アテンションマップを用いたKS検定によるデータのドリフト検知

新田常顧† 史宇植† 平川翼† 山下隆義† 藤吉弘亘†
† 中部大学
E-mail: ntsunemi0122@mprg.cs.chubu.ac.jp

1 はじめに

データのドリフトは，運用時のデータ分布がモデル学習時のデータ分布から時間とともに変化することであり，その検知は重要な課題である．ドリフトは，カメラのレンズの経年劣化により発生するノイズやピントのズレ，カメラの交換によって発生する位置ズレや反転などにより生じる．このようなドリフトが発生すると，学習時と異なるデータ分布となり，機械学習モデルの性能低下の原因となる [1]．これらのことから，特に，外観検査等の運用の際にデータ分布の変化を監視してドリフトを検知することが重要となる．

ドリフト検知は，学習時のデータ分布と運用時のデータ分布を 2 標本検定により比較することで行われる．Lipton らは，ラベル分布が変化することにより発生するラベルドリフトの検知手法を提案した [2]．また，Rabanser らは，入力データの特徴分布が変化することにより発生するデータドリフトに対するドリフト検知手法を提案した [3]．両者の手法は，どちらも学習済みモデルの出力するクラス確率分布に対して 2 標本検定を行うことでドリフト検知を行う．しかし，これらの手法は 2 標本検定の出力する分布の差からのみドリフト検知結果に対する分析が可能であり，実際にドリフトによって機械学習モデルの予測性能が低下していることへの分析が行われていない．

そこで本研究では，モデルの判断根拠の可視化が可能である Attention Branch Network (ABN) [4] と 2 標本検定によるドリフト検知法を提案する．本手法では，ABN を構成する Attention branch と Perception branch の出力するクラス確率分布と，モデル推論時に高く貢献した領域を示すアテンションマップに対して 2 標本検定を行い，ドリフトの検知を行う．また，ABN を用いることによりドリフトデータに対してアテンションマップを獲得し，検知結果に対する分析を行う．

2 関連研究

本章では，ドリフトの検知手法と，モデルの注視領域の可視化方法について述べる．

2.1 ドリフト検知

ドリフト検知は，モデル学習時のデータ分布と運用時のデータ分布を 2 標本検定により比較することで行われる．2 標本検定は，2 つの分布に有意な差があるかどうかを確かめる手法である．ドリフト検知に用いられる代表的な 2 標本検定として，Kolmogorov-Smirnov (KS) 検定や Maximum Mean Discrepancy (MMD) [5] が挙げられる．

KS 検定は，式 (1) に示すように，2 つの分布を累積して得られる累積分布の最大差を統計検定量とし，有意な差があるかの判定を行う．

$$D = \max |F_X(x) - F_Y(x)| \tag{1}$$

ここで，D は検定統計量，$(F_X(x), F_Y(x))$ はそれぞれ標本 (X, Y) から得られる累積分布関数を意味する．

MMD は，カーネル関数を用いて分布間の距離を推定し，2 つの分布に有意な差があるかの判定を行う．カーネル関数とはデータの類似度を計算するための関数である．カーネル関数には様々な種類があるが，ここでのカーネル関数は式 (2) に示すガウシアンカーネルとする．

$$k(x, x') = e^{\frac{1}{\sigma}\|x - x'\|^2} \tag{2}$$

ここで，σ は正の定数，(x, x') は 2 つの分布の入力である．MMD を式 (3) に示す．

$$\widehat{MMD} = ||\hat{P_X} - \hat{P_Y}||_H^2 \tag{3}$$

ここで，$\widehat{MMD}$ は検定統計量，$(\hat{P_X}\ \hat{P_Y})$ は 2 つの分布 (X, Y) をカーネル関数により変換した特徴分布である．H は $(\hat{P_X}, \hat{P_Y})$ の存在する空間を意味する．

ドリフトは，Zhao ら [6] によって，コンセプトドリフト，ラベルドリフト，データドリフトの 3 つが定義されている．ここで，コンセプトドリフトは，モデル学習時と運用時で入力データに対するデータの解釈や概念が変化することを意味する．ラベルドリフトは，モデル学習時のラベル分布と運用時のラベル分布が変化することを意味する．このようなデータの分布の一見些細な変化が，モデルの分類器の性能に影響を与えることが知られており [7]，不確実性の下でモデルの意思決

定が行われる場合，ラベルの分布が変化するだけでもモデルの性能に影響を与えることが知られている [8].

