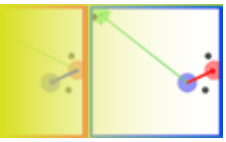
機械知覚&ロボティクスグループ
中部大学　工学部 ロボット理工学科

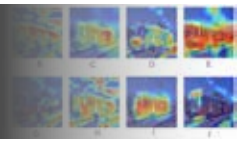
MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

2022年度　藤吉研究室　卒業論文発表　アブストラクト

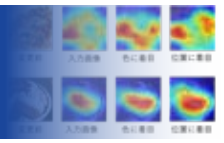
Deep Leaning, Reinforcement Learning, Multi Agent
Actor attention による協調タスクにおけるマルチエージェントの行動説明
佐川 統夢



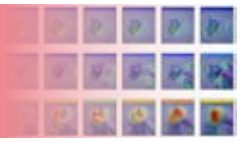
Deep Learning, Explainable AI, Self-Supervised Learning
人の順位付けによる深層学習モデルの視覚的説明の傾向調査
若山 浩之



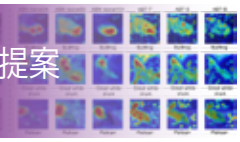
Deep Learning, Explainable AI, image recognition
シングルモーダル学習モデルとマルチモーダル学習モデルにおける判断根拠の分析
岡島 滉樹



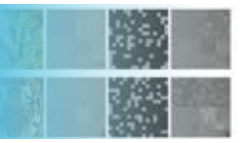
Deep Learning, Action Recognition, Explainable AI
人の知見を導入した ST-ABN による動画像認識の高精度化
野口 紗季



Deep Learning, Explainable AI, Robustness
ノイズに対する頑健性と視覚的説明性を獲得する Attention Branch Transformer の提案
鈴木 雅司



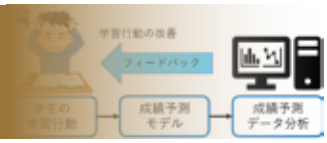
Deep Learning，Self-Supervised Learning，Explainable AI
Wavelet Masked Autoencoder による Wavelet Transformer の精度向上
福井 雅弥



Self-Supervised Learning，Data Augmentation
SimSiam におけるデータ拡張の効果と mixup の導入による向上
岩垣 渚

Deep Learning, Sports, Gaze Prediction
深層学習モデルによる視線予測のための教師データの作成法とその効果
梅谷 将成

Deep Learning, Explainable AI
Transformer の Attention weight を用いた成績予測の傾向調査と改善
伴 優輝



Deep Learning，object detection, simulation
物体検出における実環境と CG データの差異の調査
鈴木 陽太郎

Deep Learning, Semantic Segmentation
Multi Object Tracking を用いた点滅信号の識別
櫻井 洋貴

Deep Learning, Semantic Segmentation, Semi-Supervised Learning
マルチドメインに対応した半教師あり学習によるセマンティックセグメンテーション
中村 泰智

Actor attention による協調タスクにおけるマルチエージェントの行動説明

ER19035 佐川統夢　　指導教授：藤吉弘亘

1. はじめに

マルチエージェント深層強化学習は，複数のエージェントを同一環境で学習する手法である．マルチエージェント深層強化学習の応用例には，サッカーロボットなどが挙げられる [1]．サッカーロボットはエージェント間でのパス回しなど協調行動を獲得することで，目的であるゴールを達成している．一方で，マルチエージェント深層強化学習において，獲得した協調行動に対する判断根拠が不明という問題がある．そこで本研究では，Actor-Critic 法の Actor に対し，アテンションヘッドを導入する．これにより，対象エージェントの注視度を表す Actor attention を獲得でき，協調行動に対する判断根拠の解析が可能となる．

2. Multi-Actor-Attention-Critic

Multi-Actor-Attention-Critic (MAAC) [2] とは，Actor-Critic 法の Critic にアテンションヘッドを導入したマルチエージェント深層強化学習手法である．MAAC Critic は，アテンションヘッドを用いて他エージェントの貢献度を算出する．そして，その貢献度から対象エージェントの行動価値関数を求めることで協調行動を獲得している．このとき，MAAC Critic に用いる各エージェントの状態と行動を入力とする MLP とアテンションヘッドのパラメータはエージェント間で共有している．これにより，効率的に協調行動を獲得できる．

3. 提案手法

本研究は協調行動に対する判断根拠の解析を目的とし，Actor にアテンションヘッドを導入した手法を提案する．Actor では，各エージェントの行動を選択するために，アテンションヘッドを導入し，対象エージェントの行動選択に対する他エージェントの貢献度を獲得する．提案手法の構造を図 1 に示す．

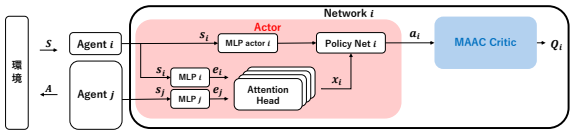


図 1：提案手法の構造

Actor の MLP は，対象エージェント i と対象エージェントを除いた他エージェント j の状態 $\mathbf{s}_i$，$\mathbf{s}_j$ を入力し，特徴量 $\mathbf{e}_i$，$\mathbf{e}_j$ を算出する．MLP で算出される特徴量 $\mathbf{e}_i$，$\mathbf{e}_j$ をアテンションヘッドに入力し，各要素間で内積し，次元数で正規化する．そして，Softmax 関数によりアテンションを算出する．また，アテンションと他エージェント j の特徴量 $\mathbf{e}_j$ から内積により他エージェント j の貢献度 $\mathbf{x}_i$ を算出する．本手法ではアテンションヘッドで算出したアテンションを Actor attention と呼称する．Actor attention は，対象エージェントがどの他エージェントを注視しているかを示す値である．Actor attention の算出式を式 (1) に示す．

$$att_{actor} = \mathrm{Softmax}\left(\mathbf{e}_i \cdot \mathbf{e}_j^T / \sqrt{dim}\right) \tag{1}$$

ここで $\mathbf{e}_j^T$ は転置した $\mathbf{e}_j$，dim は特徴量 e_i の次元数である．獲得した Actor attention を可視化することで，協調行動に対する判断根拠の解析が可能となる．

4. 評価実験

提案手法の有効性を確認するため，Multi Particle Environments の Simple_spread タスクを対象として評価する．

4.1. 実験環境

Simple_spread は各エージェントがそれぞれランドマークに向かうタスクであり，エージェントとランドマークはランダムな位置に複数生成される．各エージェントは他エージェントとの衝突を避けながらゴールであるランドマークに向かうことを目標とする．報酬設計は対象エージェントと全ランドマーク間における最小距離の負の報酬，衝突に対する負の報酬 (-1) とする．

4.2. 実験概要

Simple_spread を用いて MAAC と提案手法の学習を行い，最大報酬値と協調行動に対する定量的評価により比較する．学習条件は，エージェント数 3 (赤，緑，青)，ランドマーク数 3 (黒)，1 エピソード間の step 数を 25 とし，100,000 エピソード学習する．本実験では，他エージェントにランドマークを譲る行動を協調行動と定義する．この協調行動に対する定量的評価は，ランドマークが一定距離以上離れた環境において 100 エピソード間で対象エージェントが最も近いランドマーク以外へ移動した合計回数とする．この時，他エージェントは最も近いランドマークに移動するルールベース制御とする．また，Actor attention を可視化し協調行動に対する判断根拠を解析する．

4.3. 最大報酬と協調行動に対する評価

MAAC と提案手法の最大報酬値を表 1 に示す．表 1 から提案手法は MAAC と比較して最大報酬値が約 -1.5 低下している．協調行動に対する評価結果を表 2 に示す．表 2 から提案手法は MAAC より協調行動の合計回数が 56 回向上している．このことから Actor にアテンションヘッドを導入することで，最大報酬は低下するが，より良い協調行動を獲得できると考えられる．

表 1：最大報酬値

手法	最大報酬値
MAAC	-35.620
提案手法	-37.025

表 2：協調行動に対する評価

手法	赤	緑	青	合計
MAAC	50	63	69	182
提案手法	72	85	81	238

4.4. Actor attention の可視化

対象エージェントを赤と青とした場合の Actor attention の可視化例を図 2 に示す．可視化例は赤と青が協調行動を行ったシーンである．Actor attention の度合いを矢印の太さで表す．線が太いほど，Actor attention が強く注視していることを意味する．図 2 から step1，2，3 で赤と青は互いを注視している．ここから，赤は同じランドマークに向かう青を注視したことでランドマークを譲る行動をとり，青は赤を注視したことで赤の協調行動に合わせることができ，互いがランドマークに向かうことができたと考えられる．このように Actor attention を可視化することでエージェントの協調行動に対する判断根拠の解析を実現した．

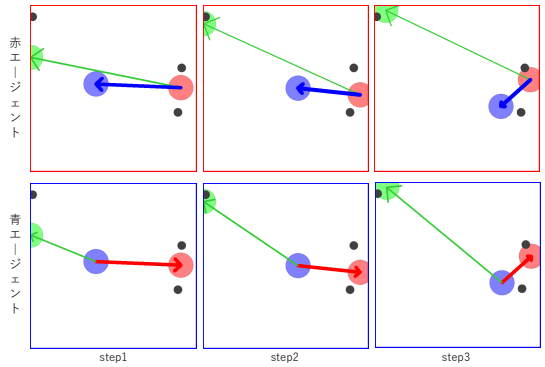


図 2：Actor attention の可視化

5. おわりに

本研究では，Actor-Critic 法の Actor に対しアテンションヘッドを導入した手法を提案した．評価実験から Actor にアテンションヘッドを導入することで，協調行動に対する判断根拠の可視化が可能となった．今後は，より良い協調行動の獲得と提案手法の汎用性の調査を行う．

参考文献

[1] J. Kim, *et al.*, “A cooperative multi-agent system and its real time application to robot soccer”, ICRA, 1997.

[2] S. Iqbal, *et al.*, “Actor-Attention-Critic for Multi-Agent Reinforcement Learning”, ICML, 2019.

