Dynamic Creative Knowledge, Vol.3 (2025)

騙される AI および AI を騙す技術の解説

中田　尚[† a)]

Introduction of Misled AI and Adversarial Techniques

Takashi NAKADA[† a)]

あらまし　人工知能の発展に伴い，人工知能の脆弱性が新たな課題として浮上している．人工知能の一つの致命的な欠点として，人間には認識できない程度のノイズであっても，その出力に大きな影響を与える敵対的攻撃の存在である．本稿では画像認識を題材として敵対的攻撃とその防御手法について簡単な例を用いて説明する．具体的には，画像認識モデルに対する代表的攻撃手法である FGSM 攻撃の効果を分析し，人工知能モデルの堅牢性を高める防御策の一つとして BwoBL を紹介する．特に大規模データセットである ImageNet を用いた実験を通じて，その有効性を示す．

キーワード　人工知能，敵対的攻撃，防御手法，統計的機械学習

Abstract　Artificial intelligence (AI) systems are susceptible to adversarial attacks, where minor input changes cause significant errors. This paper reviews these attacks in image classification and evaluates defense methods like adversarial training and Bayesian Neural Networks (BNNs). We introduce “Bayes without Bayesian Learning (BwoBL),” a novel approach that improves model robustness with low computational cost. Experiments on the ImageNet dataset demonstrate its effectiveness against advanced attacks. This work contributes to building more secure and reliable AI systems.

Keywords　Artificial Intelligence, Adversarial Attack, Robustness, Statistical Machine Learning

1. まえがき

近年，人工知能技術は急速な進歩を遂げ，画像認識や自然言語処理などの分野を中心に，これまでにない高い性能を実現している．人工知能は，医療，交通，セキュリティなど多岐にわたる分野で人間の能力を補完し，時には凌駕する役割を果たしている．しかし, これらの技術の進展に伴い，人工知能モデルの脆弱性が新たな課題として浮上している．

特に，敵対的攻撃（Adversarial Attack）は，入力データに微小な改変を加えるだけで人工知能モデルの出力を意図的に誤らせる技術である．人間の目にはほとんど認識できないこの種の改変は，人工知能システムにとって致命的な影響を与えうる．例えば，画像認識システムにおいて，攻撃者がノイズを付加することで，人工知能はネコをイヌと認識することがある．

敵対的攻撃に対する研究は，防御手法の開発と合わせて重要性を増している．本稿では，画像認識を題材に，敵対的攻撃の仕組みとそれに対する防御策について紹介する．特に，敵対的攻撃の中でも広く知られている FGSM（Fast Gradient Sign Method）を取り上げ，その影響を検証するとともに，新しい防御アプローチの紹介を行う．

本論文の目的は，敵対的攻撃の脅威を明らかにし，それに対処するための効率的かつ実用的な防御手法を示すことである．本稿を通じて，人工知能モデルの堅牢性向上に向けた具体的なアプロー

† 大阪国際工科専門職大学工科学部情報工学科，大阪市
Department of Information Technology, Faculty of Technology, International Professional University of Technology in Osaka, 3-3-1 Umeda, Kita-ku, Osaka 530-0001 Japan

a) E-mail: nakada.takashi@iput.ac.jp

チを提示する．

2. 画像分類の基本

画像分類とは，画像データを入力として，あらかじめ定められた複数の出力ラベルの中から適切なものを選択する処理である．一般的には画像に写っている物体の種類をラベルとすることが多い．入力画像が 100 × 100 画素のモノクロデータであれば，その次元数は 10,000 次元となる．出力はそのラベルを識別する一意の ID，すなわち単一の数値であるが，本稿ではその 1 つ手前の状態を対象とする．具体的には，最終段の 1 段手前では，100 分類であれば 100 個の出力値が得られる．この値を argmax 関数により，最大値のインデックスに変換することで，出力ラベルを得る．以上のことから画像分類は多次元（この例では 10,000 次元）を低次元（この例では 100 次元）に次元圧縮する処理とみなすことができる．

画像データが画像分類モデルに入力されると，分類対象となる各クラスに対してスコアが計算される．例えば, 分類するクラスが 100 種類の場合，モデルの出力は「長さ 100 の数値のリスト」となる. このリストの中で最も高いスコアを持つ値が，画像が属すると判断されたクラスに対応する．

