データの「ロゴラベル」データとして，インターネット上から企業のロゴやシンボル等を 135 枚，重複なく収集した．訓練実行後のモデルの精度を検証するための検証データとして，標識ラベルデータには標識見本一覧 [7] を拡大・回転加工した標識画像 105 枚，ロゴラベルデータには訓練データには用いていない企業等のロゴを 105 枚使用している．さらに，後に生成する Adversarial Sign のみからなるデータセットをテストデータと定義する．

3.2 誤認識を防ぐ対策手法

本節では，ロゴ等が標識であると誤認識される現象を防ぐ対策手法を提案する．前節の誤認識リスクの評価で標識であると誤認識されるようなロゴを複数事前に用意し，それらをロゴラベルのついた訓練データに取り込むことによって，誤認識を防ぐ Adversarial Training (AT) を実施する．AT のための準備として，標識であると誤認識されるようなロゴを複数用意する必要がある．前節の既存ロゴ等の評価の段階でも誤認識を誘発するようなロゴは複数存在すると考えられるが，街中に掲示されるようなロゴやシンボルは絶えず新しく創出されるものであり，今後も実在する標識に誤認識される例が生じると考えられる．したがって，既存のリスクの高いロゴ等を訓練データに含めるだけでは AT としては不十分であり架空の敵対的ロゴ (Adversarial Sign) を人工的に生成することによって AT の効果を検証する必要がある．Adversarial Sign の生成方法は次節で述べる．Adversarial Sign は標識検出器の誤検出を意図的に誘発するようなサンプルであるため，検出器の推論結果は標識ラベルになる可能性が高い．はじめに，作成した Adversarial Sign を標識検出器に推論させたときの誤り率を調査する．次に，訓練データのロゴデータセットの一部を人工的に生成した Adversarial Sign に置き換え，標識・ロゴ二値分類器を再訓練させ，3.1 節と同様の検証データで二値分類器自体の精度を評価する．さらに，テストデータセットを用いて，未知の Adversarial Sign や前節で誤識別したリスクの高い既存ロゴが AT によって正しく推論されるようになるかを検証する．

3.3 Adversarial Sign の作成

本節では，Adversarial Sign の生成方法について説明する．第 2 章で述べたように，誤認識を意図的に誘発させるために Adversarial Sign は標識のブランク (以下「ベース」と呼ぶ) に加工を行うことで生成する．図 3 に示すように，実在する標識から画像編集ソフト等を用いてベースを作成する．

次に，ブランクに重ねるためのシンプルな図形やアイコン (以下「パーツ」と呼ぶ) を作成または収集する．最後に作成・収集したベースとパーツを重ねて Adversarial Sign を作成する．このとき，標識に関して表 1 に示すようにベースとパーツには色の組があることに着目し，パーツの色を変換する前処理を行った上でベースに重畳する．Adversarial Sign 生成のプロセスを図 4 に示す．

なお，ここで生成した Adversarial Sign は識別器に対して誤 sikibetu の誘発を目的とした敵対的データに相当する．生成した Adversarial Sign は AT によって識別器の堅牢性が向上するかを検証するための，テストデータとして用いる．また，AT のために訓練データに混入させる Adverarial Sign と評価測定のためのテストデータとする Adversarial Sign ではパーツ部分に重複が生じないように分離し，リークを防いでいる．