人の順位付けによる深層学習モデルの視覚的説明の傾向調査

ER19090 **若山浩之**　　**指導教授：藤吉弘亘**

1. はじめに

深層学習モデルの推論結果に対する注視領域をアテンションマップとして可視化することで，モデルが入力画像のどこに注目して判断したかを人が理解できるようになる．このような視覚的説明性の良し悪しを表すアテンションマップの評価基準はないため，どのようなアテンションマップを獲得すると有効であるかは不明である．そこで本研究では，人の順位付けによるアテンションマップの評価法を提案し，様々な深層学習モデルのアテンション特性を調査する．

2. Grad-CAM

Grad-CAM [1] は，ネットワークの勾配を利用して出力クラスに対するアテンションマップを可視化する手法である．はじめに，画像をモデルに入力して計算した特徴マップから，クラス出力に対する勾配を計算する．次に，特徴マップの各チャネルごとに勾配の平均を求め，特徴マップに重み付けする．重み付けした特徴マップに ReLU 関数を適用し，その出力をアテンションマップとする．

3. アテンションマップの評価方法

提案するアテンションマップの評価方法を図 1 に示す．本研究では以下の手順により，人の判断根拠と類似しているアテンションマップを示すモデルの順位付けを行う．

Step 1: Grad-CAM でアテンションマップを可視化

Step 2: 被験者が分類時に注目すると考える順で順位付け

Step 3: 順位付け結果からスコアを集計

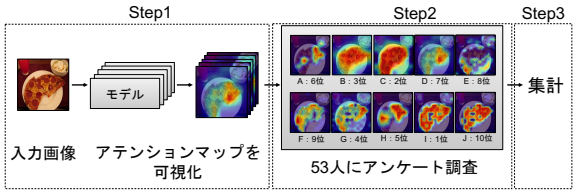


図 1: 評価方法

評価指標として，図 2 に示す順位スコアを用いる．被験者が順位付けしたアテンションマップの順位に対応した点数を小計スコアとして，被験者数分の小計スコアを加算し，合計スコアとする．合計スコアの順位を合計順位として，合計順位に対応した点数を順位スコアとする．順位スコアは合計スコアを正規化するため，合計スコアと比べて外れ値の影響を受けにくい．

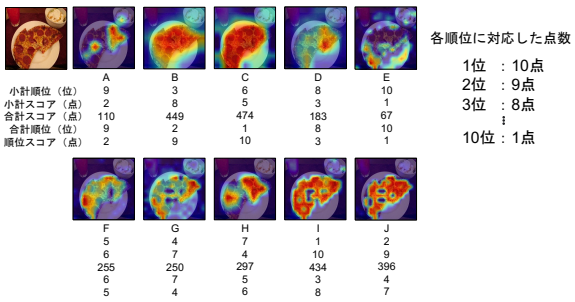


図 2: 順位スコア

4. 評価実験

本実験では，人の順位付けにより深層学習モデルのアテンションマップの傾向を調査する．

4.1. 実験条件

10 種の深層学習モデルのアテンションマップを Grad-CAM で可視化し，53 人の被験者にモデル情報を伏せて公平な評価の下，調査を行う．比較するモデルを表 1 に示す．SA は Self-Attention，Conv は畳み込み，精度は ImageNet-1k の評価示す．それぞれのモデルは ImageNet-1k で学習済みであり，パラメータ数が互いに近いモデルを用いて実験する．評価画像は，10 クラス 5 枚ずつの計 50 画像を使用する．

表 1: 比較するモデル

モデル	学習方法		特徴抽出		パラメータ数	精度 [%]
	教師あり学習	自己教師あり学習	Conv	SA		
ResNet	√	-	√	-	23M	75.8
DINOResNet	-	√	√	-	23M	75.3
ConvNeXT	√	-	√	-	28M	82.0
Swin	√	-	-	√	29M	81.1
CaiT	√	-	-	√	17M	80.7
DeiT	√	-	-	√	22M	79.9
DINODeiT	-	√	-	√	22M	77.0
CoAT	√	-	√	√	20M	81.9
XCiT	√	-	√	√	26M	81.8
DINOXCiT	-	√	√	√	26M	77.8

4.2. 実験結果

各評価画像で獲得した順位スコアを各クラス別に加算して比較する．評価結果例を表 2 に示す．全データは，50 枚の評価画像で獲得した順位スコアの合計である．合計データの括弧内に分散を示す．全データの結果から，XCiT が最もスコアが良いモデルとなることがわかる．次に CoAT，DINOResNet，ResNet と続いている．畳み込み構造の ResNet と Self-Attention 構造の DeiT を比較すると ResNet の方が，順位スコアの順位が高い．Swin と，ResNet の構造から Swin に近づけるよう設計した ConvNeXT を比較しても，畳み込みベースである ConvNeXT の順位が高い．1 位と 2 位である XCiT と CoAT は，畳み込み処理と Self-Attention の構造の両方を取り込んだモデル設計であり，最も人の注視領域と類似した傾向となる．

表 2: 評価結果 [位]

モデル	ピザ	電車	封筒	金魚	チワワ	アトリ	ホルン	全データ
ResNet	2	2	5	4	4	4	5	4 (5.64)
DINOResNet	5	2	5	3	2	2	5	3 (5.61)
ConvNeXT	7	7	7	5	8	3	4	7 (7.93)
CoAT	1	1	2	6	1	6	5	**2** (8.14)
Swin	9	8	10	9	6	8	9	10 (5.13)
DeiT	6	5	3	2	5	4	3	5 (4.9)
DINODeiT	10	10	8	8	10	10	8	10 (5.55)
CaiT	3	6	1	6	7	8	2	6 (7.36)
XCiT	4	4	3	1	3	1	1	**1** (3.72)
DINOXCiT	8	9	9	10	9	7	10	8 (6.29)

ラベルありで学習した XCiT とラベルなしの自己教師あり学習（DINO）で学習した DINOXCiT のアテンションマップの可視化例を図 3 に示す．括弧内に表 2 の全データの順位を示す．DINO を使用した場合，教師ありの XCiT と比較して，認識対象とアテンションの領域が類似していないことがわかる．また，DINOXCiT は XCiT と比べて順位が低いことから，人の注視領域と異なるといえる．

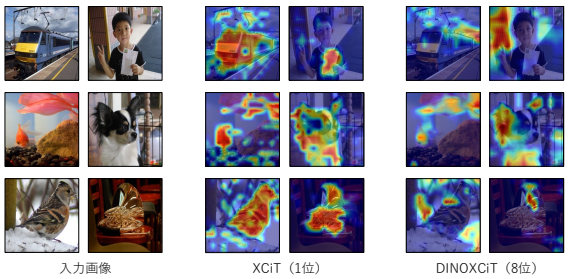


図 3: XCiT と DINOXCiT のアテンションマップ

5. おわりに

本研究では，深層学習モデルの認識特性の傾向調査した．実験結果から，人の注視領域と類似したモデルは XCiT，CoAT になることが分かった．今後は，アテンションの定量的な評価を行う．

参考文献

[1] R. Selvaraju, *et al.*,“Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”，ICCV，2019.

シングルモーダル学習モデルとマルチモーダル学習モデルにおける判断根拠の分析

ER19021 岡島滉樹　　指導教授：藤吉弘亘

1. はじめに

深層学習によるマルチモーダル学習モデルは，複数のモーダルを用いて学習し，分類する処理を行う．各モーダルの特徴を捉えることで，精度向上することが知られている．しかし，画像とテキストを用いたマルチモーダル学習モデルは，シングルモーダル学習モデルと比べて，テキストがどのように画像認識モデルに作用しているのか明らかではない．そこで本研究では，シングルモーダル学習モデルとマルチモーダル学習モデルにおける判断根拠を可視化し，モーダルを追加することによる判断根拠への影響について分析する．

2. マルチモーダル学習

モーダルとは，テキスト・音声・静止画・動画などの入力される情報のことである．シングルモーダルとは，画像，テキスト，音声などの 1 つの情報のことである．また，1 つの情報のみを用いて学習して，分類するモデルをシングルモーダル学習モデルと呼ぶ．一方，マルチモーダルとは，複数の情報のことであり，複数のモーダルを用いて学習して，分類するモデルをマルチモーダル学習モデルと呼ぶ．マルチモーダル学習モデルは，複数の情報を考慮して問題を解くことができるため，シングルモーダル学習モデルと比べて高い認識性能を発揮することが知られている．代表的な手法の 1 つに Gallo らの Early fusion モデル [1] がある．Early fusion モデルのネットワーク構造を図 1 に示す．Early fusion モデルは，画像とテキストを入力データとし，各モーダル用のモデルで抽出した特徴量を連結して分類器に入力することで，最終予測を出力する．

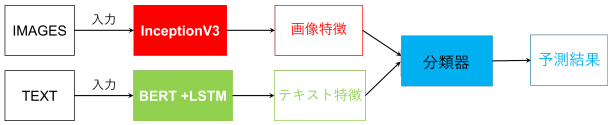


図 1：Early fusion のネットワーク構造

3. マルチモーダル学習モデルの判断根拠の分析

Early fusion モデルは，テキストのモーダル情報がどのように画像認識モデルに作用しているのか明らかではない．そこで，本研究ではシングルモーダル学習モデルとマルチモーダル学習モデルにおける判断根拠の可視化に Grad-CAM [2] を用いて分析する．

3.1. 実験概要

画像とテキストを用いる Early fusion モデルと画像のみを用いる InceptionV3 の判断根拠がどのように変化するかを調査する．また，Early fusion モデルのテキストを変更した場合に判断根拠へどのような影響を与えるのかを比較する．InceptionV3 の学習回数は 10 エポック，バッチサイズは 75，学習率 0.01 である．Early fusion の学習回数は 15 エポック，バッチサイズは 8，学習率は 0.01 である．データセットには UPMC Food-101 を使用する．UPMC Food-101 [3] は 101 クラスの食べ物の分類を行うデータセットであり，画像とテキストをペアで構成される．

3.2. 実験結果と可視化の比較

判断根拠の可視化に使用する各モデルの精度を表 1 に示す．画像とテキストの情報を用いて学習することで，画像のみを用いた場合と比べて，精度が 19.23 ポイント向上した．

表 1：各モーダルモデルの分類精度

手法	画像	テキスト	精度 [%]
Inception V3	✓		70.67
Early Fusion	✓	✓	89.90

InceptionV3 と Early fusion のアテンションマップの可視化結果を表 2 に示す．アテンションマップは，色が赤色に近いほどその領域を注視していることを表す．Early fusion モデルと比べて，分類精度が低い InceptionV3 は画像の中心領域を注視する傾向がある．多くの画像で画像の中央に食べ物があるため，画像のみを使用して学習すると注視領域が中央に偏ったと考えられる．一方で，画像とテキストを使用する Early fusion モデルでは物体を注視した判断根拠が出力されている．このことから，テキスト情報を追加することで，画像とテキストの内容を紐づけて学習し，テキストの内容に応じた箇所に注視したアテンションマップを獲得できるようになったと考えられる．

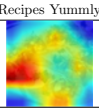
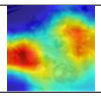
表 2：アテンションマップの可視化結果