この一連の処理は, 画像を入力として受け取り，クラススコアを出力する関数の集合として表現できる．具体的には，「高さ，幅，チャネル数で構成された画像データを入力として受け取り，それをクラスのスコアに変換する」となる．100 種類のクラスがあった場合には，この処理を 100 回繰り返すことに相当するが，実際の機械学習ではお互いの計算結果に依存関係があるため，100 倍の処理量にはならない．

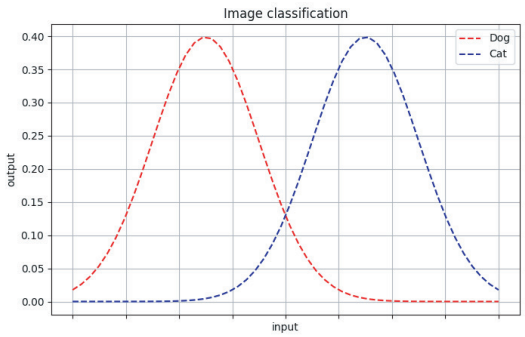

図 1　画像分類の理想的な入出力例
Figure 1　An ideal example input/output of an image classification.

たとえば，224 ピクセルの高さと幅を持つ RGB 画像を入力し，100 種類のクラスに分類する場合，入力は「224 × 224 × 3 次元」, 出力は「100 次元」になる．この過程を通じて，画像が特定のクラスに分類される．

図 1 に入力を 1 次元，出力をイヌとネコの 2 次元に簡略化した場合の理想的な結果を示す．赤色の折れ線が高い所がイヌ，青色の折れ線が高い所がネコの可能性が高い画像である．重なっている部分は「イヌのようなネコ」または「ネコのようなイヌ」に相当する．以上のことから攻撃者からの視点では，青と赤のグラフの大小を逆転させることが目的となる．

3. 敵対的攻撃

システムに対する攻撃は，大きく分けてホワイトボックス攻撃とブラックボックス攻撃に分けられるが，本稿ではホワイトボックス攻撃を対象とする．

ホワイトボックス攻撃とは，攻撃者が対象となるシステムやアルゴリズムに関する完全な内部情報を保有している状況下で実施される攻撃手法である．この攻撃モデルにおいて，攻撃者はシステムのソースコード，アルゴリズム設計，モデルパラメータ，トレーニングデータ，内部処理の挙動といった詳細情報を十分に理解し，それらを最大限に活用して攻撃を設計･実行する. ホワイトボックス攻撃の特徴的な要素は，システムに関する情報が攻撃者に対して完全に透明である点にある．この透明性は，例えば設計上の不備，リバースエンジニアリングによる解析，内部からの情報漏洩などによって生じる．こうした情報を前提に攻撃者は，対象の設計上の弱点や実装ミスを正確に突き，そのシステムの挙動を意図的に操作することが可能となる．このように，ホワイトボックス攻撃の方が攻撃者に有利であるため，ホワイトボックス攻撃を対象として議論しても一般性を失わない．

また，ホワイトボックス攻撃は攻撃者にとって

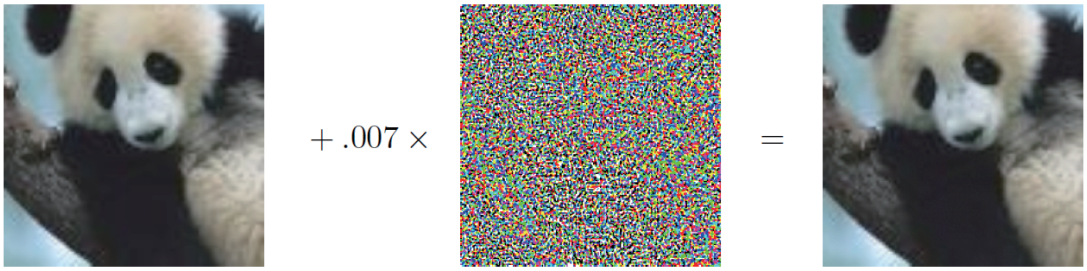

図 2　敵対的攻撃の例 [1]
Figure 2　An example of an adversarial attack [1].

