とらえるべき特徴に着目した深層学習モデルの構築

浅野 正義 川越 佑太　吉岡 信和

自動運転を始めとした, ミッションクリティカルな分野への深層学習の活用が期待されている. しかし, ミッションクリティカルなシステムの開発に現在の深層学習を用いたシステム開発を適用するには課題がある. 深層学習を用いたモデルの訓練手順では, 作成したモデルの評価として, 訓練済みモデルのアウトプットを中心にモデルの完成を評価している. アウトプットでモデルを評価した場合にはインプットのどのような特徴に注目してモデルがアウトプットを出したかを判断できない. このような手順により作成したモデルをミッションクリティカルなシステムに用いると, システム稼働時のリスクに気づけなくなる. これは, モデルが適切でない特徴に注目していた場合, そのことに気づく手段が無いためである. 本稿ではこの問題を解決するため, モデルがとらえるべき特徴に着目した, 深層学習の学習済みモデル作成手法を提案する. 本手法は, ゴール指向分析により, モデルがとらえるべき特徴を分析する. さらに,XAI によりモデルがとらえた特徴を可視化する. これらを比較することにより, モデルがとらえるべき特徴をとらえていることが確認できる. 実験を行い, 本手法を用いて道路標識の識別モデルを作成した. モデルがとらえるべき特徴を分析し, モデルの作成過程でモデルがとらえた特徴と比較した. これにより, 本手法を用いることで訓練済みモデルがとらえるべき特徴に着目できていることの判断が可能であることが確認された.

Deep learning is expected to be utilized in mission-critical fields such as autonomous driving. However, the current system development using deep learning cannot be applied to the development of mission-critical systems. In the model training procedure using deep learning, the model is evaluated by focusing on the output of the trained model as the evaluation of the created model. When the model is evaluated by the output, the basis of the model output cannot be confirmed. If a model created by such a procedure is used in a mission-critical system, the risks during system operation will not be noticed. This is because we have no way to notice that the model focuses on inappropriate features. In order to solve this problem, this paper proposes a method for creating a trained model in deep learning that focuses on the essential features of the model. This method analyzes essential features by goal-oriented analysis. Then the features captured by the model are visualized by Explainable AI(XAI). By comparing these features, it can be confirmed that the model focuses on the essential features. In this method, we analysis the essential features of the model and confirm that outputs is based on it. We conducted an experiment and created an identification model of road signs using our method. During the course of learning, we compared the essential features we analyzed with the features the model focused on. Experiments have shown that it is possible to judge that the trained model focuses on essential features by using this method.

1　はじめに

深層学習が注目され, 多くのシステムに深層学習の技術が導入されるようになってきた. 特に近年では, 自動運転を始めとした, ミッションクリティカルな分野へも深層学習の活用が期待されている. しかし, ミッションクリティカルなシステムの開発に現在の訓練手順を適用するには課題がある. 現在の深層学習を利用した訓練手順では, 作成したモデルの評価として, 訓練済みモデルのアウトプットを中心にモデルの完成を

Creating a construction method of deep learning model that focuses on the essential features.

Masayoshi Asano, 株式会社デンソー, DENSO Corporation.

Yuta Kawagoe, NEC ソリューションイノベータ株式会社, NEC Solution Innovators.

Nobukazu Yoshioka, 国立情報学研究所, National Institute of Informatics.

評価している. しかし, ミッションクリティカルなシステムの場合, モデルのアウトプットのみでモデルを評価するのは危険である. これは, アウトプットの確認のみでは訓練済みモデルがとらえるべき特徴を学習していることが確認できないためである. 例えば, オオカミを識別するモデルを作成する際に, 識別精度を要件とした場合には, オオカミの背景にある, 雪を重要な特徴として, モデルが生成される場合がある [3]. この雪という特徴はオオカミとハスキーを識別するモデルがとらえるべき特徴とは言えない. そのため, このモデルは背景に雪が映っていれば, 他の動物もオオカミとして識別結果を出力する. このようなことがミッションクリティカルなシステムで起こると, 誤識別による重大な事故を引き起こす. そのため, 訓練済みモデルが意図しない特徴ではなく, とらえるべき特徴に注目していることを確認しながら学習を進める必要がある. とらえるべき特徴は, 作成する訓練済みモデルが, 適切なアウトプットを出すために必要な特徴である. そのため, 訓練済みモデルがとらえるべき特徴を根拠に出力していることが確認できれば, 訓練済みモデルが適切な学習を行えていることを判断できる.