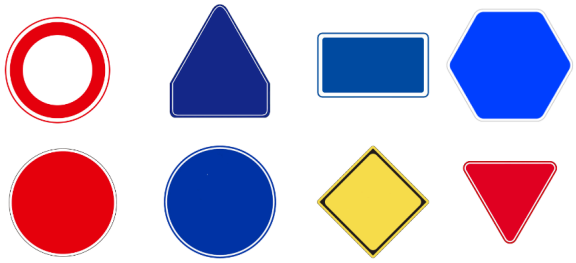


図 3: ベースの作成例

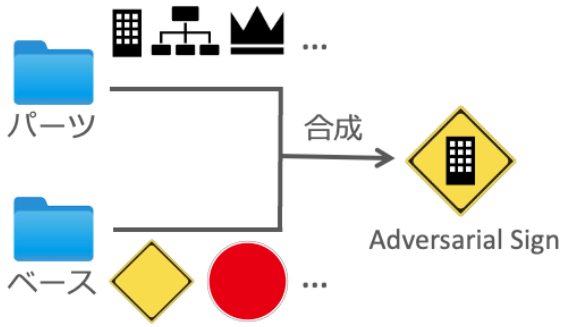


図 4: Adversarial Sign 生成プロセス

表 1: 標識におけるベースとパーツの色の組み合わせ

ベース	パーツ
赤	白
青	白
黄	黒
白	青

4 評価実験

本章では，3 章提案手法に基づくロゴ・標識の二値分類器による評価と誤識別を防止する対策に関して評価実験を行った．さらに，転移学習済みモデルが推論したラベルの判断根拠の指標として，Grad-CAM[8] による可視化を行った．

4.1 ロゴ・標識二値分類器の作成

標識画像とロゴ画像を用いて二値分類器を訓練した結果を表 2，図 5，図 6 に示す．表 2，図 5，図 6 に示すように作成した識別器はロゴ等と標識を 95% 以上の精度で識別することができる．一方で誤識別を引き起こしたサンプルも複数存在し，それらには以下の特徴があることが分かった．

- 円形，四角形などの大枠でロゴが構成されている
- 赤，青などの標識で主に使用されている色を使用している
- 単純な図形でシンボルマークが形成されている

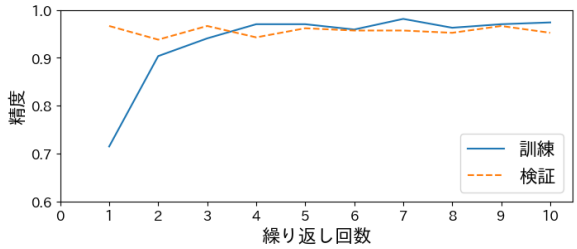


図 5: 二値分類器の通常訓練結果（精度）

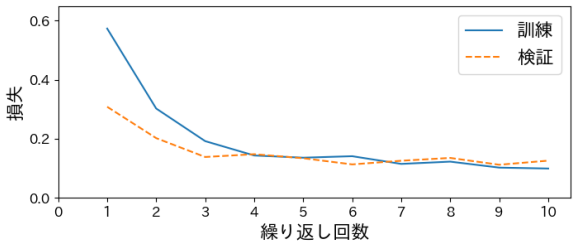


図 6: 二値分類器の通常訓練結果（損失）

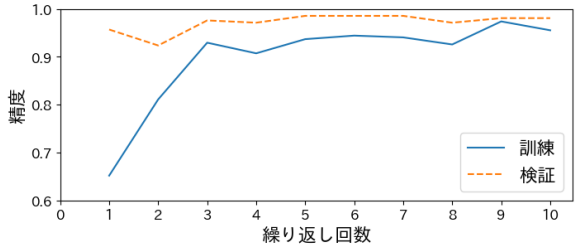


図 7: 二値分類器の AT 結果（精度）

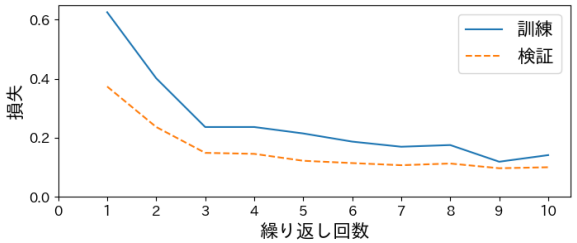


図 8: 二値分類器の AT 結果（損失）

4.2 Adversarial Sign の作成と評価

本節では Adversarial Sign を人工的に生成した結果と，前節で作成した二値分類器に Adversarial Sign を推論させたときの精度（攻撃成功率），AT の効果ついて述べる．パーツを 84 種類，ベースを 6 種類用意し，$84 \times 6 = 504$ 枚の Adversarial Sign を作成した．その一部を図 9 に示す．このうち 30 枚の Adverarial Sign を AT 用の訓練データに，残りの 474 枚をテストデータとして分離した．このとき，1 つのパーツに対して，6 枚のベースに埋め込まれることになるため，同じパーツをもつ Adversarial Sign が訓練データとテストデータの両方に存在してしまうとリークが生じて正しい判別ができなくなる．そこで，先にパーツを訓練用，テスト用に分離してから，ベースへの埋め込みを行いデータセットを作成している．