Lipton ら [2] は，ラベルドリフトの検知方法として，学習済みモデルの出力するクラス確率分布に対して KS 検定および MMD を用いる手法を提案した.

しかしドリフトには，ラベルドリフトだけでなく，入力の特徴分布が変化することを意味するデータドリフトと呼ばれるものがある．データドリフトの例として，カメラのレンズの経年劣化によるノイズや，ピントのズレなどが挙げられる．そこで Rabanser ら [3] は，ラベルドリフトとデータドリフトの双方の検知を目的とした研究を行った．Rabanser らの手法では，それぞれのデータに対して次元削減を行い，2 標本検定により分布を比較することでドリフト検知が行われる．結果として，データドリフトの検知においても，ResNet-18 [9] が出力するクラス確率分布に対して，KS 検定を用いる検知手法が最も高精度であることを示した.

2.2 モデルの注視領域の可視化

モデルの注視領域の可視化方法として，Class Activation Mapping (CAM) [10] が挙げられる．CAM は，畳み込み層の応答値を用いて，ネットワークが認識において高く貢献した領域をヒートマップで表現することができる．この可視化されたヒートマップは，Class Activation Map と呼ばれる．CAM では，Global Average Pooling (GAP) により出力された各チャンネルにおける特徴マップの平均を重みとし，それぞれの特徴マップの重み付き和から Class Activation Map を生成する．このように，CAM では Class Activation Map を出力するために GAP を必要とするため，モデル構造に制限がかかる．Gradient-weighted Class Activation Mapping (Grad-CAM) [11] では，各チャンネルの重みを勾配から計算する．そのため，モデル構造に対する制限がなく，モデルの一般化に成功している．しかし，これらの手法は全結合層を畳み込み層に入れ替える等の処理が必要なため，画像分類においては性能低下を引き起こしやすいとされている.

そこで，提案されたのが Attention Branch Network (ABN) である．ABN は，従来の視覚的説明モデルが性能低下を引き起こしやすい問題を解決するために，視覚的説明モデルから生成されるアテンションマップを Attention 機構へ応用し，視覚的説明モデルの性能向上とアテンションマップによる注視領域の可視化を同時に実現したモデルである．Attention 機構は特定の領域の特徴を強調することでネットワークの汎化性能を向上させる手法である．ABN は Feature extractor，Attention branch，Perception branch の 3 つのモジュールから構成される．Feature extractor は入力画像に対する特徴マップを出力する．Attention branch は CAM と同様に畳み込み層と GAP から構成し，アテンションマップを出力する．Perception Branch では，Feature extractor から出力される特徴マップと，Attention branch から出力されるアテンションマップの内積を取ったものを新たな特徴マップとし，最終的な各クラスにおける確率を出力する．また，Feature extractor と Perception branch はベースラインのネットワークを特定の層で分割することで構築し，Attention branch と Perception branch から出力される学習誤差を用いることで学習する.

3 提案手法

本研究では，ABN を用いたドリフト検知法を提案する．ABN では，Attention branch から得られるアテンションマップを Attention 機構に入力し，特定領域の特徴を強調して推論を行う．そのため，ドリフトによるアテンションマップの変化によって，入力画像の変化に対する特徴が強調され，ドリフトの検知精度が向上すると考える．ABN を用いたドリフト検知の流れを図 1 に示す.

3.1 クラス確率分布を用いたドリフト検知

本手法では，ABN を構成する層の特徴空間が異なる点に着目し，3 つの出力からドリフトの検知を行う．1 つ目は Attention branch の GAP に対して softmax から出力されるクラス確率分布，2 つ目は Perception branch から出力されるクラス確率分布である.

