動画像認識における深層学習モデルへの人の知見の導入に関する研究

TR23005 野口 紗季　　指導教授：藤吉 弘亘

1. はじめに

動画像認識は，複数フレームからなる動画中の動作を認識するタスクである．動画像認識に深層学習を用いることで高い認識性能を実現できる．静止画像の物体認識では，人の知見を導入することでモデルの注視領域が改善し，認識精度が向上することが知られている [1]．一方，動画像認識では，時空間を考慮する必要があるため，人の知見の導入に関する研究はされていない．そこで本研究では，動画像認識モデル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. 深層学習による動画像認識

深層学習による動画像認識は，CNN または Transformer を用いるモデルが提案されている．CNN ベースのモデルは，畳み込み処理により各画素に隣接するピクセル間の関係性から局所的な特徴を抽出する．Transformer ベースのモデルは，入力を複数のトークンに分割し Self-Attention でトークン間の関係性から大域的な特徴を抽出する．本章では，CNN ベースのモデルである Spatio-Temporal Attention Branch Network (ST-ABN) [2] と，Transformer ベースのモデルである Video Transformer Network (VTN) [3] について述べる．

2.1 ST-ABN

Spatio-Temporal Attention Branch Network (ST-ABN) [2] は，空間情報と時間情報を同時に考慮した CNN ベースの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，モデル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．

2.2 VTN

Video Transformer Network (VTN) [3] は，フレーム単位で抽出した特徴をトークンとして扱う Transformer ベースの動画像認識手法である．VTN はフレームごとに空間特徴を抽出する 2D backbone と，2D backbone の出力をトークンとして時間特徴を抽出する Temporal attention-based Encoder，クラス確率を出力する MLP Head で構成される．

3. 人の知見データの作成

複数フレームの映像から動作を認識するためには，各フレームの空間情報だけでなく，フレーム間の関係性を表す時間情報も重要である．空間情報に関する人の知見データは提案されている [1] が，時間情報に関する人の知見データは存在しない．そこで，動画像認識におけるベンチマークである Something-Something v.2 データセットに対して時間情報に関する人の知見データを作成する．具体的には，動画像をフレーム単位に分割し，各フレームの重要度を人の知見に基づいて付与する．このとき，認識に不要なフレームに 0.0，認識に必要な動きのあるフレームに 0.5，認識に重要なフレームに 1.0 を付与する．

4. 提案手法

本章では，CNN ベースの ST-ABN と Transformer ベースの VTN への時間情報に対する人の知見導入方法について述べる．

4.1 ST-ABN への人の知見の導入

ST-ABN は，動画像の認識に有効な Spatial Attention や Temporal Attention を獲得できない場合，誤認識を誘発することがある．静止画像による物体認識において，空間情報に関するモデルの注視領域への人の知見導入の有効性は既に確認されている [1]．そこで本節では，図 1 に示すように ST-ABN をファインチューニングすることで，時間情報に関するモデルの注視領域へ人の知見を導入する．ファインチューニング時の ST-ABN の学習誤差を式 (1) に示す．

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

ここで，$\mathcal{L}_{att}$ と $\mathcal{L}_{per}$ は ST-ABN の ST Attention Branch と Perception Branch の学習誤差である．また，$\mathcal{L}_{temp}$ は式 (2) に示すように，モデルから出力される Temporal Attention M_t と，時間情報に関する人の知見データ M'_t の平均二乗誤差で算出する．ここで，n は入力フレーム数，γ_t は学習誤差 $\mathcal{L}_{temp}$ を調整する係数である．

$$\mathcal{L}_{temp} = \gamma_t \frac{1}{n} \sum_{i=1}^{n} (M'_{t\_i} - M_{t\_i})^2 \tag{2}$$

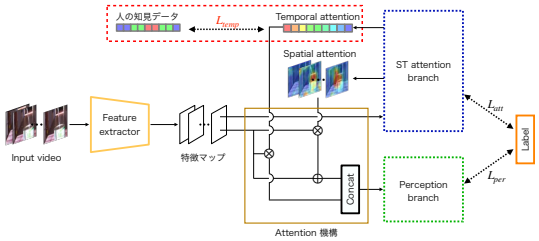


図 1: ST-ABN への人の知見の導入

4.2 VTN への人の知見の導入

Transformer ベースのモデルの Self-Attention には，トークン間の関係性を特徴マップに重み付けする Attention 機構の役割が含まれている．そのため Self-Attention で適切なトークン間の関係性が獲得できない場合，誤認識を誘発する．VTN の Temporal attention-based Encoder には Self-Attention が用いられている．そこで Temporal attention-based Encoder 内の Self-Attention に人の知見を導入するために，Learnable Bias と Reactive Bias を以下の手順で追加する．