以下本稿では,2 章ではこれまで深層学習やシステム開発で用いられてきた, 手法を説明する.3 章では, モデルがとらえるべき特徴を明確化し, 深層学習によって生成されたモデルが, 妥当な特徴を根拠に識別を行っているかを確認する手法を説明する.4 章では, 手法の効果を確認するために, 行った実験について説明する.5 章では, 実験結果についての評価について説明する.6 章では本稿のまとめと今後の展望について説明する.

2 関連研究

2.1 訓練済みモデルの可視化

近年, 訓練済みモデルの解釈性が低いことが問題視されており, これを可能にする SHAP [2] などの Explainable AI(XAI) の技術が注目されている.XAI は, 訓練済みモデルに対する元のテストデータ上の一部をランダムに変化・欠損させ, 編集されたテストデータが訓練済みモデルのアウトプットに与える影響を分析し, テストデータ上に含まれるどの要素が答えを導出する際に重要であったかを判断させ, 機械訓練済みモデルが識別している特徴を可視化する. これにより, 訓練済みモデルがどのような特徴に注目して, アウトプットを導出しているかを分析することが可能となる手法である.

しかし,XAI を使用するだけではモデルが満たすべき特徴に注目しているかを確認することはできない. 前述したとおり,XAI は訓練済みモデルが注目している特徴を可視化することは可能であるが,XAI によって明らかにした注目している特徴がシステムの要求を満たすものであるかについては XAI だけでは評価することはできないためである. モデルがとらえるべき特徴をとらえているかを判断するには, 事前にシステムの要求と特徴が紐づくように分析しておかなければならない. そして, システムの要求を満たすためにモデルがとらえるべき特徴を抽出する必要がある. そのうえで,XAI によって可視化されたモデルが注目している特徴とモデルがとらえるべき特徴の比較により, 訓練済みモデルがとらえるべき特徴に注目しているかを判断する必要がある.

2.2 システムの要求分析

システム開発において, 要求を抽出する手法の一つとしてはゴール指向分析が広く用いられている [1]. この手法は達成すべき最上位の目標をトップゴールとして設定し, トップゴールを達成するために必要な要素をさらに下位目標（サブゴール）へと分割していく. 分解したサブゴールをさらにサブゴールへ分解し, 実現可能な要素まで継続していくことで, 抽象的なトップゴールを達成可能である具体的な要求へ整理していく手法である. また, このサブゴールへの分割する方法として,AND 分割と OR 分割の二種類がある.AND 分割は上位ゴールから分割した全てのサブゴールを満たしたときに上位ゴールを達成できる際に用いられる. 対して OR 分割はいずれか一つのサブゴールが満足されると, 上位ゴールを達成できる場合に用いられる. この手法では, 最下層のサブゴールを満たすことで, システムの要求が満たされることとなる. すなわち, サブゴールを分割することでシステムに対する要求を抽出することを可能とする.

しかし, ゴール指向分析をそのまま深層学習を用いたシステム開発に適用することは難しい. これは, 深層学習の学習過程がブラックボックスになっているためである. 通常ゴール指向分析を用いた場合, サブゴールを充足するよう, 開発を行う. しかし, 深層学習の場合モデルがどのように学習をすすめるかを, 制御することはできない. このため, 訓練済みモデルに対する要求を分析したとしても, 訓練済みモデルが要求分析に従った学習するとは限らない. そのため, 学習を繰り返しても, サブゴールを達成できないため, 学習を終了できない可能性がある.