まず，テストデータ用として作成した 474 枚の Adversarial Sign を前節で作成した二値分類器に推論させた結果，精度は 0.03947 であった．次に AT 用に作成した 30 枚の Adversarial Sign を訓練データのロゴラベル画像からランダムに選んだ 30 枚と入れ替えたデータセットで AT を実施した．検証精度は約 2.857% 向上した．また，テストデータを用いて Adversarial Sign を AT 実施後の二値分類器に推論させた結果，精度は 0.94315 であり，大幅に向上した．さらに，表 3，表 4 に示す通り，二値分類器の False Positive すなわち，「実際にはロゴであるにもかかわらず標識であると誤識別した」例が減少していることが分かった．具体的には，提案手法で示したような標識のベースから任意のパーツを重畳して生成した Adversarial Sign の一部をロゴデータに学習させることで，誤識別リスクの高い既存のロゴの誤識別を是正できることが分かった．

4.3 転移学習済みモデルの推論根拠の可視化

本節では，二値分類器の推論根拠の指標として Grad-CAM[8][9] による可視化を行った．通常訓練時と，AT 実施時における推論根拠のヒートマップを作成し，比較した．テストデータにおける推論根拠の可視化の例を図 10，図 11 に示す．

図 9: 作成した Adversarial Sign の例

表 2: ロゴ・標識二値分類器の訓練結果

	AT 実施前	AT 実施後
訓練精度	0.97407	0.95238
検証精度	0.95556	0.98095
テスト精度	0.03947	0.94315

図 10，図 11 の例は，Adversarial Sign に対する推論根拠を可視化している．通常訓練時は標識だと誤識別したが，AT 実施後はロゴであると正しく分類された．通常訓練時のモデルと AT 実施後のモデルの推論根拠を比較すると，図 10 においては，推論に影響を与えた領域がパーツの中心部分からベースとパーツの境界へとシフトしている様子が確認できた．一方で，図 11 においては，通常訓練時と AT 実施後で予測ラベルは変化しているにもかかわらず，着目している推論根拠は大きく変化していない結果が得られた．

また、通常訓練時と AT 実施後の Grad-CAM による

表 3: 検証データの混同行列（AT 実施前）

真 \ 予測	logo	sign
logo	95	10
sign	0	105

表 4: 検証データの混同行列（AT 実施後）

真 \ 予測	logo	sign
logo	101	4
sign	0	105

可視化の差分を図 12 に示す。図 12 に示す通り、AT 実施後は画像のベースとパーツの境界付近を重要視していることが確認された。これは、図 11 のような推論根拠が大きく変化していない例でも同様の傾向が示された。実際のロゴ画像に関しても同様の可視化を行った結果，同様にしてパーツとベースの境界部分を重視している傾向が観察された．

5 考察

実験結果における表 2，表 3 から，ロゴ等と標識を区別できる識別器を構築しても一部のロゴは標識と誤識別される事例が存在していることが分かる．誤識別される例を個別に検証した結果，標識の構成パーツやベースに類似している形状，色が一部に含まれていることから，標識をベースとして任意の図形やアイコンを重ね合わせて生成した架空のロゴである Adversarial Sign は誤識別を効果的に引き起こすことが明らかになった．したがって，実社会においても，Adversarial Sign と同様，実在する標識に部分的に構成が一致しているロゴはリスクが高いと考えられる．

本研究で提起した「街中で掲示されるようなロゴ等が標識であると誤認識される現象」は標識識別器の標識に対する過剰適合が問題であると言える．したがって，Adversarial Sign のように，標識をベースとした架空のロゴを人工的に生成し，訓練データに取り込む AT が有効である．実際に，表 3，表 4 の比較から，通常の訓練時ではロゴを標識であると誤識別する例が減少することが実験的に示された．また，AT によって，未知の Adversarial Sign も 90% 以上検知できるようになり，識別器の堅牢性が大幅に向上していることが分かる．

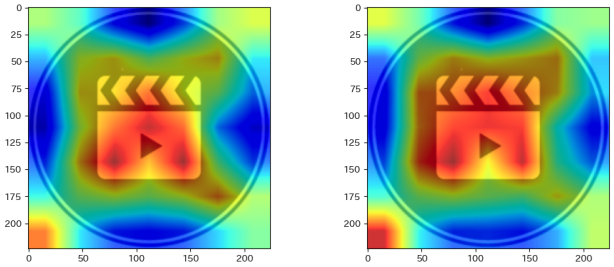
Adversarial Sign は標識のベースの上に標識に実際に利用されているパーツと同じ色の図形が配置されている。通常訓練時は、パーツの中心部分やベースとパーツの局所的な境界部分に着目して誤認識を引き起こした。AT による精度改善は、図 12 に示す結果から、ベースとパーツの大局的な境界部分に重点をおいて推論するように訓練されたためであると考えることができる．