クラス名	shrimp and grits	french onion soup	chocolate mousse	tuna tartare	deviled eggs
入力画像					
入力テキスト	Wild Patagonian Shrimp and Grits Cooking and Recipes	french onion soup In Jennie's Kitchen	Christmas Dessert Recipes: Chocolate Mousse Pie CHOW	Ahi Ahi Tuna Tartare Recipes Yummly	Mexican Deviled Eggs Recipe MyRecipes.com
InceptionV3					
Early fusion					

3.3. テキスト入力の変化による可視化の比較

InceptionV3 と Early fusion モデルでアテンションマップが大きく異なったデータについて，テキストを変更してモデルへ入力した場合のアテンションマップの可視化結果を表 3 に示す．これより，テキストの変更内容に応じてアテンションマップが変化することが確認できた．色に関するテキストに変更した場合，より広い領域を注視するアテンションマップとなることが確認できた．細かい位置を指定するテキスト変更した場合，色に関するテキストに変更したときに比べ，アテンションマップがわずかに変化することが確認できた．この結果から Early fusion モデルは，位置を指定したテキスト情報に比べ，色に関するテキスト情報を重要視するモデルであると言える．

表 3：テキスト変更時のアテンションマップ

	変更前	変更後：色に着目	変更後：位置に着目
入力テキスト	Ahi Ahi Tuna Tartare Recipes Yummly	red tomato ketchup	top left white plate
attention map			

4. おわりに

本稿では，シングルモーダルで学習したモデルとマルチモーダルで学習したモデルのアテンションマップを比較した．画像のみを用いた学習と比べて，画像とテキストを用いた学習は，テキストに応じたアンテンションマップを獲得できることを確認した．今後は，シングルモーダル学習モデルとマルチモーダル学習モデルの種類を増やし，異なるモデル構造においても同様の傾向にあるのか調査を行う．

参考文献

[1] I.Gallo, *et. al.*, “Image and Text fusion for UPMC Food-101 using BERT and CNNs”, IVCVZ, 2020.

[2] Ramprasaath R. Selvaraju, *et. al.*, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, ICCV, 2019.

[3] X.Wang, *et. al.*, “Recipe recognition with large multimodal food dataset”, ICMEW, 2015.

人の知見を導入した ST-ABN による動画像認識の高精度化

ER19062 野口紗季　　指導教授：藤吉弘亘

1. はじめに

動画像認識は，複数フレームの映像からその動画で行われる動作などを認識するタスクである．これまで静止画像による物体認識では，認識時における判断根拠の可視化や，人の知見の導入の有効性について示されている．本研究では，複数フレームの映像を対象とする動画像認識に対してネットワークの注視領域に人の知見を導入することで，より適切な注視領域の獲得と認識精度の向上を目指す．

2. ST-ABN

Spatio-Temporal Attention Branch Network (ST-ABN) [1] は空間情報と時間情報を同時に考慮した，動画像認識手法である．ST-ABN では，空間情報における重要度を示す Spatial attention と時間情報における重要度を示す Temporal attention を特徴マップに重み付けすることで認識精度の高精度化を実現している．ST-ABN は入力動画から特徴マップを獲得する Feature extractor, ネットワークの注視領域を獲得する ST attention branch, ST attention branch で獲得した注視領域を特徴マップに重み付けする Attention 機構，クラス確率を出力する Perception branch で構成される．

3. 提案手法

ST-ABN は，動画像の認識に有効な Spatial attention と Temporal attention を獲得できない場合，誤認識を誘発する．本研究では，動画像を対象とした認識に有効な注視領域の獲得と認識精度の向上を目的とし，人の知見を導入する手法を提案する．提案手法の流れを以下に示す．

Step1　ST-ABN の学習済みモデルを使用し，評価時に誤認識した学習サンプルの Temporal attention を収集する．

Step2　収集した Temporal attention を人手によって修正する．Temporal attention の修正例を図 1 に示す．ここで，Temporal attention は各フレーム画像の上部にあるカラーバーとして表す．修正時には，認識に不要なフレームに青色を，認識に必要な動きのあるフレームには緑色を，認識時に特に重要なフレームには赤色を付与する．

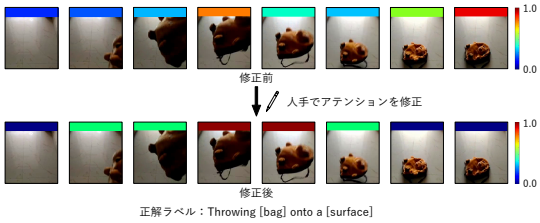


図 1：Temporal attention 修正例

Step3　修正したアテンションを真値として ST-ABN の再学習を行うことで人の知見を導入する．再学習する際のネットワーク構造を図 2 に示す．再学習時には，式 (1) に示すように，ST-ABN の学習誤差 $L = L_{per} + L_{att}$ に L_{temp} を追加する．

$$L = L_{per} + L_{att} + L_{temp} \tag{1}$$

L_{temp} はネットワークから出力される Temporal attention M_t と，修正した Temporal attention M'_t の平均二乗誤差から算出する．ここで，n は入力フレーム数を示している．また，γ_t は学習誤差 L_{temp} を調整する係数である．

$$L_{temp} = \gamma_t \frac{1}{n} \sum_{i=1}^{n} (M'_{t_i} - M_{t_i})^2 \tag{2}$$

4. 評価実験

提案手法の有効性を検証するために評価実験を行う．

4.1. 実験条件

本実験では Something-Something v.2 データセットの誤認識動画に対してアテンションを修正し，実験を行う．

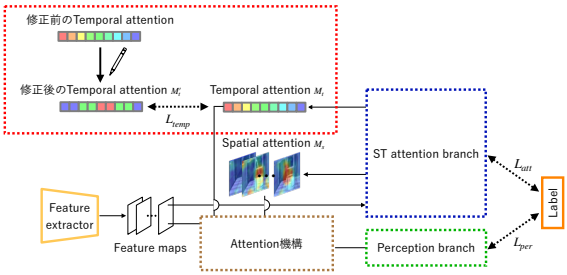


図 2：人の知見を導入した ST-ABN

Temporal attention の修正は全 174 クラスのうち，学習用データと評価用データにおける認識精度が低い 8 クラスの誤認識動画を対象とする．ベースネットワークとして 3D ResNet-50 を使用し，入力フレーム数は 32 フレームとする．また，学習誤差 L_{temp} 調整する係数 γ_t は 10 とする．

4.2. 実験結果

再学習前後の認識精度の比較を表 1 に示す．従来の ST-ABN に提案手法を取り入れることで精度が向上していることが確認できる．また，Temporal attention を修正したクラスの方が修正していないクラスよりも精度が向上していることから，ST-ABN に人の知見を導入することで，認識により有効な注視領域が獲得できるようになったと考えられる．

表 1：ST-ABN の再学習による認識精度の比較 [%]

	全クラス	修正対象クラス	その他クラス
再学習前	58.62	20.50	59.76
再学習後	**60.65**	**26.32**	**61.68**

4.3. アテンションマップの比較結果

ST-ABN の再学習前後のアテンションマップの可視化結果を図 3 に示す．空間情報の重要度を表す Spatial attention は各フレーム画像に重ねたヒートマップ，時間情報の重要度を表す Temporal attention はヒートマップの上部にあるカラーバーとして可視化している．再学習前は動きの有無に関わらず隣接するフレームのカラーバーの色変化が大きいのに対し，再学習後は徐々に色が変化しており，より適切なアテンションの獲得ができたといえる．また，Spatial attention は認識に不要な注視領域が減少していることから，Temporal attention の修正により Spatial attention も改善されると考えられる．

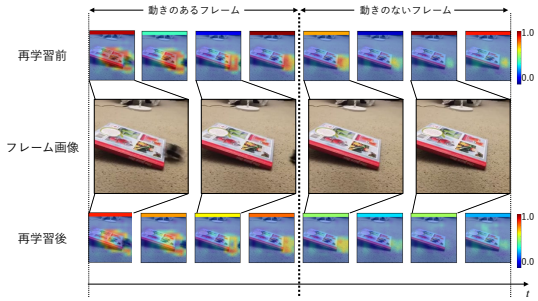


図 3：アテンションマップの変化

5. おわりに

本研究では，人の知見を導入した Temporal attention を用いた ST-ABN の再学習法を提案し，より適切な注視領域の獲得と ST-ABN 認識精度の向上が確認できた．今後は Spatial attention に人の知見を導入することで ST-ABN のさらなる高精度化を目指す．

参考文献

[1] 三津原ら，“Spatio-Temporal Attention Branch Network による時空間情報を考慮した視覚的説明”, MIRU, 2021.

ノイズに対する頑健性と視覚的説明性を獲得する Attention Branch Transformer の提案

ER19043　鈴木雅司　　指導教授：藤吉弘亘

1. はじめに

Attention Branch Network (ABN) [1] は，Attention 機構による精度の向上とアテンションマップの可視化による視覚的説明を両立する手法である．ABN のベースモデルである畳み込みニューラルネットワーク (CNN) は，局所的な帰納的バイアスを持つため物体のテクスチャを重視して学習する．そのため，画像の意味的な構造やスタイルの変化等に対応できない．これらの問題を解決するために，本研究では局所的な帰納的バイアスの影響を受けにくいモデルである Vision Transformer (ViT) に Attention Branch を導入した Attention Branch Transformer (ABT) を提案する．

2. Vision Transformer

ViT は自然言語処理の Transformer を画像処理に応用した手法である．ViT は画像を均等なパッチに分割し，埋め込み処理で各パッチをベクトルの系列である Patch Token に変換する．そして，学習可能な Token である Class Token を追加し，これらの Token を Transformer Encoder (TE) へ入力する．TE では Multi-Head Attention で全ての Token 間の関係性を捉え，Token 間で相互に値を更新し特徴抽出を行う．ViT は Token 間の広域的な特徴を捉えるため，CNN と比べ局所的な帰納的バイアスの影響が少ない．そのため，ViT は画像の意味的な構造やスタイルの変化およびテクスチャノイズ等のノイズに対して頑健となる．