3 とらえるべき特徴に着目した訓練済みモデルの作成

本稿では図 1 の手順で訓練済みモデルを作成することで, とらえるべき特徴を根拠とした出力ができる訓練済みモデルを作成する. 手順 (1) では 3.1 節に示す手順でゴール指向分析を用い, とらえるべき特徴を抽出する. 手順 (2) ではデータセットを訓練データとテストデータに分割し, 訓練データを使用して, 深層学習アルゴリズムを使用してモデルを作成する. 手順 (3) では XAI を用いて, 手順 (2) で作成したモデルがアウトプットを出すために注目している特徴を可視化する. 手順 (4) では手順 3 の結果を人が分析し, モデルがとらえている特徴を分析する手順 (5) では 3.2 節に示す手順で学習結果をもとにとらえるべき特徴を更新する. 手順 (6) では手順 (5) で更新されたとらえるべき特徴の充足度を確認し, 学習の継続を判断する. 学習を継続する場合, 再度手順 (2) に戻り, 再度手順を進める.

3.1 ゴール指向分析によるとらえるべき特徴の分析

2.2 節で示したゴール指向分析を深層学習を用いるシステムのとらえるべき特徴分析に使用する. まずは, システムのゴール指向分析を行い, システムの要求を分析する. そして, 分析結果のうち, 訓練済みモデルによって対応する範囲を決定する. さらに, 訓練済みモデルのよって対応する範囲をサブゴールが特徴をとらえている状態の集合となるまで分割をする. 通常のゴール指向分析では, 上位ゴールを満たすための状態に制限を設けない. しかし, 本手法ではとらえるべき特徴を抽出し, 訓練済みモデルがとらている特徴と比較することが目的であるため, 特徴をとらえている状態まで分割する. そして, 最下層に現れた特徴を訓練済みモデルがとらえるべき特徴とする. これにより, モデルがとらえるべき特徴を抽出できる.

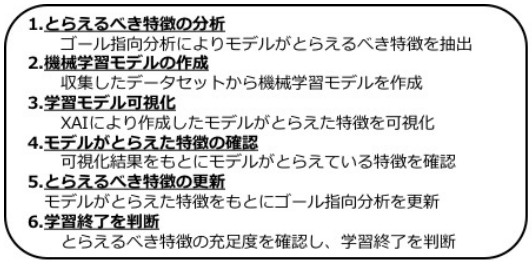


図 1 とらえるべき特徴に着目した訓練済みモデルの作成手順

3.2 学習結果に応じたとらえるべき特徴の更新

3.1 節の手法で抽出した特徴と手順 (4) で抽出した特徴を比較することで, 訓練済みモデルの作成時にモデルがとらえるべき特徴をとらえていることを確認することができる. しかし,2.2 節の通り, この手法を深層学習を用いたシステム開発に適用するためには課題がある. ゴール指向分析でとらえるべき特徴を分析したとしても, 訓練済みモデルがとらえるべき特徴に注目して学習するとは限らない.

そこで, 本稿ではこの問題を解決するため, 本稿では可視化された特徴をもとにとらえるべき特徴を再構築する手法を提案する. 手順 (4) で抽出したモデルがとらえた特徴をゴール指向分析の結果と比較する. そして, モデルとらえた特徴が, いづれかのサブゴールを達成するために有効であると判断できた場合, モデルがとらえた特徴を OR ゴールとしてゴール指向分析に追加する. 追加された特徴は,OR ゴールとして追加されるため, 手順 (1) で抽出したすべてのサブゴールを満たさない場合でも, トップゴールの達成が可能となる. 学習結果からとらえるべき特徴を再構築することで, ゴール指向分析の段階では発見できなかった, とらえるべき特徴を抽出することが可能となる. これ

表 1　実験で使用した手法およびデータセット

	名称	バージョン
データセット	GTSRB	-
ゴール指向分析	KAOS	-
学習アルゴリズム	TensorFlow	1.15.2
XAI	SHAP	0.34.0

により, 手順 (1) のサブゴールを達成しない場合でも, 学習の終了を判断できるようになる.