試行錯誤の必要性を大幅に低減する．ブラックボックス攻撃とは異なり，内部情報を利用することで，攻撃の効率が飛躍的に向上し，より短期間で攻撃を成功させることができる．加えて，設計情報に基づく綿密な計画を立てることが可能であり，攻撃の成功率が高まる傾向にある．

ホワイトボックス攻撃への対策は, その特性上, ブラックボックス攻撃への対策としても有効に機能する．この理由は，ホワイトボックス攻撃においては攻撃者がシステムの内部情報を完全に把握している一方で，ブラックボックス攻撃ではその情報が限定的または欠如しているためである．したがって，ホワイトボックス攻撃を防ぐために講じられる高度な対策は，より制約された情報環境下で行われるブラックボックス攻撃に対しても十分な防御を提供する．

3.1　FGSM

画像認識における FGSM（Fast Gradient Sign Method）攻撃 [1] は，ニューラルネットワークに基づく画像認識モデルに対する敵対的攻撃（Adversarial Attack）の一つである．この攻撃手法は，モデルの誤分類を引き起こすために入力画像に対して小さなノイズを加えることを目的とする．FGSM は，その効率性とシンプルさから広く研究されている．

図 2 はこの攻撃に関する多くの論文で引用されたため，広く知られる図である．AI が正しくパンダ（Panda）と認識できる画像（左）に計算されたノイズ（中央）を加えることにより，認識結果をテナガザル（Gibbon）に変えることに成功している．

FGSM 攻撃では, モデルの損失関数（誤差関数）の勾配情報を用いて敵対的ノイズを生成する．このノイズは，入力画像に対してわずかな変化であるが，モデルの出力結果を大きく変化させるよう設計される．具体的には，ニューラルネットワークの損失関数 $J(x, y)$ を最大化する方向にノイズを付加する．

FGSM による敵対的サンプル x_{adv} は次式で定義される．図 2 では $\varepsilon = 0.007$ としている．

$$x_{\text{adv}} = x + \varepsilon \cdot \text{sign}(\nabla J(x, y)) \tag{1}$$

ここで，各変数は以下を表す：

x：元の入力
y：正しい出力ラベル
x_{adv}：敵対的サンプル（ノイズが加えられた画像）
ε：ノイズの大きさを調整するパラメータ
$\nabla J(x, y)$：入力画像 x, 出力ラベル y に関する損失関数の勾配
$\text{sign}(\cdot)$：符号関数

損失関数の本来の目的は，損失すなわち正解との誤差を最小化することである．そのような関数の勾配が正の方向に入力を変化させると，損失は増大する．この計算により，画像の各ピクセルに対して損失が増加する方向にノイズが加えられ，損失関数が最大化される．その結果，正解ラベルに対応する出力値は小さくなり，結果として他のラベルの出力値が大きくなり，正解ラベルと逆転することが期待される．

図 2 の左が元の図であり，中央がノイズで，右

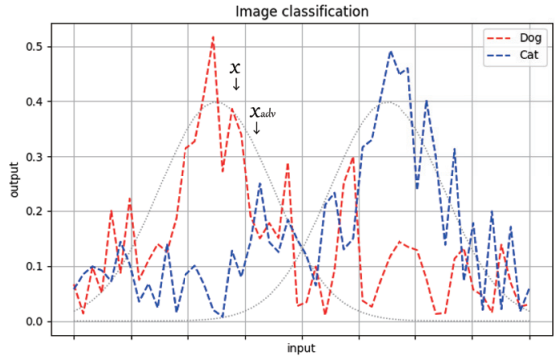

図 3　画像分類の現実的な入出力例
Figure 3　A realistic example of an image classification.

が元の図にノイズを加えたものである．ノイズを加える際には ε を乗算する．ε が小さいので左右のパンダの違いを目視できないであろう．ましてや，右の図がテナガザルに見えることはあり得ない，

実際の入出力で起こっている現象を模式的に表示したのが図 3 である．これは図 1 と同じ入出力関係をより現実的に示しており，図 1 ではなめらかな正規分布であったが，ここでは実際の学習結果を反映してなめらかにはなっていない．このような現実的な画像分類モデルに対して，図中の x で示した入力が与えられた時に FGSM による攻撃を行うと，x における Dog の勾配が負であることを利用し，右に移動するようなノイズを加えることにより，入力を x_{adv} に変化させる．その結果，Dog の出力値は大幅に小さくなり，青線の Cat の出力値と逆転しており，攻撃に成功して誤認識が発生することとなる．