まず，KS 検定を行う際に用いる累積分布を式 (4) に示す．ここで，n はサンプル数，$X_i(x)$ は度数，x_i はサンプルを意味する．x は階級を意味し，x_i に対する度数を算出する．累積分布は，この度数の総和をサンプル数で割ったもので表される.

$$
\begin{aligned}
F_n(x) &= \frac{1}{n}\sum_{i=1}^{n} X_i(x)\ , \\
X_i(x) &= \begin{cases} 1\ (x_i \leq x) \\ 0\ (x_i > x) \end{cases}
\end{aligned}
\tag{4}
$$

Attention branch 及び Perception branch の出力するクラス確率は，クラス数 C 個分存在するため，1 枚のサンプル当たりに出力されるクラス確率 x は，$\{x_1, x_2, ..., x_C\}$ となる．クラス確率分布を用いるドリフト検知では，各クラスのクラス確率ごとに KS 検定を行うため，Attention branch の出力するクラス確率の累積分布 D_c^{AB} は式 (5) のように表される．ここで，c は $1 \sim C$ のクラス番号，n はサンプル数，x_c^i は i 番目のサンプルに対するクラス番号 c のクラス確率である．また，同様に Perception branch の出力するクラス確率分布 D_c^{PB}，運用時データに対する Attention branch，Perception branch の出力するクラス確率分布

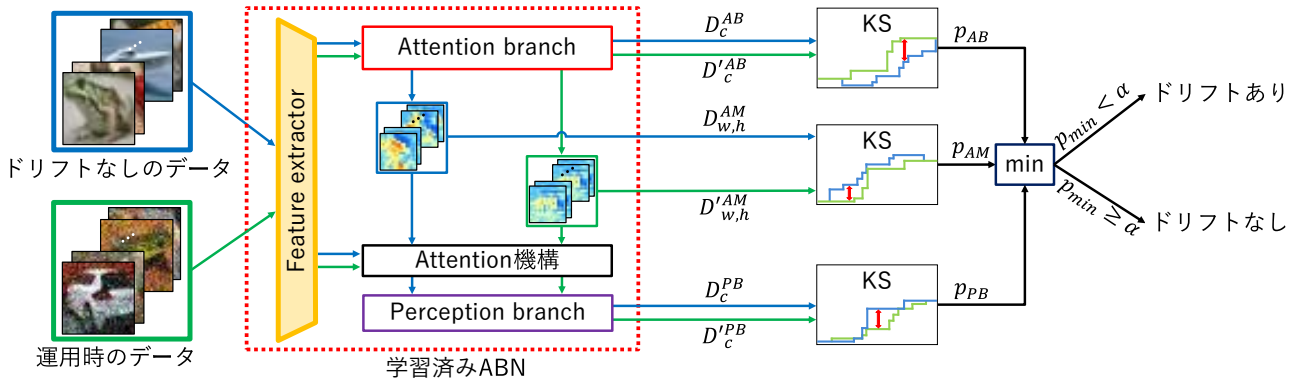


図 1: 提案手法によるドリフト検知の流れ

$(D_c'^{AB}, D_c'^{PB})$ も式 (5) と同様に算出する．

$$D_c^{AB} = \frac{1}{n}\sum_{i=1}^{n} X_c^i(x)\ ,$$
$$X_c^i(x) = \begin{cases} 1\ (x_c^i \le x) \\ 0\ (x_c^i > x) \end{cases} \tag{5}$$

次に，式 (5) で算出した $(D_c^{AB}, D_c'^{AB})$ に対して，式 (6) に示すように各クラスのクラス確率ごとに KS 検定を行い，結果の最小値を p_{AB} とする．p 値は，2 つの分布の母集団が同じであるという仮定の下で，KS から得られる検定統計量がそれ以上の値となる確率である．また，KS 検定をクラス数分繰り返して p 値を算出する多重検定を行うため，ドリフトの判定に用いる有意水準 α に対してボンフェローニ補正 [12] が行われる．また，Perception branch の出力するクラス確率分布に対しても，式 (7) に示すように p_{PB} を算出する．

$$p_{AB} = \min_c(\mathrm{KS}(D_c^{AB}, D_c'^{AB})) \tag{6}$$

$$p_{PB} = \min_c(\mathrm{KS}(D_c^{PB}, D_c'^{PB})) \tag{7}$$

3.2 アテンションマップを用いたドリフト検知

本手法では，モデルの注視領域を示すアテンションマップもドリフト検知に用いる．アテンションマップをドリフト検知に応用することで，Rabanser らの手法 [3] と異なり，モデルの実際に注視している領域の変化の特徴を捉えることが可能である．