3. 提案手法

本研究では，より頑健であり説明性の高い説明可能 AI (XAI) の実現を目指し，ViT にアテンションマップを出力する Attention Branch を導入した手法 ABT を提案する．図 1 に提案手法のネットワーク構造を示す．提案手法の TE では Patch Token のみで特徴抽出を行い，各 Branch の最終層で Class Attention (CA) および Class Token を用いて推論を行う．提案手法の損失関数 $L(\mathbf{x}_i)$ として，式 (1) のように各 Branch の学習損失の和を用いる．ここで，L_{att}, L_{per} は各 Branch の学習損失，x_i は入力サンプルを表す．また，各 Branch の学習損失は，各 Branch より出力されるクラス確率と教師ラベルとのクロスエントロピー誤差として算出される．

$$L(\mathbf{x}_i) = L_{att}(\mathbf{x}_i) + L_{per}(\mathbf{x}_i) \tag{1}$$

Attention Branch では，式 (2) に示すように CA に基づきアテンションマップを生成する．ここで，$\mathbf{Q}_a, \mathbf{K}_a$ は Attention Branch における CA の Query，Key，d, h は各パッチの埋め込み次元数および head 数，$\mathbf{M}_{att}$ はアテンションマップを表す．

$$\mathbf{M}_{att} = \mathrm{Sigmoid}(\mathbf{Q}_a\mathbf{K}_a^\top/\sqrt{d/h}) \tag{2}$$

また，Perception Branch では，式 (3) に示すように 1 層目の TE の Attention weight とアテンションマップの要素積をとることで特徴マップの強調を行う．ここで，$\mathbf{A}, \mathbf{Q}_p, \mathbf{K}_p$ は Perception Branch における 1 層目の Multi-Head Self Attention の Attention weight および Query，Key を表す．

$$\mathbf{A} = \mathrm{Softmax}(\mathbf{Q}_p\mathbf{K}_p^\top/\sqrt{d/h})\mathbf{M}_{att} \tag{3}$$

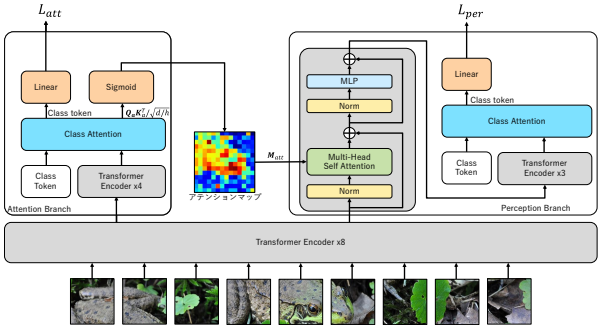


図 1: 提案手法のネットワーク構造

4. 評価実験

本章では，提案手法と ABN，ViT の精度，耐ノイズ性およびアテンションマップによる視覚的説明性の比較結果について述べる．

4.1. 実験条件

学習用データセットには ImageNet を用いる．評価用データセットには，ImageNet と耐ノイズ性の評価に ImageNet-A，ImageNet-R，ImageNet-C を用いる．ここで，ImageNet-A は ImageNet から誤識別の多い画像，ImageNet-R は創作物の画像を集めたデータセットであり，ImageNet-C は ImageNet に 19 種類のノイズを付加したデータセットである．ImageNet-A および ImageNet-R では Top 1 accuracy，ImageNet-C では各ノイズに対する Corruption Error (CE) の平均である mean Corruption Error (mCE) で評価する．ABT-B は埋め込み次元数を 786，head 数は 12 であり，ABT-S と ABT-T は，埋め込み次元数と head 数をそれぞれ 1/2,1/4 に変更したモデルとする．

4.2. 実験結果

提案手法と ABN，ViT の精度と頑健性の比較結果を表 1 に示す．表 1 より，精度，頑健性共に ABT-B が最も優れている事が確認できる．

表 1: 提案手法と従来手法の精度と頑健性の比較

Model	Param	INet Top-1 ↑	INet-A Top-1 ↑	INet-R Top-1 ↑	INet-C mCE ↓
ABN-34	36.4 M	71.4	3.3	34.2	77.9
ABN-50	43.6 M	79.9	11.3	41.2	65.4
ABN-101	62.6 M	81.5	18.4	44.3	58.8
ViT-T	5.7 M	72.2	7.0	33.2	71.3
ViT-S	22.1 M	79.9	19.2	42.5	55.7
ViT-B	86.6 M	82.0	27.9	45.3	49.4
ABT-T	8.6 M	71.3	7.5	32.6	71.1
ABT-S	33.1 M	80.1	23.6	43.5	53.6
ABT-B	129.9 M	**82.8**	**32.5**	**46.2**	**47.9**

ABN と提案手法の出力したアテンションマップを図 2 に示す．図 2 は左から入力，ABN および提案手法の Attention Branch から出力したアテンションマップである．図 2 より，提案手法のアテンションマップは推定物体の形状に沿った Attention が得られている事が確認できる．この結果より，提案手法は視覚的説明において形状の表現に有効であると確認できる．

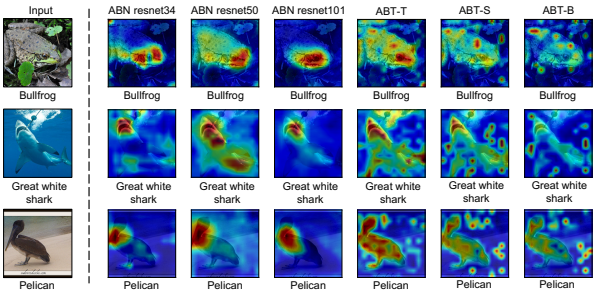


図 2: アテンションマップの可視化

5. おわりに

本研究では，より頑健であり説明性の高い XAI の実現を目指し，ViT に Attention Branch を導入した手法 ABT を提案した．本研究では ABT が精度，耐ノイズ性，視覚的説明性に有効であることを確認した．今後は人から AI への働きかけが可能となるよう，提案手法へ人の知見を導入する手法を検討する．

参考文献

[1] H. Fukui, *et al.*, “Attention Branch network: Learning of attention mechanism for visual explanation”, CVPR, 2019.

Wavelet Masked Autoencoder による Wavelet Transformer の精度向上

ER19069 福井 雅弥　　指導教授：藤吉 弘亘

1. はじめに

Wavelet Transformer [1] は，Wavelet 変換した画像を Vision Transformer (ViT) に入力する手法であり，明示的に物体の高域成分や低域成分，スケールの詳細な補助情報との関係性を考慮した特徴量の抽出を可能とした．ViT と比べて視覚的説明性が向上する一方で，精度が低下するという問題がある．そこで本研究では，Wavelet Transformer の事前学習に Masked Autoencoder (MAE) [2] を導入した Wavelet Masked Autoencoder (Wavelet MAE) を提案する．MAE は入力画像の一部にマスクを行い，マスクした領域のピクセルを予測することで学習を行う．そのため，Wavelet MAE は成分間やスケール間の対応関係を捉える事前学習が可能となる．

2. Wavelet Transformer

Wavelet Transformer は，Wavelet 変換により入力情報を予め人間が解釈可能な単位へ分割して ViT に入力する．このとき，ViT に入力する各パッチは Wavelet 変換によってスケール情報やテクスチャ情報に対応しているため，Self-Attention 機構の Attention から，スケールやテクスチャなどの詳細な情報と判断根拠の関係を人間へフィードバックできる．

3. 提案手法

提案する Wavelet MAE のネットワーク構造を図 1 に示す．Wavelet MAE は，入力画像に対して Wavelet 変換を低域成分に 4 回再帰的に適用する．Wavelet 変換後の画像をパッチ分割し，ランダムにマスク処理を適用する．マスクしたパッチのピクセルを予測することを pre-text タスクとして自己教師あり学習を行う．

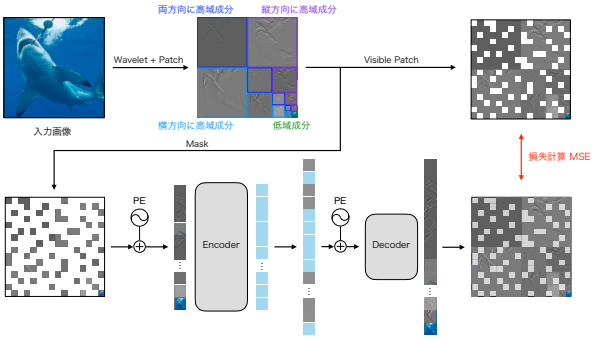


図 1：Wavelet MAE のネットワーク構造

Position Embedding は sin-cos 関数による固定のパラメータで位置情報を表現する．式 (1) に Position Embedding の計算式を示す．

$$
\begin{aligned}
PE_{(pos,2i)} &= \sin(pos/10000^{4i/D}) \\
PE_{(pos,2i+1)} &= \cos(pos/10000^{4i/D})
\end{aligned}
\tag{1}
$$

Encoder はマスクされていないパッチから特徴を抽出し，Decoder は抽出した特徴量とマスクしたパッチを表すマスクトークンから各パッチのピクセルを出力する．損失関数は，マスクトークンの出力と Wavelet 変換後の画像の平均二乗誤差である．Wavelet MAE で事前学習した重みを Wavelet Transformer の初期値として使用し，ファインチューニングを行う．ファインチューニングでは Decoder 部分を破棄し，Encoder のみを利用する．

4. 評価実験

事前学習において，マスク処理は画像全体の 75%をマスクする．事前学習とファインチューニングのデータセットには ImageNet-1k を使用する． Wavelet MAE のネットワーク構造は，Encoder に ViT-L/14，Decoder に Transformer block を 8 層，埋め込み次元数を 512 次元とした小規模な構造を使用する．

4.1 精度比較

学習結果を表 1 に示す．Wavelet Transformer の精度は 72.2%であるのに対し，提案手法では 79.5%と 7.3pt 精度向上した．

表 1：従来手法との精度比較（*は [2] より引用）

手法	事前学習	精度
ViT	-	82.5*
	✓	84.9*
Wavelet Transformer	-	72.2
	✓	79.5

図 2 に，Wavelet MAE の事前学習により得られた再構成画像を示す．ここでは Wavelet 変換後の画像に対して 75%の確率でランダムにマスク処理を適用する．図 2 より，Wavelet 変換した画像の各成分をみると，スケールやテクスチャ，色情報をある程度復元可能であることが確認できる．

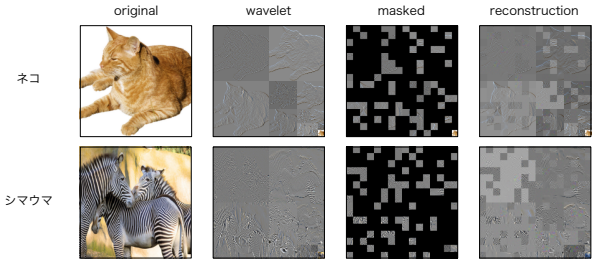


図 2：Wavelet MAE で事前学習した際の再構成画像

4.2 アテンションを用いた逆 Wavelet 変換

Wavelet Transformer から得られたアテンションを Wavelet 変換後のパッチ特徴に重み付けし，逆 Wavelet 変換を施すことで，Wavelet MAE がどのパッチ情報を重要としているかを確認する．図 3 にアテンションを用いた逆 Wavelet 変換画像を示す．アテンションの可視化には Attention Rollout を用いる．図 3 より，低域成分にアテンションが高くなる傾向にあることが分かる．これは，ImageNet で学習した場合，特定のパッチ情報を重視するような特徴量の捉え方をしていると考えられる．