Step1: Learnable Bias の学習　Self-Attention に人の知見を導入するために，はじめに Learnable Bias で下流データセットのバイアスを学習する．Learnable Bias は図 2 に示すように，Self-Attention における Query と Key の内積演算によって求めた Attention Weight に加算する学習可能なパラメータである．Learnable Bias 学習時の Self-Attention は式 (3) のように定式化する．このとき，Temporal attention-based Encoder のパラメータは更新しない．

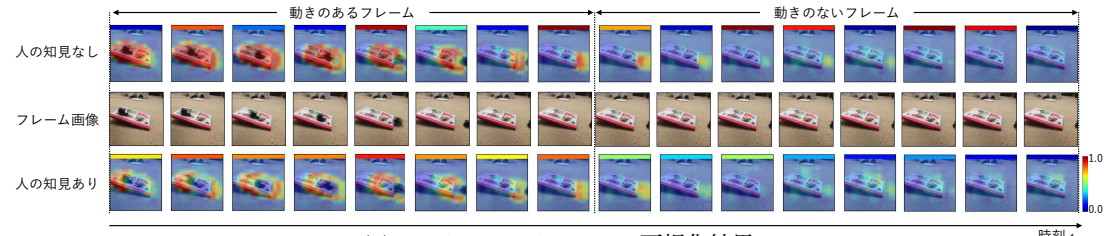
$$\mathbf{SA} = \mathrm{Softmax}\left(\frac{\boldsymbol{QK}^T}{\sqrt{d_h}} + \boldsymbol{l}\right)\boldsymbol{V} \tag{3}$$

ここで，$\boldsymbol{l}$ は Learnable Bias，$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ は Transformer における Self-Attention の Query，Key，Value，d_h は Query と Key の次元数である．

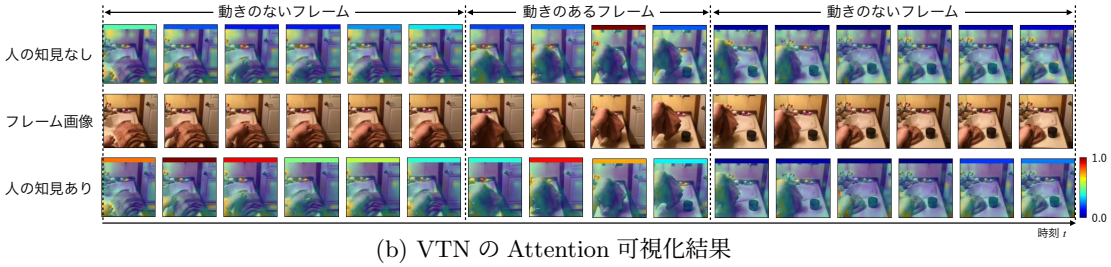
Step2: Reactive Bias の学習　Learnable Bias の学習をした後，Reactive Bias を学習する．Reactive Bias は，Learnable Bias を制御することで個々のデータに適応させるモジュールである．図 2 に示すように Reactive Bias は，Self-Attention における Query と Key の内積演算をベースとした構造をしている．Reactive Bias は，入力 $\boldsymbol{z}$，線形層の重み $\boldsymbol{W}^{\boldsymbol{Q}'_{\mathrm{RB}}}$, $\boldsymbol{W}^{\boldsymbol{K}'_{\mathrm{RB}}}$ を用いて式 (4) ように定式化する．

$$\boldsymbol{R}(\boldsymbol{z}) = \frac{(\boldsymbol{z}\boldsymbol{W}^{\boldsymbol{Q}'_{\mathrm{RB}}})(\boldsymbol{z}\boldsymbol{W}^{\boldsymbol{K}'_{\mathrm{RB}}})^T}{\sqrt{d_h}} \tag{4}$$

Reactive Bias の学習時に，人手によって作成した Attention をマスクとして適用することで人の知見を導入する．人手によって作成した Attention を $\boldsymbol{M}$ としたとき，Self-Attnetion は式 (5) のように定式化する．このとき，Reactive Bias と MLP Head のパラメータのみを更新する．



(a) ST-ABN の Attention 可視化結果



(b) VTN の Attention 可視化結果

図 3: 人の知見導入による Attention の変化

$$\mathbf{SA} = \mathrm{Softmax}\left(\frac{\boldsymbol{QK}^T}{\sqrt{d_h}} + \boldsymbol{l} \odot \boldsymbol{R} \odot \boldsymbol{M}\right)\boldsymbol{V} \qquad (5)$$