アテンションマップを用いたドリフト検知では，各画素ごとに KS 検定を行う必要があるため，アテンションマップの特徴分布 $D_{w,h}^{AM}$ は式 (8) のように表される．ここで，(w,h) はアテンションマップの各画素の座標値を表す．また，運用時データに対するアテンションマップの特徴分布 $D_{w,h}'^{AM}$ も式 (8) と同様に算出する．

$$D_{w,h}^{AM} = \frac{1}{n}\sum_{i=1}^{n} X_{w,h}^i(x)\ ,$$
$$X_{w,h}^i(x) = \begin{cases} 1\ (x_{w,h}^i \le x) \\ 0\ (x_{w,h}^i > x) \end{cases} \tag{8}$$

次に，式 (8) で算出した $(D_{w,h}^{AM}, D_{w,h}'^{AM})$ に対して，式 (9) に示すようにアテンションマップの各画素ごとに KS 検定を行い，結果の最小値を p_{AM} とする．また，クラス確率分布を用いたドリフト検知と同様に，KS 検定を行う回数に応じて有意水準 α がボンフェローニ補正によって調整される．

$$p_{AM} = \min_{w,h}(\mathrm{KS}(D_{w,h}^{AM}, D_{w,h}'^{AM})) \tag{9}$$

3.3 提案手法の流れ

提案手法では，以下のデータセットの前処理と Step1 から Step3 の流れでドリフト検知を行う．

Step1. ABN の学習

学習用データを用いて ABN を学習する．ここで，ABN は学習用データにはドリフトのシミュレーションは行われず，学習用データに対して画像分類の精度を向上させるように学習する．

Step2. クラス確率分布とアテンションマップの算出

ドリフトなしのデータと運用時のデータから，指定したサンプル数を取得し，学習済み ABN に入力する．それぞれのデータ群に対して Attention branch，Perception branch の出力するクラス確率分布を式 (5) のように算出する．また，アテンションマップの特徴分布を式 (8) のように算出する．算出した分布は，ドリフト検知を行うために一時的に保存する．

Step3. ドリフト検知

Step2 で求めた各分布に対して，それぞれ KS 検定を行い，p 値 (p_{AB}, p_{AM}, p_{PB}) を算出する．

次に，式 (10) に示すように，各 p 値の最小値を取ることで結果を統合し，p_{min} とする．求められた p_{min} は，3 つの出力に対して KS 検定を行った際，最もドリフトなしのデータと運用時のデータの分布との差が大きいと判断されたものである．これにより，3 つの出力を用いて共同でドリフトの検知を行う．

$$p_{min} = \min(p_{AB}, p_{AM}, p_{PB}) \tag{10}$$

最後に，$p_{min} < \alpha$ であればドリフトありと判定する．ここで，α は Rabanser らの手法 [3] と同値の閾値であ

り，KS 検定を行う回数に応じてボンフェローニ補正が行われる．この値は 2 標本検定で一般的に用いられる有意水準 5% である．

4 評価実験

提案手法の有効性を調査するために，Rabanser らの手法 [3] とドリフト検知精度の比較を行う．

4.1 データセット

本実験では，MNIST データセットと CIFAR-10 データセットを用いる．MNIST データセットは，0~9 の手書き白黒画像で構成されるデータセットで，学習用データ 5 万枚，検証用データ 1 万枚，テスト用データ 1 万枚で構成される．CIFAR-10 データセットは，10 種類の物体カラー画像で構成されるデータセットで，学習用データ 4 万枚，検証用データ 1 万枚，テスト用データ 1 万枚で構成される．

各データセットの学習用データは ABN の学習に用いる．また，検証用データは学習用データと同じデータ分布であると仮定し，ドリフトなしのデータとする．テスト用データには，ドリフトのシミュレーションを施し，運用時のデータとする．