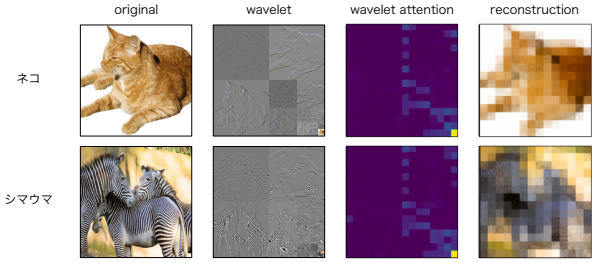


図 3：Attention を用いた逆 Wavelet 変換

5. おわりに

本研究では，Wavelet 変換による成分間やスケール間の対応関係を獲得する事前学習法である Wavelet MAE を提案した．提案手法では，Wavelet Transformer と比べて 7.3pt 精度が向上した．一方で，アテンションマップは偏りのある関係性となった．今後は，アテンションマップの改善を行う．

参考文献

[1] T. Ogata, *et al.*, “Wavelet Transformer: Exploit Interpretable Attentions by Splitting Input Component into Human-Explainable Information”, In MIRU, 2022.

[2] H. Kaiming, *et al.*, “Masked Autoencoders Are Scalable Vision Learners”, In CVPR, 2022.

SimSiam におけるデータ拡張の効果と Mixup の導入による向上

ER19009 岩垣渚　　指導教授：藤吉弘亘

1. はじめに

自己教師あり学習は，pretext タスクを設定し入力データに疑似ラベルを用いることで，正解ラベルを必要としない学習方法である．代表的な手法である SimSiam[1] は，データ拡張として，入力画像から異なる 2 枚の変換画像を作成し，各画像から抽出した特徴量の類似度が高くなるように学習する．データ拡張にクロップと色変換が採用されており，その他のデータ拡張手法を使用した場合の学習への影響は深く調査されていない．そこで，本研究では SimSiam におけるデータ拡張の影響について分析を行い，適切なデータ拡張による SimSiam の性能改善を目指す．

2. SimSiam

SimSiamは, 1 つの入力データからデータ拡張により 2 つのデータを作成し, 2 つのデータに対する特徴量が一致するように学習を行う．作成した 2 つのデータのうちの 1 つから，対となるもう 1 つのデータの特徴量をMLPにより予測することで 2 つのデータの類似度を損失関数により計算する．損失関数を式 (1) に示す．

$$\mathcal{L}_{siam} = \frac{1}{2}\mathcal{D}(p_1, stopgrad(z_2)) + \frac{1}{2}\mathcal{D}(p_2, stopgrad(z_1)) \tag{1}$$

ここで $\mathcal{D}$ は負のコサイン類似度, z はネットワークが抽出した特徴量, p は特徴量 z を入力とするMLPが出力した特徴量である．$stopgrad$ は勾配の停止を表し，予測対象 z を予測出力 p に近づける学習を行わないようにしている．

3. データ拡張の組み合わせによる精度変化

SimSiam で使用されているデータ拡張の組み合わせによる影響については深く考慮されていない．そこで，SimSiam におけるデータ拡張の組み合わせによる精度への影響を調査する．ネットワークには ResNet-50，データセットには CIFAR-100 を使用する．データ拡張には Crop, Cutout, Color distort, Sobel filtering, Gaussian noise, Gaussian blur, Rotate を用いる．

2 つのデータ拡張を組み合わせた時の評価結果を図 1 に示す．縦軸と横軸は適用したデータ拡張手法，各値は認識精度を表し，色は明るい程精度が高いことを示す．切り抜いた画像を作成する Crop と色の変換を行う Color を組み合わせた場合が最も高い精度となった．一方で Crop を用いない場合に精度が低下する傾向にある．以上より，SimSiam において. 最適なデータ拡張の組み合わせは Crop と Color であることが分かった．

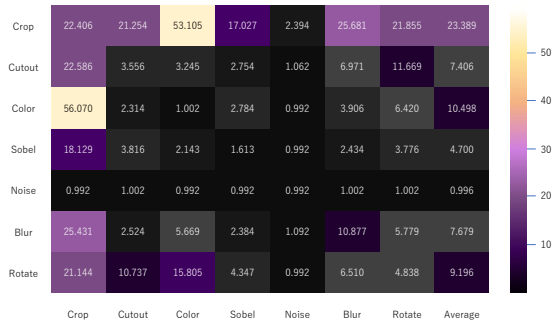


図 1：SimSiam におけるデータ拡張手法の精度表

4. Mixup タスクを導入した SimSiam

現在 SimSiam に用いられている Color は Crop と元画を共有しているため，変換後の色が限定的である．そこで従来の SimSiam に mixup を用いたタスクを導入した手法を提案する．mixup タスクでは，元画像が異なる画像を mixup して入力することで，色変換を多様化し，より頑健な特徴の抽出を行う．また，mixup の混合比率によって異なる画像の特徴を近づけることのないように，動的に学習を切り替える．

4.1. mixup を用いた動的な自己教師あり学習

動的な自己教師あり学習を導入した SimSiam を図 2 に示す．データ拡張として異なる画像を合成する mixup[2] により色情報を大きく変更する．mixup によるデータ拡張を式 (2) に示す．

$$Mix(x, y) = \lambda x + (1 - \lambda)y \tag{2}$$

ここで, x, y はミニバッチ内の画像,λはmixupの混合比率である.λの値が大きい場合，データ拡張前の情報をある程度残したデータが作成されるため，近づける学習を行う．一方でλの値が小さい場合，データ拡張前のデータから大きく異なるデータが作成されるため，離す学習を行う.mixup を適用したデータに対する損失関数を式 (3) に示す．

$$\mathcal{L}_{mix} = \begin{cases} \mathcal{D}(p_m, stopgrad(z)) & \lambda > 0.5 \\ -\mathcal{D}(p_m, stopgrad(z)) & otherwise \end{cases} \tag{3}$$

ここで, $\mathcal{D}$ は負のコサイン類似度関数, z はネットワークが抽出した特徴量, p_mは特徴量 z_mを入力とするMLPが出力した特徴量である．最終的な損失関数は，従来のSimSiam の損失 $\mathcal{L}_{siam}$である式 (1) とmixupによる損失 $\mathcal{L}_{mix}$である式 (3) を足し合わせたものとする．

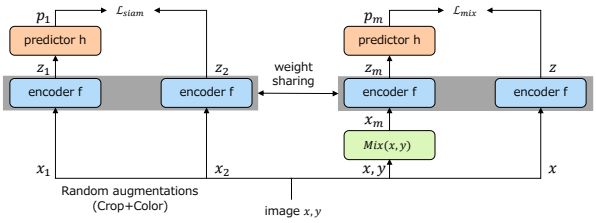


図 2：mixup タスクを導入した SimSiam の学習方法

4.2. 評価実験

提案手法と SimSiam の精度を比較する．ネットワークとして ResNet-50，データセットとして CIFAR-100 を使用し，200 エポックで学習する．評価方法は k-nn 法を用いる．評価結果を表 1 に示す．SimSiam に mixup を用いた動的な学習を追加することで従来の SimSiam に比べて精度が約 1pt 向上した．また，x に対してデータ拡張を適用し，0.3 から 0.7 の乱数で混合した画像を入力とすることで最も高い精度を獲得した．

表 1：従来法との精度比較

	mixup タスク	混合する画像	x に対して適用するデータ拡張	画像の混合比率	Top1 Accuracy[%]
SimSiam	-	-	Crop+Color	-	56.07
提案手法	✓	x, y	-	$0.4 \leq \lambda \leq 0.6$	57.02
	✓	x_1, y	Crop+Color	$0.3 \leq \lambda \leq 0.7$	**57.87**

5. おわりに

本研究では SimSiam におけるデータ拡張による精度変化について調査を行い，調査結果に基づいた mixup の導入を行った．提案手法により従来の SimSiam と比べて精度が向上することを示した．今後は，SimSiam のデータ拡張による更なる精度向上に取り組む．

参考文献

[1] X.Chen, *et al.*, “Exploring Simple Siamese Representation Learning”, CVPR, 2021.

[2] H.Zhang, *et al.*, “mixup: BEYOND EMPIRICAL RISK MINIMIZATION”, ICLR, 2018.

深層学習モデルによる視線予測のための教師データの作成法とその効果

ER19015 梅谷将成　　指導教授：藤吉弘亘

1. はじめに

視線予測は人間の視覚注意システムを機械学習により模倣する試みであり，スポーツの競技力向上や自動車運転時の安全性向上などへの応用が期待されている．特にスポーツでは，審判や指導者が競技中の注視している領域を選手や育成中の審判，指導者にフィードバックすることは重要である．しかしながら，審判や指導者の視線を常時計測することは困難である．

そこで本研究では審判や指導者の視線予測モデルを構築する，この時，教師となる視線範囲を変更した際の精度比較と，時系列情報を考慮した場合の傾向調査を行う．

2. 深層学習モデルによる視線予測モデル

2.1. DINet

静止画から視線予測を行うモデルとして，Dailated Inception Network（DINet）[2] がある．DINet は，特徴抽出モデルである Dilated Residual Network（DRN）に Dailated Inception Module (DIM) を追加したモデルである．DIM では，異なるカーネルサイズで畳み込み処理を行い，拡張率を変更することでカーネルサイズを調整することができる．DINet のネットワーク構造を図 1 に示す．

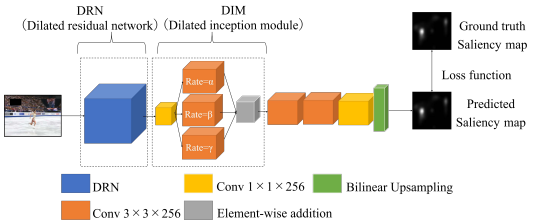


図 1：DINet の構造

2.2. ViNet

時系列情報を考慮した視線予測を行うモデルとして，ViNet[1] がある．ViNet は S3D ブロックにて 4 つの異なるカーネルサイズで 3 次元畳み込み処理を行う．それぞれの畳み込み層から予測ネットワークに直接情報を渡すことで，異なる時間的，空間的特徴量を抽出して推論を行う．ViNet のネットワーク構造を図 2 に示す．