4　実験

本手法の有効性を確認するため, とらえるべき特徴に着目した訓練済みモデルを作成する実験を行った. この実験では表 1 に示す, 手法およびデータセットに対し, 本手法を適用した. 以下に示す手順に従い, ミッションクリティカルなシステムを想定し, 自動運転システムの一部機能である, 道路標識識別モデル作成した. 本実験では道路標識のデータセットとして,German Traffic Sign Recognition Benchmark(GTSRB) [4] を用いて 43 種類の道路標識を識別する訓練済みモデルを作成した.

1. KAOS 法により, 道路標識を識別するためにとらえるべき特徴を分析
2. TensorFlow を用いて道路標識を識別するモデルを作成
3. SHAP を用いてモデルが注目している特徴を可視化
4. 可視化されたモデルが注目している特徴をもとに KAOS の分析結果を更新
5. とらえるべき特徴の分析結果と可視化された特徴を比較し, モデルがとらえるべき特徴を根拠に出力ができているかを確認
6. とらえるべき特徴を根拠に出力ができていると判断できた場合は学習を終了し, そうでない場合はデータセット数を増やし, 手順 (2) から再度学習

4.1　道路標識識別システムにおけるとらえるべき特徴抽出

実験手順 (1) では,KAOS 分析により道路標識を識別するモデルが注目すべき特徴を分析した. 分析結果の一例として,50km の速度表示を識別するために, とらえるべき特徴を表 2 に示す. このような分析を 43 種類のラベルすべてに適用し, 道路標識を識別するモデルが注目すべき特徴を抽出した.

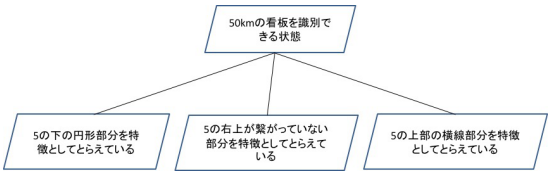


図 2　50km 速度表示のとらえるべき特徴

4.2　学習済みモデルの作成

本実験では道路標識のデータセットである,GTSRB のデータセット [4] と, 表 1 に示したバージョンの TensorFlow を使用して訓練済みモデルを作成した.GTSRB のデータセットのうち全体の約 80%にあたる,31367 件のデータをランダムに抽出し, 訓練データとした. また, 残りの 7842 件をテストデータとした. また, 学習の過程で訓練済みモデルのがとらえるべき特徴をとらえていることを確認するために訓練データ全体から 5%,30%,50%,80%,100%のデータを取り出し, 学習させたモデルを作成した.

4.3　SHAP による特徴の可視化

実験手順 (3) では XAI の手法として,SHAP を用いて作成したモデルを可視化した.SHAP の出力ではモデルが答えを出すために寄与率の高い特徴が SHAP 値で出力される. これを, 訓練データの画像上にプロットした. 出力の一例として 50km の速度表示看板の例を図 3 に示す. 図中の赤く着色された部分は, その画像を正しく識別するために重要と判断される特徴を示す. 青く着色された部分は, その画像の識別に対して誤った判断を誘発している特徴を示す. 図 3 では,50km の速度表示看板を識別するために,5 の上部の横線や, 下部の丸くあいた空間が重要視されている. 一方で,5 の左側の空間は 5 を識別するために, 誤識別

図 3　SHAP により可視化された特徴

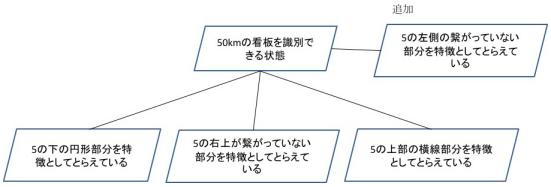


図 4　更新された 50km 速度表示のとらえるべき特徴

を誘発している特徴として抽出されている. このように, 学習毎に GTSRB のデータについて, 図 3 のように特徴を可視化した.