4.2 ドリフトのシミュレーション

本実験で用いるドリフトのシミュレーション方法は以下の通りである．

ガウスぼかし

ガウスぼかしは，ガウス関数を用いて画像をぼかす処理である．本実験では，ぼかしの強度を 3 段階用意し，それぞれの強度に対しテスト用データにシミュレーションを施す割合を {10%,50%,100%} の 3 段階用意する．

ガウシアンノイズ

ガウシアンノイズは，画像の各画素の輝度を正規分布に従い変更することでノイズを付加する処理である．本実験では，ノイズの強さを 3 段階用意し，それぞれの強度に対しテスト用データにシミュレーションを施す割合を {10%,50%,100%} の 3 段階用意する．

幾何変換

幾何変換では，画像に対して回転，水平・垂直移動，せん断，拡大・縮小，水平・垂直方向への反転を組み合わせて変換を行う．本実験では，幾何変換の各要素の強さを 3 段階用意し，それぞれの強度に対しテスト用データにシミュレーションを施す割合を {10%,50%,100%} の 3 段階用意する．

データの不均衡化

データの不均衡化では，特定の 1 つクラスのサンプル数を削減する処理を行う．実験では，特定のクラスのサンプル数を削減する割合を {10%,50%,100%} の 3 段階用意する．

4.3 評価指標

本実験では，全てのドリフトのシミュレーションに対しての検知率を平均した値を用いて，提案手法と Rabanser らの手法の比較を行う．また，ドリフト検知に必要とするサンプル数を段階的に評価するために，ドリフト検知に用いるサンプル数を 8 段階用意する．実験では，ドリフトなしのデータと運用時のデータから指定したサンプル数を取得し，ABN に入力する処理を行う．

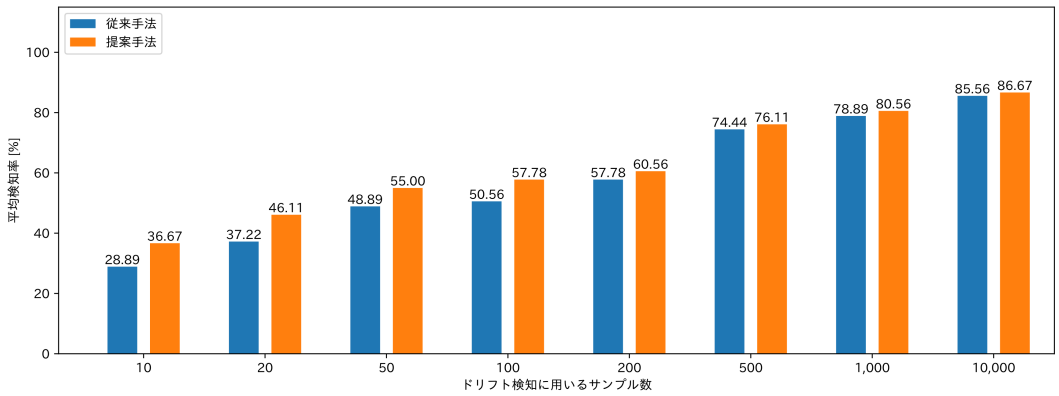
4.4 ドリフト検知率の比較

まず，MNIST データセットでの Rabanser らの手法と提案手法のドリフト平均検知率の比較を図 2(a) に示す．図 2(a) より，提案手法は従来手法より概ね平均検知率が向上していることが分かる．次に，CIFAR-10 データセットでのドリフト平均検知率の比較を図 2(b) に示す．図 2(b) より，平均検知率が向上していることが分かる．また，MNIST データセットでの検知と比較し，CIFAR-10 データセットでの検知の方が平均検知率の向上率が大きいことが分かる．これらの結果より，提案手法によるドリフト検知は有効であると言える．