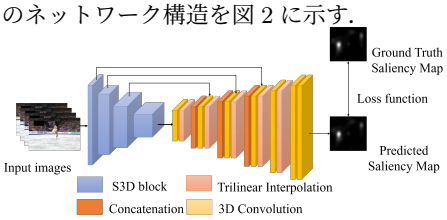


図 2：ViNet の構造

3. 提案手法

本研究では，視線予測モデルの学習に用いる教師データのプロットサイズが精度に与える影響を調査する．視線の教師マップの作成方法を図 3 に示す．視線予測モデルの学習には，入力画像に対して視線となる位置を 1，それ以外を 0 としたマップを正解として用いる．この時，予測位置に対してガウシアン処理し，正解との位置のずれを考慮する．真値となる領域が少ないと顕著性が小さくなるため，30px まで拡大した視線を教師データとする．

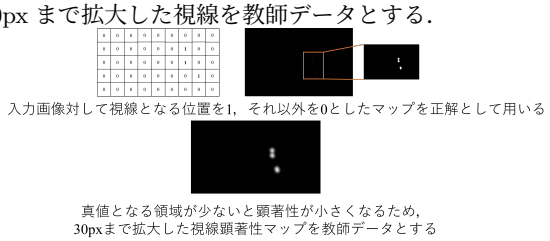


図 3：教師データの作成

4. 評価実験

提案手法を評価するために 2 つの実験を行う．1 つ目は，使用する教師データのプロットサイズを変更した際の精度比較を行う．サイズは視線座標を中心に 20，30，40px とする．2 つ目は，ViNet と DINet の予測精度を比較し，時系列情報が与える影響の調査を行う．なお，比較指標には，Similarity Metric（SIM）と Linear Correlation Coefficient（CC）を使用する．各モデルの学習条件は，エポック数 1000，学習率 0.04，バッチサイズ 8 で統一し，最適化手法には DINet に Adam，ViNet に SGD を使用する．

4.1. 使用するデータセット

フィギュアスケートのジャンプ動作を撮影した 30 本の動画像を用いる．フレームレートは，29.97FPS で 1 つのジャンプは 240 フレームで構成される．視線データは 60Hz である．また，視線データは審判群 3 名と選手群 3 名から取得されている．本実験では，27 本を学習用に使用し，3 本をテスト用データとする．

4.2. 実験結果

教師データのプロットサイズを変更した ViNet での結果を表 1 に示す．表 1 からプロットサイズを 30px とした時に最も高い精度であった．

表 1：プロットサイズを変更した際の定量的評価

サイズ [px]	20	30	40
SIM	-0.0246	0.0045	0.0027
CC	0.0059	0.0143	0.0108

次に，ViNet と DINet の精度を比較した結果を表 2 に示す．なお，プロットサイズは 30px で統一した．表 2 から ViNet の方が精度が高く，時系列情報を用いることは有効であることが分かった．

表 2：ViNet と DINet の精度比較

モデル	ViNet	DINet
SIM	0.0027	-0.2735
CC	0.0108	-0.0295

定性的評価の結果を図 3 に示す．プロットサイズを 20px とすると注視領域が小さいことが分かる．一方プロットサイズを 40px まで大きくすると，本来の注視点から離れた範囲を注視領域とするため，精度が低下した．その中間である 30px が最もよい精度となることが分かった．

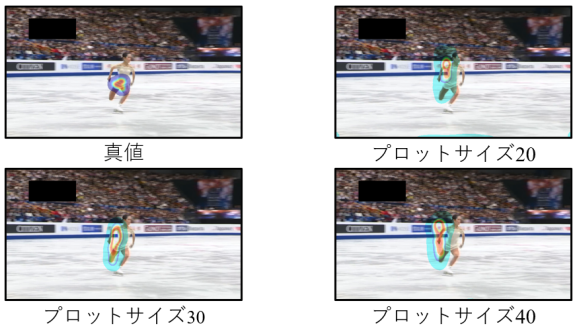


図 4：定性的評価

5. おわりに

本研究では，プロットサイズを変更した際の精度比較と時系列情報を考慮したモデルでの比較を行った．評価実験では，プロットサイズ 30px とした時に最もよい精度が得られ，時系列情報の有効性を確認した．今後はモデルのパラメータ調整を行い，競技理解が深い人の視線予測精度向上を目指す．

参考文献

[1] S. Jain, *et al.*, “ViNet: Pushing the limits of Visual Modality for Audio-Visual Saliency Prediction”, IEEE, pp.3520-3527, 2021.

[2] Y. Yu, *et al.*, “A review of recurrent neural networks: Lstm cells and network architectures”, Neural Computation, pp.1235-2019, 2019.

Transformer の Attention weight を用いた成績予測の傾向調査と改善

ER19066 伴優輝　　指導教授：藤吉弘亘

1. はじめに

教育環境のデジタル化に伴い，学生の学習行動データの大規模な収集が可能となった．これらの学習行動を分析し，成績が良い学習パターンとの相違を発見することで成績不良者の学習行動の改善につながることが期待されている．文献 [1] では，Attention ベースの機械学習モデルである Transformer[2] を用いて成績予測を行い，Attention の可視化により重要な講義や操作を分析している．そこで本研究では，Transformer による成績予測，重要な教材の特定に加え，学習行動の改善を図る．具体的には学習した Self-Attention の Attention weight を解析することで，学習者の行動パターンの傾向を捉え，学習者にフィードバックすることを目指す．

2. 提案手法

本研究では，分類型の Transformer による成績予測を行う．またデータ拡張を行い精度向上を図る．さらに，Attention weight の可視化により，成績ごとの傾向調査を行い，傾向の違いから成績不振の生徒の改善を図る．

2.1. データセットと特徴量作成

実験には，九州大学で行われた講義を受講した学生の学習行動データを用いる．学習行動データとは，電子教材を閲覧した際の操作ログである．操作ログとして収集できる情報は，教材を開く，閉じる，ページ送り，戻り，メモ，マーカーなど 21 種類あり，操作した時間も含まれる．また，各学生の成績も含まれており，成績は A から F の 5 段階である．本実験では，2020 年から 2022 年の 3 年分，165 人の学習行動データを使用して成績予測を行う．入力する特徴量には 12 回の講義資料それぞれに対する各操作の頻度をベクトルとして用いる．

2.2. データ拡張

2.1 節で述べたように，データセットの人数が 165 人とデータ数が少ない．データ数が少ないと汎化性能が低くなるため，データ拡張を行い汎化性能の向上を図る．

データ拡張には Mixup や CutMix を使用する．Mixup の新しいサンプルは式 (1) で算出する．ここで，$\boldsymbol{X}$，$\boldsymbol{y}$ は入力と正解ラベル，λ は混合比である．混合比 λ はベータ分布を用いてランダムに決定する．

$$\boldsymbol{X} = \lambda \boldsymbol{X_1} + (1-\lambda)\boldsymbol{X_2}, \boldsymbol{y} = \lambda \boldsymbol{y_1} + (1-\lambda)\boldsymbol{y_2} \quad (1)$$

また，CutMix は式 (2) で算出する．ここで，$\boldsymbol{M}$ はバイナリ行列である．

$$\boldsymbol{X} = \boldsymbol{M} \odot \boldsymbol{X_1} + (1-\boldsymbol{M}) \odot \boldsymbol{X_2}, \boldsymbol{y} = \lambda \boldsymbol{y_1} + (1-\lambda)\boldsymbol{y_2} \quad (2)$$

2.3. Attention weight の可視化

成績を正しく予測できた学生について，成績ごとの Attention weight の平均を (3) 式で算出する．ここで，Q，K は Self-Attention の Query と Key，d_k はその次元数であり，n は分類に成功したそれぞれの成績の人数である．

$$Attention(Q, K) = \frac{1}{n}\sum_{i}^{n} softmax(\frac{Q_i K_i^T}{\sqrt{d_k}}) \quad (3)$$

3. 精度比較実験

本実験では，Transformer のエンコーダを用いる．ヘッド数 1，埋め込み次元数 128 のエンコーダを 4 層積層する．評価手法として 3 年のうちの 1 年を評価データ，残りを訓練データとして交差検証を行い，分類精度の平均と標準偏差を求める．また，訓練データに対してデータ拡張を適用し，モデルの訓練を行う．ここで，Mixup の適用率は 20%，CutMix の適用率は 40%とし，両手法を用いる場合は適用率を 10%とする．

データ拡張を用いて訓練した Transformer の分類精度を表 1 に示す．データ拡張後の成績分類の精度は Mixup を適用した際に 2.52pt 向上し，データ拡張の有効性を確認することができる．

表 1: Transformer の分類精度

Mixup	CutMix	精度 [%]
-	-	43.28 ± 8.44
✓	-	45.80 ± 7.91
-	✓	43.24 ± 5.87
✓	✓	44.54 ± 7.89

4. Attention weight の分析

Self-Attention の Attention weight を用いて成績の良い生徒と悪い生徒の分析を行う．ここで，Attention weight の値が大きい教材は成績を予測する上で重要度が高いデータとなる．

4.1. 成績の傾向分析

成績ごとに予測が成功した学生の Attention weight の平均を求め，傾向の分析を行う．2020 年，2021 年のデータで訓練・評価を行った際に正しく予測できた学生の成績ごとの Attention weight の平均を図 1 に示す．図 1 より，成績の良い学生は 12 番目の教材を注視し，成績の悪い学生は 7 番目から 12 番目の教材のように広範囲を注視している傾向があることが分かる．

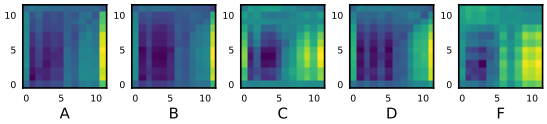


図 1: 成績ごとの Attention weight の平均

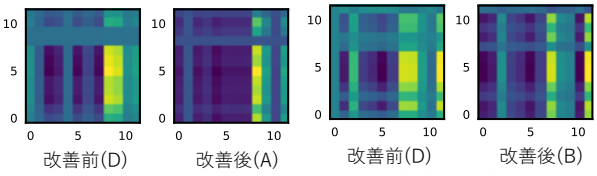
4.2. 成績不振の学生の改善

4.1 節で得られた傾向より，成績不振の各学生について Attention weight が大きい教材の操作回数を増加させて，擬似入力データを生成する．擬似データで再度予測を行い，分類結果が向上すれば，その操作が有効であると判断できる．12 番目の教材を除いて，Attention weight が大きい 6 つの教材について，操作回数を 5 倍に増やし，再度予測を行った．