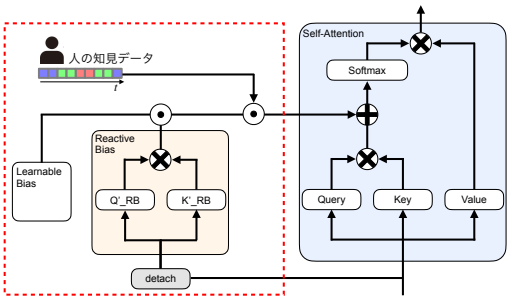


図 2: VTN への人の知見の導入

5. 評価実験

時系列情報を介した人の知見導入の有効性を確認するために，評価実験を行う．

5.1 実験条件

本実験では，Something-Something v.2 データセットを用いて実験を行う．人の知見データの作成は全 174 クラスのうち，ST-ABN の認識精度が特に低い 8 クラスの動画を対象とする．ST-ABN のベースネットワークとして 3D ResNet-50 を使用し，学習誤差 $\mathcal{L}_{temp}$ を調整する係数 γ_t は 10 とする．また，VTN の 2D backbone には Vision Transformer を使用し，事前学習には Time Is MattEr [4] フレームワークを用いる．

5.2 定性評価

人の知見導入による Attention の変化を図 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．図 3(a) より，人の知見の導入前の ST-ABN の Temporal Attention 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．それに対し人の知見の導入後は，動きの大きさとカラーバーの色が一致している．そのため，人の知見を導入することでより適切な注視領域が獲得できたといえる．また，Spatial Attention は人の知見の導入によって対象物体上の注視領域が減少し，フレーム前後で変化した領域に注視している．このことから Temporal Attention の修正により Spatial Attention も改善されたといえる．

VTN へ人の知見を導入することによる Attention 可視化結果を図 3(b) に示す．人の知見の導入前は単一のフレームのカラーバーのみが赤色となっており，モデルは特定のフレームに依存した認識を行っている．それに対し，人の知見の導入後は，動きのある複数のフレームのカラーバーが赤色となっている．人間による動画像認識も複数のフレームの情報に基づいて行われることから，より人間に近い注視領域が獲得できるようになったといえる．しかし，動画前半の動きのないフレームのカラーバーも一部赤色になっていることから，更なる改善が必要であると考える．

5.3 定量評価

人の知見の導入による全 174 クラスの認識精度の変化と人の知見データを作成した 8 クラスの認識精度の変化をそれぞれ表 1 に示す．ここで，HK は人の知見の有無を示しており，チェックマークは人の知見が導入されていることを表す．表 1 より ST-ABN は人の知見の導入により，人の知見データを作成した 8 クラスの認識精度が約 6pt 向上している．そのため，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．一方 VTN は，人の知見データを作成した 8 クラスの認識精度が約 5pt 低下した．これは人の知見の導入により，モデルが動作を認識するのに必要のないフレームにも注視するようになったためだと考えられる．

表 1: 人の知見の導入による認識精度の変化 [%]

		全 174 クラス		8 クラス
ベースモデル	HK	Top-1	Top-5	Top-1
ST-ABN		58.62	85.45	20.50
ST-ABN	✓	**60.65**	**86.93**	**26.32**
VTN		**56.73**	**84.20**	**29.64**
VTN	✓	54.28	81.04	24.52

6. おわりに

本研究では，動画像認識における深層学習モデルへ人の知見を導入する方法を提案した．評価実験より，ST-ABN に人の知見を組み込むことでモデルの注視領域が改善し，認識精度が向上することが確認できた．今後は Transformer ベースのモデルにおいて，人の知見を導入した際に動きのないフレームにも注視するようになる原因調査を行う．

参考文献

[1] M. Mitsuhara, *et al.*, “Embedding Human Knowledge into Deep Neural Network via Attention Map”, VISAPP, 2021.

[2] M. Mitsuhara, *et al.*, “ST-ABN: Visual Explanation Taking into Account Spatio-temporal Information for Video Recognition”, arXiv, 2021

[3] D. Neimark, *et al.*, “Video transformer network”, ICCV, 2021.

[4] S. Yun, *et al.*, “Time Is MattEr: Temporal Self-supervision for Video Transformers”, ICML, 2022.

研究業績

[1] S. Noguchi, *et al.*, ”Embedding Human Knowledge into Spatio-Temproal Attention Branch Network in Video Recognition via Temporal Attention”, BMVC, 2023.
(他 1 件)