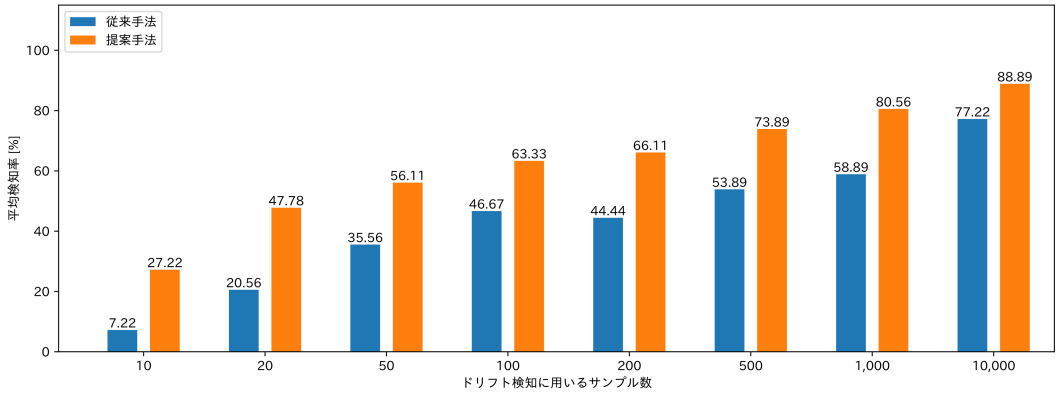
次に，ドリフトをシミュレーションした画像に対するアテンションマップの変化を確認する．図 3 に，ドリフトをシミュレーションしていない画像に対するアテンションマップと，ガウスぼかし及びガウシアンノイズを付加した画像を入力した際に得られるアテンションマップを示す．図 3 より，ガウスぼかし，ガウシアンノイズによるドリフトのシミュレーションを行った際は，注視領域が縮小する傾向があることが分かる．また，図 3(c) のようなガウスノイズ付加前と付加後の入力画像の変化を見分けにくい例であっても，アテンションマップを獲得することでモデルの注視領域から変化を理解することが可能である．提案手法では，アテンションマップを用いて推論を行う ABN を用いるため，これらの変化を得ることでドリフトの検知精度が向上していると考えられる．

次に，平均検知率の分析を行うために，向上率が大きかった CIFAR-10 データセットに対する各ドリフトのシミュレーション方法ごとの検知率の比較を行う．ここで，検知率の評価を容易にするため，ドリフト検知に用いるサンプル数は 100 枚，1,000 枚とする．実験結果を表 1 に示す．表 1 から，平均検知率の向上に最も寄与しているドリフトのシミュレーション方法は，ガウシアンノイズであることが分かる．また，ガウスぼかし 1,000 枚に対する検知率は 26.6pt の向上と，ガウシアンノイズの次に向上率が高いことが分かる．

次に，各ドリフトのシミュレーション方法における検知率に対して，提案手法の有効性の検証を行う



(a) MNIST



(b) CIFAR10

図 2: ドリフト平均検知率の比較

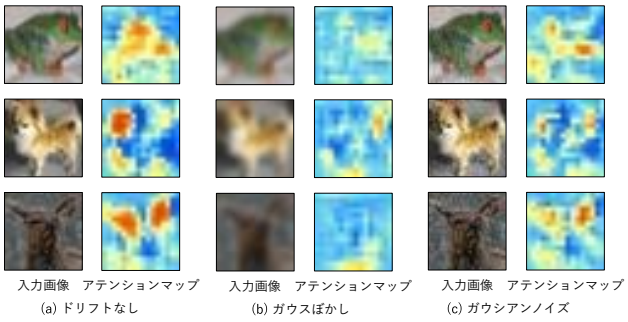


図 3: アテンションマップの比較

ために，提案手法において算出される p 値である (p_{AB}, p_{AM}, p_{PB}) のそれぞれの貢献度を調査する．ここで，貢献度はドリフトの検知に成功した際に選ばれた p 値の回数をそれぞれカウントし，全ての選出回数から割った割合とする．実験結果を表 2 に示す．表 2 より，全体的に p_{PB} の貢献度が高いことが分かる．しかし，ガウスぼかしにおいては p_{AB} が約 48% 採用されており，3 つの中で最も高く，p_{AM} が 12.6% 採用されていることが分かる．ガウシアンノイズにおいては p_{AM} が約 9% 採用されていることが分かる．このことから，3 つの出力を統合する方法はドリフト検知において有効であることが分かる．また，ABN から獲得できるアテンションマップをドリフト検知に用いることが可能であり，特定の条件下ではモデルのクラス確率を用いる手法より有効であり，ドリフトの平均検知率の向上に貢献していると言える．