FGSM 攻撃はシンプルでありながら効果的であり，後続の多くの発展的攻撃手法（例：PGD[2], C&W 攻撃 [3]）の基礎となる技術である．この攻撃手法は，AI モデルが現実世界で遭遇する可能性のある脅威に対する脆弱性を明らかにする一方, 防御策を開発する上で重要な知見を提供する．

4. 防御手法

4.1　敵対的攻撃への対応

敵対的攻撃の特徴は人間には差分が識別できないにもかかわらず，AI には全く異なる物体として認識される点にあり，防御手法としては，ノイズが加わった場合であっても答えが変化しないことが求められる．

4.2　既存防御手法

既存手法としては，事前の学習画像にランダムのノイズを加える手法（Adversarial Training）と，モデル自体を確率分布で表現する手法（BNN）が挙げられる．

4.2.1　Adversarial Training

Adversarial Training は，事前の学習画像にランダムのノイズを加えることにより，どのようなノイズが加わったとしても，出来る限り正しく認識できるように学習する方法である．これは単純にノイズパターンの数だけ学習コストの増大を招く．学習効率改善のためにランダムなノイズを加えるのではなく，FGSM 攻撃を用いて生成された画像を学習データに加えることも有効である．ただし，モデルが修正されるとそれに追従して FGSM 攻撃のノイズも変化する点に注意が必要である．いずれにせよ，ノイズパターンは膨大であり，あらゆる場合について事前に網羅的に学習しておくということは非現実的であるため，学習コストと防御性能のトレードオフがある．

また，ホワイトボックス攻撃においては攻撃が成功したかどうかは攻撃者にとっても明らかであるため，「誤認識が発生するまで」攻撃処理を続けることにより，必ず攻撃が成功するようにできる．ただし，この場合は加えるノイズが過大になる危険性があり，人間に攻撃の有無を気付かれてしまう可能性が高まるため，一定の抑止効果は得られる．

4.2.2　BNN

BNN（BayesNeuralNetwork）[4] ～ [6] は，通常の深層ニューラルネットワーク（Deep Neural Network；DNN）とは異なり，モデルの内部パラメータである重みを確率分布として扱う．このアプローチにより，モデルが予測にどの程度自信を持っているかを表現できるようになる．

BNN は敵対的攻撃に対する防御力を向上させる可能性がある．3. 章で述べた通り敵対的攻撃とは，入力データに人間にはほとんど認識できない

微小な改変を加えることで，モデルの予測結果を意図的に誤らせる．DNNでは重みが固定されているため，このような攻撃に対して脆弱だが，BNNでは重みが確率的に定義されるため，攻撃者がモデルの振る舞いを予測しにくくなり，防御力が向上する．

さらに，BNNは多数の異なるモデルが重なり合わさったような性質を持っている．これにより，モデルの予測性能が向上する．従来のアンサンブル学習では互いに異なる複数のモデルを実際に構築する必要があったが，BNNでは単一のモデル内でアンサンブル効果を実現できるため，計算資源の効率的な利用が可能である．

しかし，BNNにはいくつかの課題が存在する．まず，モデルの重みを確率分布として扱うことで，通常のDNNと比べて学習に必要な計算量が大幅に増加する．特に大規模なデータセットや複雑なネットワーク構造の学習では，実用的ではない場合が多くなる．また，学習の際には多くのパラメータを最適化する必要があるため，モデルのトレーニングを正しく完了すること自体が難しくなることもある．

具体的には28 × 28ピクセルのMNISTや32 × 32ピクセルのCIFAR-10といった小さなデータセットによる評価のみがされており，それよりも大きな画像サイズに対応出来るかどうかは未検証あることが，その限界を暗に示している．

4.3 提案方針

原理に立ち返ってみると，敵対的攻撃の特徴は人間には差分が識別できないにもかかわらず，AIには全く異なる物体として認識される点にある．

人間に違和感が無いということは元画像との差分が少ないということである．式におけるεが小さいことに対応する．

また，敵対的攻撃が成功するということは，元の画像認識の精度がある程度高い必要がある．もし，認識精度が低ければ誤認識が敵対的攻撃によるものなのかどうかの判別が難しくなるため，当然の仮定である．

以上のことから，解空間の多くの部分で正しく認識が可能であり，ある入力を1つに固定したという前提において，その近傍で間違った認識結果を返す点（画像）を効率的に発見することができることが，敵対的攻撃の狙いである．