今後も新しくロゴは創出され続けるため，既存のロゴだけにとどまらず，Adversarial Sign のように人工的に生成したロゴを AT に取り込むことで未知のロゴを正確に識別する対策は効果的であると考えられる．

6 結論

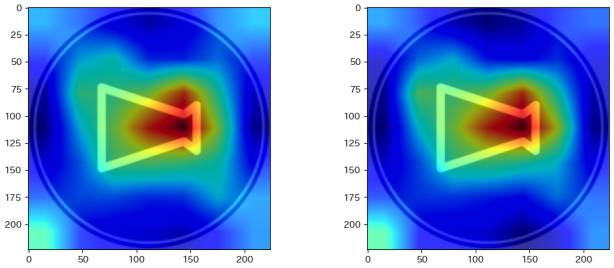
本研究では，「既存のロゴの一部が標識であると誤認識される現象」に着目し，誤識別を引き起こす構成の特徴の抽出と誤認識率の評価，対策を行った．誤認識を引き起こすロゴには標識と部分的に構成や色が一致する例が多く存在することが分かった．標識の構成は視認のしやすさから単純なベースとパーツであるため，ロゴに同様の構成が含まれていると通常の検出器では標識への過剰適合から誤認識を起こしやすいと考えられる．Adversarial Sign により誤認識リスクの高い架空のロゴを複数 AT に取り込むことで，モデルの堅牢性が向上することが示された．

今後の課題としては次の 2 点が考えられる．1 点目は，標識ラベルのデータが現実の環境に対応していない点である．本研究の実験で使用した標識ラベルのデータはすべて国交省が公開している見本を用いている．したがって，現実の環境で入力される標識データのように，光の反射，周囲の障害物，標識の経年劣化などによる変色，変形が考慮されていない．理想的には，標識見本画像から変形，変色していても正しく検知する条件の幅広さを維持したまま，ロゴやのぼりのような掲示物は正しく除外できる状態が求められる．2 点目は，使用するデータの拡充が必要な点である．本研究では，自動車運転支援システムに搭載される標識検出器を想定して，運転中に



(a) 通常訓練時　(b) AT 実施時

図 10: 推論根拠の可視化（判断根拠の変動が大きい例）



(a) 通常訓練時　(b) AT 実施時

図 11: 推論根拠の可視化（判断根拠の変動が小さい例）

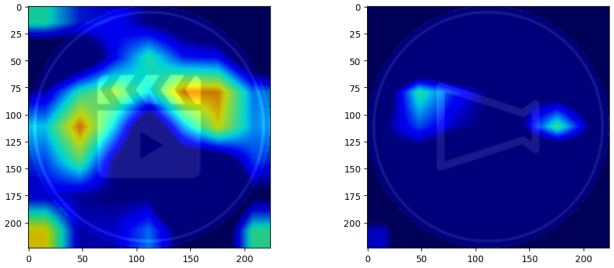


図 12: 推論根拠の変化の可視化

視界に入りやすい標識とロゴに焦点を当てて画像識別を行った．実際には，ロゴ以外にも様々なタイプの掲示物が入力の候補になり得る．モデルの構築には，実際に運用されている自動車などから膨大な画像データを収集し，より実践的な環境で実験する必要がある．

謝辞

本研究の一部は，日本学術振興会における科学研究費補助金基盤研究（C）（課題番号 23K11111）による支援を受けている．ここに記し謝意を表す．

参考文献

[1] 本田技研工業株式会社．アクティブセーフティー標識認識機能．https://www.honda.co.jp/tech/auto/safety/traffic-sign-recognition.html.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. land Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[3] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 2015.

[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[5] C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal. Darts: Deceiving autonomous cars with toxic signs, 2018.

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

[7] 国土交通省. 道路標識一覧. https://www.mlit.go.jp/road/sign/sign/douro/ichiran.pdf.

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017.

[9] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.