5 おわりに

本研究では，ABN と Kolmogorov-Smirnov 検定によるデータのドリフト検知手法を提案した．提案手法では，ABN を構築する Attention branch と Perception branch の出力するクラス確率分布と，アテンションマップを用いて KS 検定を行い，その結果を統合することでドリフト検知を行う．評価実験により，提案手法は従来手法よりドリフトの平均検知率が向上したことを確認した．また，ドリフトデータに対するアテンションマップを獲得することでドリフト検知結果の分析を行った．また，それぞれの出力がドリフト検知へどれほど貢献しているかを調査し，クラス確率分布だけでなく，アテンションマップを用いてドリフト検知を行うことの

表 1: 各ドリフト検知率の比較 (CIFAR-10)

ドリフトの種類	手法	ドリフト検知に用いるサンプル数	
		100	1000
ガウスぼかし	従来手法	53.3%	57.8%
	提案手法	60.0%	84.4%
ガウシアンノイズ	従来手法	24.4%	26.7%
	提案手法	53.3%	68.9%
幾何変換	従来手法	46.7%	73.3%
	提案手法	66.7%	86.7%
データの不均衡化	従来手法	62.2%	77.8%
	提案手法	73.3%	82.2%

表 2: 各 p 値の貢献度 (CIFAR-10)

ドリフトの種類	p_{AB}	p_{AM}	p_{PB}
ガウスぼかし	47.9%	12.6%	39.5%
ガウシアンノイズ	34.9%	8.9%	56.2%
幾何変換	41.6%	2.6%	55.8%
データの不均衡化	46.7%	2.2%	51.1%

有効性を示した．今後の課題としては，ドリフト検知結果に対する詳細な分析及び，ドリフトのシミュレーション方法の追加，各分布に適した 2 標本検定の導入による高精度化などが挙げられる．

参考文献

[1] D. Hendrycks and T. Dietterich: “Benchmarking neural network robustness to common corruptions and perturbations”, International Conference on Learning Representations (2019).

[2] Z. Lipton, Y.-X. Wang and A. Smola: “Detecting and correcting for label shift with black box predictors”, Proceedings of the 35th International Conference on Machine Learning (Eds. by J. Dy and A. Krause), Vol. 80 of Proceedings of Machine Learning Research, PMLR, pp. 3122–3130 (2018).

[3] S. Rabanser, S. Günnemann and Z. Lipton: “Failing loudly: An empirical study of methods for detecting dataset shift”, Advances in Neural Information Processing Systems (Eds. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett), Vol. 32, Curran Associates, Inc. (2019).

[4] H. Fukui, T. Hirakawa, T. Yamashita and H. Fujiyoshi: “Attention branch network: Learning of attention mechanism for visual explanation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[5] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf and A. Smola: “A kernel two-sample test”, Journal of Machine Learning Research, **13**, 25, pp. 723–773 (2012).

[6] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia and K. Keutzer: “A review of single-source deep unsupervised visual domain adaptation”, IEEE Transactions on Neural Networks and Learning Systems, **33**, 2, pp. 473–493 (2022).

[7] D. Zügner, A. Akbarnejad and S. Günnemann: “Adversarial attacks on neural networks for graph data”, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining, ACM (2018).

[8] K. Zhang, B. Schölkopf, K. Muandet and Z. Wang: “Domain adaptation under target and conditional shift”, Proceedings of the 30th International Conference on Machine Learning (Eds. by S. Dasgupta and D. McAllester), Vol. 28 of Proceedings of Machine Learning Research, Atlanta, Georgia, USA, PMLR, pp. 819–827 (2013).

[9] K. He, X. Zhang, S. Ren and J. Sun: “Deep residual learning for image recognition”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba: “Learning deep features for discriminative localization”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).

[11] R. R. Selvaraju, A. Das, R. Vedantam,

M. Cogswell, D. Parikh and D. Batra: "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization", CoRR, **abs/1610.02391**, (2016).

[12] O. J. Dunn: "Multiple comparisons among means", Journal of the American Statistical Association, **56**, 293, pp. 52–64 (1961).