各手法の再分類結果を表 2 に示す．CutMix が最も分類結果の向上する人数が多く，成績 D の学生の 50%の改善が確認できた．また，分類結果が A，B になった 2 人の学生の Attention weight を図 2 に示す．分類結果が向上した学生の Attention weight は，特定の教材を注視しているため成績が不振であったことが分かる．

表 2: 成績 D の学生の再分類精度 [%]

手法	A	B	C	D	F
データ拡張なし	21.05	21.05	0.00	57.90	0.00
Mixup	15.79	26.31	0.00	57.90	0.00
CutMix	15.00	35.00	0.00	50.00	0.00
Mixup&CutMix	20.00	20.00	0.00	60.00	0.00



(a) 学生の改善例 1　　**(b)** 学生の改善例 2

図 2: 成績が改善した学生の Attention weight の比較

5. おわりに

本研究では，Transformer による成績分類を行った．Mixup を行うことで成績分類精度が 2.52pt 高い結果となり，汎化性能の向上が確認できた．さらに，Attention weight の可視化により，成績が良い生徒，悪い生徒の傾向の違いを発見でき，それを活かした学習の改善方法を示した．今後は，成績予測に適した入力特徴について検討する．

参考文献

[1] A. Satake *et al.*, “Performance Prediction and Importance Analysis Using Transformer”, ICCE, 2021

[2] A. Vaswani *et al.*, “Attention is all you need”, NIPS, 2017

物体検出における実環境と CG データの差異の調査

ER19044　鈴木陽太郎　　指導教授：藤吉弘亘

1. はじめに

自動車に搭載する認識アルゴリズムを評価する際，同一シーンにおいて天候等の異なる状況下のデータを収集することは難しい．そこで，CG を用いて実環境を再現することで，異なる天候下の同一シーンデータの生成が可能となる．一方で，実環境の画像と CG の生成画像には，アピアランスの違いから発生するドメイン差があると認識アルゴリズムの検出精度に影響を与えることになる．そこで本研究では，実環境と CG で作成したデータに対して，深層学習モデルによる物体検出結果を比較して両データの差異について明らかにする．

2. DIVP シミュレータによる CG データの作成

本研究では，実世界の東京臨海部を撮影したシーンを，Driving Inteligence Validation Platform (DIVP) のシミュレータを用いて再現する．DIVP は，自動運転システムの評価を行う環境を構築する内閣府のプロジェクトで開発されたシミュレータである．RAW データからの現像処理をするため，DIVP シミュレータは実環境との偽色の再現性が一般公開データセットに比べて高い．JPEG 画像は色合い等の調整後に画質が劣化してしまうが，RAW データであれば調整後も画質が劣化することがない．テストデータ数は 6 つあり，各データは実環境と CG データを含む．CG データには，曇り，雨，13 時，16 時のバリエーションも含まれ，検出対象車両ではない車両は描写されていない．

3. YOLOv4

YOLOv4 [1] は，YOLO を拡張して検出精度と計算速度を向上させたモデルであり，backbone と neck，head から構成される．backbone は CSPDarknet53 により画像の特徴を抽出する．neck は backbone から受けとった特徴マップを Feature Pyramid Network(FPN) と Path Aggregation Network (PANet) により受容野の広い特徴量に変換する．head はクラス分類の予測や bounding box の位置を予測する．

4. 評価実験

実画像と CG 画像に対して YOLOv4 を用いて物体検出を行い，遠方車両の検出精度を比較する．YOLOv4 の学習には東京臨海部で収集した物体検出用データセットを用いる．本データセットの学習枚数は 19,261 枚である．

4.1. 定量的評価

本実験では，2 つのシーンを用いて比較実験を行う．各データをシーン 1，シーン 2 とし，各シーンに対して検出対象の車両との距離ごとの検出精度を比較する．図 1 に示すようにシーン 1 は対象車両に日が当たっており，シーン 2 は対象車両に日が当たっていない状態を対象とする．シーン 1 の検出精度を図 2，シーン 2 の検出精度を図 3 に示す．横軸は検出対象の車両までの距離，縦軸は車両の検出率を表す．また，青線は実環境上の検出精度，橙線は CG の検出精度を表す．シーン 1 では，100m までは実環境と CG で同程度の検出精度だが，100m 以上遠くなると CG 内での検出精度は実環境よりも低下する．シーン 2 では 20〜60m の距離では実環境と CG で同程度の検出精度だが，60m 以上遠くなると CG 内での車両の検出は不可能であった．

(a) シーン 1

(b) シーン 2

図 1: 評価対象のシーン（CG データ）

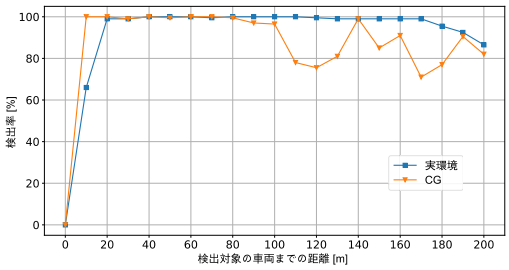


図 2: シーン 1 の検出率

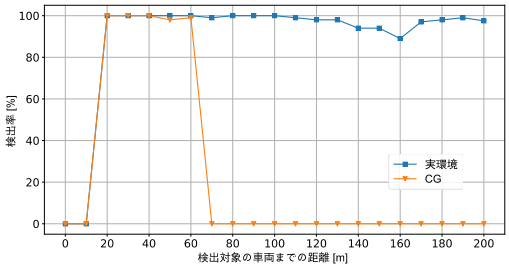
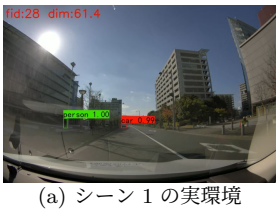


図 3: シーン 2 の検出率

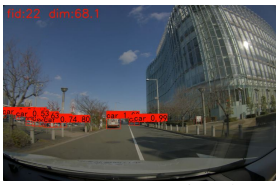
4.2. 定性的評価

各シーンで実環境と CG を比較した例を図 4 に示す．図 4 の (a) と (b) はシーン 1，図 4 の (c) と (d) はシーン 2 を示す．シーン 1 では，実環境と CG で遠方車両に対する検出が同程度であった．シーン 2 では，CG 内の検出対象の車両から遠くなると検出が不可能であったが，近くなると実環境と同程度の検出が可能であった．図 1 に示すように，シーン 1 とシーン 2 では CG 内の対象車両に対する影の影響が異なる．シーン 1 は影の影響が小さいデータ，シーン 2 は影の影響が大きいデータであり，評価実験からシーン 2 はシーン 1 に比べ検出精度が低下することから，実環境と CG 間で影がドメイン差の要因ではないかと考える．

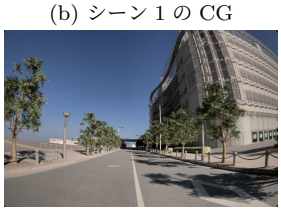


(a) シーン 1 の実環境

(b) シーン 1 の CG



(c) シーン 2 の実環境



(d) シーン 2 の CG

図 4: 実環境と CG データに対する検出例

5. おわりに

本実験では物体検出の精度から実画像と CG のドメイン差について調査した．CG 画像では，検出対象に対し影の影響が大きいことが判明した．今後は実環境と CG のドメイン差を低減するために影領域に対する CG の改良へとつなげる．

参考文献

[1] A. Bochkovskiy, *et al.*, “YOLOv4: Optimal Speed and Acuuracy of Object Detection”, CVPR, 2021.

Multi Object Tracking を用いた点滅信号の識別

ER19036 櫻井洋貴　　指導教授：藤吉弘亘

1. はじめに

自動運転の実現に向けて，信号機の識別は不可欠な機能の 1 つである．信号機を識別するためには，まず信号機の位置を特定する必要がある．信号機の識別には，デジタル地図情報を用いる方法がある．デジタル地図情報には，信号機の位置，距離，信号機 ID などが含まれており，これらの情報を基に入力画像の検出対象領域を限定することで高精度な検出が可能となる．しかし，デジタル地図情報の作成にはコストがかかり，位置ずれも起こる．そのため，デジタル地図情報に代わる画像認識による信号機識別が必要となる．また，信号機の状態には赤，青，黄，矢印，点滅など様々な状態があり，中でも点滅信号は単一のフレームから識別するのは困難である．そのため複数のフレームの色識別結果を基に点滅信号の識別を行う必要がある．そこで本研究では，セマンティックセグメンテーションにより信号機の検出を行い，複数物体追跡手法と時系列を考慮した状態識別を行う手法を提案する．

2. 提案手法

本研究では 4 つのステップにより点滅信号の識別を行う．まず第 1 ステップでは信号機領域をピクセル単位で検出するためにセマンティックセグメンテーションにより信号機領域の抽出を行う．第 2 ステップでは抽出した信号機の領域を基に信号機の色識別を行う．第 3 ステップでは複数のフレーム間で信号機の追跡を行う．最後に，複数のフレーム間での色識別結果を基に信号機の最終識別を行う．図 1 に信号機識別の流れを示す．

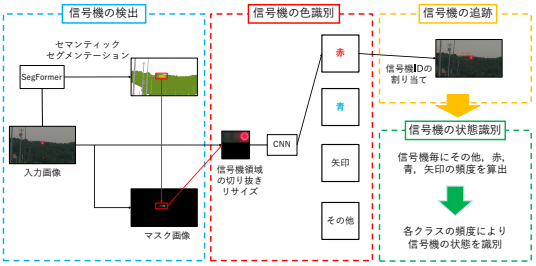


図 1：信号機識別の流れ

2.1. 信号機の検出

遠方の信号機や遮蔽物に一部が隠れた信号機の検出には，SegFormer [1] を用いたセマンティックセグメンテーションを行い，信号機の領域を検出する．セマンティックセグメンテーションにより検出した信号機領域から外接矩形を獲得し．マスク処理を行う．

2.2. 信号機の色識別

信号機の色識別には CNN を用いる．ネットワークには ResNet50 を用いる．2.1 節で検出した信号機の領域を正方形に切り出し，224×224 ピクセルにリサイズして ResNet50 に入力する．識別クラスは赤，青，矢印，その他の 4 クラスである．また，黄色信号は赤のクラスに識別する．