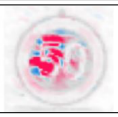
4.4　とらえるべき特徴の更新

実験手順 (4) では手順 (3) で可視化した特徴のうち, ゴールを達成するために有効であると判断できるものを抽出し, とらえるべき特徴を更新した. 更新の一例として 50km 速度表示の例を表 4 に示す. 実験手順 (1) で作成した, 図 2 の 50km の速度表示を識別するためのとらえるべき特徴について,5 の左側の繋がっていない部分を特徴としてとらえている, というサブゴールが追加されている. このようなサブゴールの更新を学習毎に GTSRB のデータについて行った.

4.5　実験結果

今回の実験では, 実験手順 (2) から実験手順 (6) を学習データセット数を増やしながら 5 回繰り返した. また, 学習毎に SHAP を用いたとらえた特徴の可視化を行った. 学習によるとらえた特徴の変化の例として,50km の速度表示の 1 回目の学習後と 2 回目の学習後のモデルがとらえた特徴を, 図 2 に示す. 図 2 で示した 2 枚の画像は同じテストデータの画像について, モデルが注目している特徴を可視化したものである.1 回目の学習よりも,2 回目の学習ではモデルがより多くの特徴に注目していることがわかる. また, とらえた特徴をもとに判断したとらえるべき特徴の充足確認結果を図 3 に示す. 図の中でとらえるべき特徴を充足したと判断した場合は, サブゴールを白抜きとして表す. 図 3 の 1 回目終了時のとらえるべき特徴の充足度分析では, 可視化結果からとらえるべき特徴として,5 の左側の繋がっていない部分を特徴としてとらえている, というサブゴールを追加した. しかし, データセット全体としてこの特徴に注目していると判断できなかったため, とらえるべき特徴の充足度としては, 未達成とした. 2 回目終了時のとらえるべき特徴の充足度分析では, とらえるべき特徴の多くを充足できたと判断した. この結果から,2 回目の学習の終了時点で,50km の道路標示については, 識別のためにモデルがとらえるべき特徴をとらえていると判断した.

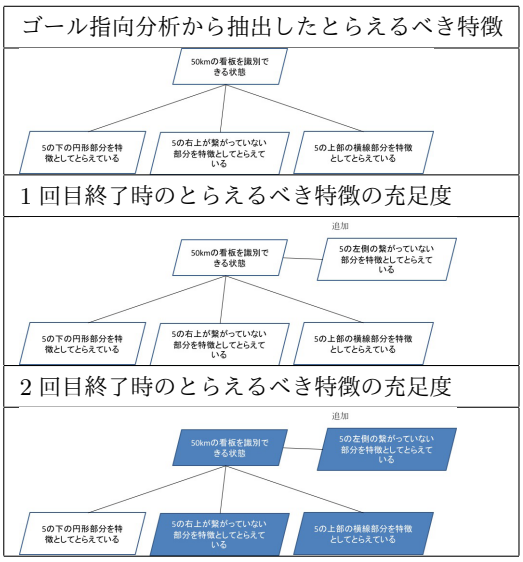
表 2　50km 道路標示の SHAP 分析結果の推移

1 回目	2 回目
	

5　評価と考察

4.5 節の実験の結果から,XAI を用いることで, ゴール指向分析によって抽出したとらえるべき特徴を訓練済みモデルがとらえていることを確認できた. これにより, ゴール指向分析によって, 訓練済みモデルがとらえるべき特徴を分析できることがわかる. また,XAI によって可視化された特徴のいためであると考えらえれる. しかし, 看板の右下が空白という情報は, 右折の情報とは関連が薄いうち, 識別に有効であるものをとらえるべき特徴に反映させることが可能であることも確認できた. 実験で使用した GTSRB のデータセットには絵に近い画像も存在する. これらについても, とらえるべき特徴の分析が行えたことから, 一般的な物体認識への適用可能性が示唆された.