本稿で紹介する防御手法は，入力画像が一つ与えられた時に，その認識処理を統計的に処理する．具体的には，入力画像や認識プロセスに故意にノイズを加えることにより，元の入力画像に敵対的攻撃が加えられたとしても，その画像にさらにノイズを加えることにより，誤判定を防ぐ．

事前に挙動の予測が不可能な乱数生成器を用いることにより，たとえ防御処理によりノイズが加えられることが事前にわかっていたとしても，具体的なノイズ値は実行時にしか判明しないため，未知のノイズパターンに対して的確に誤認識を発生させることは困難である．

4.4 BwoBL

BNNの欠点は確率変数に置き換えた重みを学習するコストにある．

そこで，文献[7], [8]では，事前学習済みのDNNを活用し，それを基にBNNを構築する方法を提案している．この手法の中心にあるのは，DNNの既存の構造を変更することなく，特定の層を「ベイズ層」と呼ばれるものに置き換えることで，不確実性を取り入れ，敵対的攻撃に対する耐性を高めることである．

まず，BNNを構築するために，事前学習済みDNNの畳み込み層や全結合層をの重み変数を確率変数に置き換える．この変更によって，モデルの重みは固定値ではなく，確率分布として扱われるようになる．このように変換された層を「ベイズ層」と総称する．他の層，例えば活性化関数やプーリング層などはそのまま維持され，元のネットワーク構造の基本的な特性が保たれる．

ベイズ層では，重みが確率分布に基づいて変動するようになり，敵対的攻撃に対する防御能力を高める効果がある．この不確実性は，勾配に基づく攻撃を難しくするため，敵対的な入力に対してモデルが過信するリスクを軽減する．

さらに，従来のベイズ学習が直面していた学習

コストの課題を解決するために, ここでは「Bayes without Bayesian Learning（BwoBL）」と呼ばれるアルゴリズムを利用する．このアルゴリズムは, 事前学習済み DNN の重みを活用して確率変数を初期化する．具体的には, 元の DNN の重みを「平均」とし，それに基づく分散を外部パラメータとして調整することで，不確実性を取り入れた重みを自動生成する．これにより，追加トレーニングを必要とせず，既存のネットワークを効率的に強化する．

また，BNN の推論では，複数回のモデル実行を通じてアンサンブルを作成し，最も頻繁に得られる予測結果を最終的な出力として採用する．この方法により，単一のモデル実行に比べて予測精度が向上する．この手法は計算リソースを追加で消費するが，その増加幅は定数倍のオーダであり追加メモリも必要としないため，データ並列による高速化も可能であるため，近年のハードウェアの高性能化により現実的に適用可能である．

このようにして構築された BNN は，元の DNN の高い性能を維持しながら，追加学習なしに敵対的攻撃に対する頑健性を飛躍的に向上させる．また，このアプローチは，事前学習済みモデルをそのまま活用するため，現時点で登場していないモデルを含めた幅広いモデル構造に容易に適用できる．この柔軟性が, 提案手法の大きな強みとなる．

実際の入出力で起こっている現象を模式的に表示したのが図 4 である．これは図 3 と全く同じ入出力関係を示している．図中の x_{adv} で示した FGSM による攻撃を受けた入力に対して，x_{adv} にさらに確率的にノイズを加えて，$x1, \ldots, x4$ を自動的に生成する．その結果，x_{adv} の出力では，青線の Cat の出力値の方が大きいが，それ以外の入力では赤線の Dog の出力値の方が大きいため, これらの多数決を取ることにより防御に成功して誤認識を防止する．

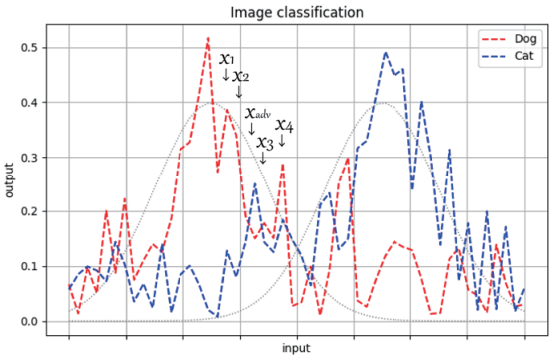

図 4　BwoBL による防御例
Figure 4　An example of protection by BwoBL.