2.3. 信号機の追跡

点滅信号を識別するためには，複数のフレーム間で同信号機であることを識別する必要がある．そのため，Deep Sort を用いた Multi Object Tracking(MOT) により信号機を追跡する．これにより，連続するフレーム間で同じ物体に対して同じ ID を割り当てることができる．ID を付与した複数のフレームの情報を用いて信号機を識別することが可能となる．

2.4. 信号機の状態識別

信号機の色識別結果を基に信号機ごとに，その他，赤，青，矢印の頻度を算出する．各クラスの頻度により信号機の最終識別を行う．以下に識別方法を示す．

青信号： 青が最も多い場合

矢印信号： 矢印が最も多い場合

赤信号： 赤が最も多いかつ (青+矢印)< 赤の場合

点滅信号： 0.5<(赤/その他)<2 かつ
(青+矢印)<(赤+その他) の場合

3. 評価実験

本実験により提案手法と，デジタル地図情報を用いた場合の信号機の識別精度を比較する．

3.1. 実験概要

データセットは学習に 620 枚，評価に 13,421 枚，信号機の数 79 基で構成される東京臨海部のセマンティックセグメンテーションデータセットを使用する．学習データセットを用いて SegFormer を学習する．評価データに対して，提案手法とデジタル地図情報を用いた際の信号機の識別精度を比較する．

3.2. 実験結果

表 1 に提案手法，表 2 にデジタル地図情報を利用した場合の信号機の識別結果を示す．点滅信号の識別率は提案手法が約 57.1%，デジタル地図情報を利用した場合が約 58.1% でほぼ同等の識別結果となった．

表 1：提案手法の識別結果

	正解数	不正解数	識別率 [%]
青信号	13	0	100
赤信号	3	0	100
点滅信号	36	27	57.1

表 2：デジタル地図情報を利用した場合の識別結果

	正解数	不正解数	識別率 [%]
青信号	13	0	100
赤信号	3	0	100
点滅信号	37	26	58.7

信号機検出の成功例と失敗例を図 2 に示す．灯火状態の信号機は，遠距離の場合でも検出が可能である．しかし，無灯火状態の信号機は背景や日の当たり方によって検出が困難ということが確認できた．これにより，信号機 ID が切り替わり，点滅信号の識別率が低下したと考える．また，無灯火信号の識別可能距離の平均は，提案手法が約 90m，デジタル地図情報を用いた場合が約 95m でほぼ同等の結果となった．

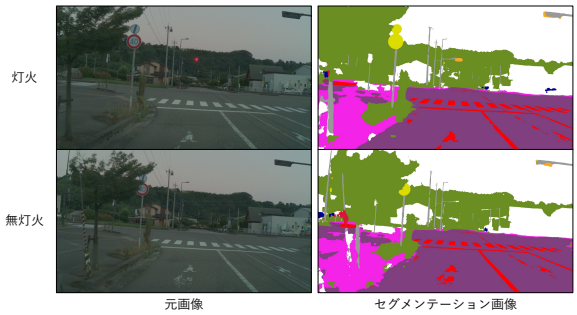


図 2：信号機検出の成功例と失敗例

4. おわりに

本研究では，提案手法と，デジタル地図情報を用いた場合の信号機の識別精度の比較を行った．提案手法とデジタル地図情報の信号機の識別精度は，ほぼ同等であるということが確認できた．これにより，提案手法はデジタル地図情報の作成コスト削減が可能であり，代替手段になると考える．今後は，遠方の無灯火信号の識別，信号機追跡手法の精度向上を行う．

参考文献

[1] E. Xie, *et al.*, “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers”, In NeurIPS, 2021.

マルチドメインに対応した半教師あり学習によるセマンティックセグメンテーション

ER19055 中村泰智　　指導教授：藤吉弘亘

1. はじめに

セマンティックセグメンテーションは，入力画像の画素単位でクラス分類するタスクである．モデルの学習には，画素ごとにアノテーションした正解データが大量に必要である．また，学習と評価するドメインが異なると精度が著しく低下する問題もある．そこで本研究では, 各ドメインに対応した出力を行うマルチヘッド構造を導入することでマルチドメインの問題を解決する．さらに，半教師あり学習を導入することで，アノテーションコストの削減を目指す．

2. マルチドメイン学習

マルチドメイン学習は，複数のドメインのデータセットを同時に学習する手法である．一般的なセグメンテーションネットワークでは, クラス数が異なるデータセットを単一のモデルで学習することは難しい．そこで，文献 [1] は全てのドメインで共通するバックボーンとドメインごとの出力を行うマルチヘッド構造を導入した．しかし, 文献 [1] は教師あり学習のため，すべてのドメインに対する大量の正解付き学習データが必要である．

3. 提案手法

本研究ではアノテーションコストを削減するために，半教師あり学習である CAC[2] をマルチドメイン構造を持つセマンティックセグメンテーション手法に導入する．

3.1. ネットワーク構造

ベースネットワーク構造に，ResNet-50 をバックボーンにした DeepLabv3 ＋を用いる．ネットワーク構造を図 1 に示す．学習時に 1 枚の正解なし学習データから重複領域が存在するランダムなパッチを 2 つ抽出する．2 つのパッチからピクセル単位で対照学習を行うことで正解なし学習データの整合性を保つことができ，環境の変化に対するロバスト性が向上する．

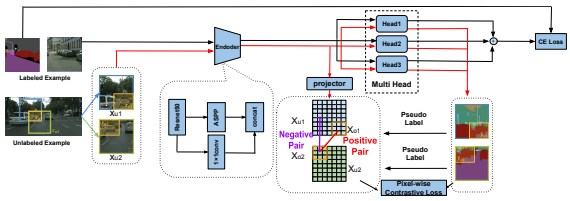


図 1：提案手法の CAC の構成

3.2. 学習方法

半教師あり学習は，教師あり学習と教師なし学習を組み合わせて学習する方法であり，CAC では擬似ラベリングを使用する．擬似ラベリングは，正解つき学習データを用いた教師あり学習を行ったモデルを用いて正解なし学習データに対して，擬似ラベルを付与する方法である．

教師あり学習

教師あり学習の損失関数は式（1）の Cross-Entropy Loss (CELoss) を用いる．このとき p は教師データ，q 正解付き学習データ，i クラス数である．

$$L_{ce} = \sum_{n=1}^{N} p(i) log(q(i)) \tag{1}$$

教師なし学習

教師なし学習の損失関数は式（2）の Directional Contrastive Loss（DCLoss）を用いる．まず，正解なし学習データから重複領域を含む 2 つのパッチをランダムに切り出して，モデルに入力し，各画素に対するクラス分類を行う．そして，重複領域に着目し，クラス分類の信頼度が低い特徴を信頼度が高い特徴に近づくように学習を行う．

$$L_{dc}(\dot{y}) = -\frac{1}{N}\sum_{h,w} M_{d,p}^{h,w} log \frac{r(\dot{y})}{r(\dot{y})\sum M_{n,1}^{h,w} r(x_{o1}, x_{u2})} \tag{2}$$

ここで，$\dot{y}$ は 2 つのパッチの重複領域,x_{o1} は 1 枚目の重複領域のピクセル，x_{u2} は 2 枚目のパッチのピクセル，r は類似度の指数関数，N は重複空間のピクセル数，$M_{n,1}^{h,w}$ は正解なしデータに擬似ラベルを付与する関数，$M_{d,p}^{h,w}$ は閾値より信頼度の低い特徴を計算から除外する関数である．

マルチヘッド構造

マルチヘッド構造において各データセットごとに逆伝播を行う場合，学習を行うデータセットの順番によって特定のデータセットにバイアスがかかる可能性がある．そこで，Mix Loss により全てのドメインデータを入力し，各ヘッドで出力した Loss を合計してから逆伝播を行う．Mix Loss を式 (3) に示す．このとき λ は，教師なし Loss の重み，L_{ce} は CELoss，l_{dc} は DCLoss，x_n は各ドメインの入力サンプル，N はデータセット数である．

$$L = \sum_{n=1}^{N} (L_{ce} + \lambda L_{dc}) x_n \tag{3}$$

4. 評価実験

本章では複数のデータセットを用いて，提案手法の有効性を評価する．

4.1. 実験概要

実験では Cityscapes，Synscapes，Berkley DeepDrive 100K（BDD100K）データセットを使用する．最適化手法に SGD を使用し，重み減衰を 0.0001，モーメンタムを 0.9，初期学習率を 0.01 とする．また，全てのモデルを各データセットにおいて 150 エポック学習する．さらに，半教師あり学習では各データセットで学習する総枚数の $\frac{1}{8}$ をラベルあり画像，$\frac{7}{8}$ をラベルなし画像で学習する．教師あり学習は各データセットで学習する総枚数の $\frac{1}{8}$ をラベルあり画像で学習する．評価指標には mIoU を用いる．

4.2. 実験結果

定量的評価を表 1，結果画像を図 2 に示す．表 1 より，提案手法は従来手法と比較し，全てのデータセットで精度の向上が見られた．また，マルチドメインはシングルドメインと比較し，BDD100K データセットの認識精度が 9.22pt 向上した．定性的評価より，図 2 の赤丸で示すように rider と bicycle が正解であるが，シングルドメインでは people と認識している，一方，提案手法は rider と bicycle を正しく認識していることが確認できる．

表 1：シングルドメインと提案手法の比較 [%]

	学習方法	Cityscaps	Synscapes	BDD100k	平均
シングルドメイン	従来手法	61.83	80.22	41.99	61.35
	提案手法	63.85	**81.45**	43.85	63.05
マルチドメイン	従来手法	63.24	73.57	50.18	62.33
	提案手法	**65.52**	77.60	**53.07**	**65.40**

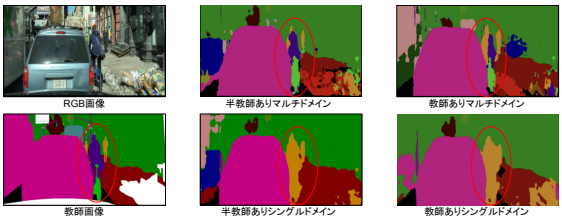


図 2：BDD100K データセットの結果画像

5. おわりに

本研究では，マルチドメインに対応した半教師あり学習によるセマンティックセグメンテーション手法の有効性を示した．今後は，データセットの数や新たな学習方法を考案し，より高い精度を得られる条件を検討する．

参考文献

[1] M.Shota, *et al.*, “Multi-Domain Semantic-Segmentation using Multi-Head Model”, ITSC, 2021.

[2] X.Lai, *et al.*, “Semi-supervised Semantic Segmentation with Directional Context-aware Consistency”, CVPR, 2021.