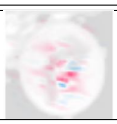
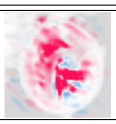
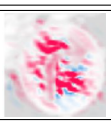
表 3　50km 道路標示のとらえるべき特徴の充足度分析結果の推移



5.1　モデルがとらえた意図しない特徴

今回の実験では, モデルが意図しない特徴をとらえた場合も存在した. 図 4 に学習段階ごとの, 右折限定の道路標示の SHAP 分析結果を示す. また, この表示のとらえるべき特徴の分析結果を図 5 に示す. この例では, とらえるべき特徴として, 矢印に注目していることを抽出した. しかし, 図 4 のように訓練済みモデルは,5 回の学習を繰り返した場合でも, 矢印の部分にはほとんど注目していない. これは,GTSRB のデータの中には看板の右下が空白のデータは含まれていな. そのため, だまされやすい部分に注目して識別結果を出しているといえる. この問題を解決するため, 学習状況に合わせた訓練データの選定などの応用が必要になると考えられる.

表 4　右折限定道路標示の SHAP 分析結果の推移

1 回目	3 回目	5 回目
		

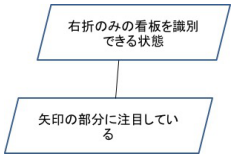


図 5　右折限定標識のとらえるべき特徴

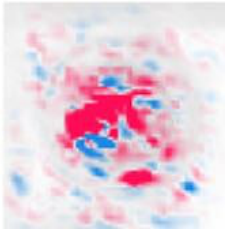


図 6　目視で意味を確認できなかった特徴

5.2　特徴可視化の課題

今回の実験の中では,XAI の出力のうち人間が理解できない特徴が存在した. 図 6 に 50km の速度表示看板がとらえた特徴のうち, 目視では意味を理解できなかったものを示す. このような特徴は, 妥当であることの確認ができないため, とらえるべき特徴の更新に使用できない. この問題を解決するため, 訓練データの特徴に対して, 分析できる手法との組み合わせを検討する必要がある. 今回の実験では, 図 6 の特徴について, 理解ができなかったが, データセットの特徴について理解することで, 妥当な特徴として再検討できる可能性がある. また, 他の XAI 手法の活用も検討ですする必要がある.XAI として今回は SHAP を用いたが, 他の XAI を用いてとらえた特徴の可視化を行うことで, 今回は理解できなかった特徴を理解できる可能性がある.

6　まとめ

本研究では深層学習を用いたシステム開発について, とらえるべき特徴に着目した, 訓練済みモデル作成のプロセスを提案した. 実験の結果, 本手法のプロセスに従い, モデルを作成することで, 訓練済みモデルがとらえるべき特徴に着目して, 学習を進めている見通しを立てることは可能であることは確認された.

今後は学習の途中でとらえるべき特徴以外の部分にモデルが注目したことが明らかになった場合の対応について, 検討を進めていく. また, 対象となるデータに対する分析手法との組み合わせや, 画像認識以外のシステムへの適用についても検討していく.

謝辞 本論文を纏めるにあたり, 国立情報学研究所トップエスイープロジェクトの関係者様に, 多大なご助力をいただきました. 厚く感謝を申し上げます.

参 考 文 献

[1] Dardenne, A., van Lamsweerde, A., and Fickas, S.: Goal-directed requirements acquisition, *Science of Computer Programming*, Vol. 20, No. 1(1993), pp. 3 – 50.

[2] Lundberg, S. and Lee, S.: A unified approach to interpreting model predictions, *CoRR*, Vol. abs/1705.07874(2017).

[3] Ribeiro, M. T., Singh, S., and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *CoRR*, Vol. abs/1602.04938(2016).

[4] Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C.: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition, *Neural Networks*, No. 0(2012), pp. –.