5. 評　　価

提案手法の有効性を検証するための実験について紹介する．使用するデータセットは大規模な画像データセットとして広く知られる ImageNet である．このような大規模データセットで敵対的訓練を実施することは計算コストが高いため，提案手法がその課題をどのように克服するかを評価することが本実験の目的である．

実験に使用するベースラインモデルとして, ResNet や EfficientNet など，DNN アーキテクチャを用いている．これらのモデルは，事前に ImageNet で訓練されたものであり，提案手法ではこれらの事前学習済みモデルを基盤に BNN を構築した．

攻撃手法としては，FGSM を拡張した Projected Gradient Descent（PGD）[2] と Carlini&Wagner attack（C&W 攻撃）[3] という二つの強力な敵対的攻撃が選択されている．PGD 攻撃では，異なる摂動サイズや反復回数を設定して攻撃の強度を調整し，モデルの耐性を評価する．一方，C&W 攻撃は，PGD よりも収束が遅いものの，非常に強力な攻撃として知られており，さらにモデルの耐久性を測定するために使用される．実験では, BNN の構築において重要な分散を調整する外部パラメータも評価されている．このパラメータの選定は，モデルが敵対的攻撃に対してどの程度耐性を持つかに直接影響を与える．

その結果，PGD 攻撃および C&W 攻撃が行われている状況においても，それぞれ最大で 92.14%，94.20% の精度を維持 [8] し，高い耐性を確認した．

この実験により，提案手法の性能を包括的に評価され，ImageNet という大規模データセットに対しても敵対的攻撃への高い耐性が示されている．

6. ま と め

本稿では，人工知能モデルに対する敵対的攻撃の脅威と，それに対する防御手法の可能性について検討を行った．特に，従来の防御手法である Adversarial Training や Bayes Neural Network（BNN）の特徴と課題を整理し，それを踏まえた新しい防御アプローチ Bayes without Bayesian Learning（BwoBL）を紹介した．

BwoBL は，事前学習済みモデルを活用し，確率的要素を取り入れることで防御力を向上させるものである．その結果，追加学習コストを一切不要にしつつ，敵対的攻撃に対する高い耐性を実現することができた．さらに，ImageNet のような大規模データセットにおいても，強力な攻撃手法（PGD および C&W 攻撃）に対する有効性を実験的に示した．

この成果は，人工知能モデルの実用性と安全性を向上させる可能性を示唆している．特に，紹介した手法は適用範囲が広く，既存の様々なモデルに適用可能である点は，実世界での応用において大きな利点であり，今後の敵対的攻撃技術の進化に伴う新たな脅威に対して有望な防御策となり得る．

この成果がより頑健な人工知能モデルの構築に寄与することを期待する．

文　　献

[1] I.J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 3rd International Conference on Learning Representations - ICLR 2015, arXiv:1412.6572, pp.1–11, 2015.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” 6th International Conference on Learning Representations - ICLR 2018, arXiv:1706.06083, pp.1–28, 2018.

[3] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” 2017 IEEE Symposium on Security and Privacy (SP)IEEE, pp.39–57 2017.

[4] C.M. Bishop, Bayesian methods for neural networks, Aston University, 1995.

[5] R.M. Neal, Bayesian learning for neural networks, vol.118, Springer Science & Business Media, 2012.

[6] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural networks,” 32nd International Conference on Machine Learning, JMLR: W&CP, arXiv:1505.05424, pp.1–10, 2015.

[7] T.T.T. Khong, T. Nakada, and Y. Nakashima, “Bayes without bayesian learning for resisting adversarial attacks,” 2020 Eighth International Symposium on Computing and Networking (CANDAR), pp.221–227, 2020.

[8] T.T.T. Khong, T. Nakada, and Y. Nakashima, “Flexible bayesian inference by weight transfer for robust deep neural networks,” IEICE Transactions on Information and Systems, vol.E104.D, no.11, pp.1981–1991, 2021.

中田　尚

豊橋技術科学大学大学院工学研究科博士後期課程電子・情報工学専攻修了．博士（工学）．奈良先端科学技術大学院大学情報科学研究科助教，東京大学大学院情報理工学系研究科助教，奈良先端科学技術大学院大学准教授を歴任．現在，大阪国際工科専門職大学工科学部情報工学科准教授．主に高速シミュレーション技術，ノーマリーオフコンピューティング，深層機械学習の最適化の研究に従事する．情報処理学会 CS 領域奨励賞，IEEE SSCS Japan Chapter Academic Research Award など受賞．