動画像認識における ST-ABN を用いた 人の知見の組み込みによる説明性の向上と高精度化

野口 紗季† 史 宇植† 平川 翼† 山下 隆義† 藤吉 弘亘†
†中部大学
E-mail: noguchi@mprg.cs.chubu.ac.jp

1 はじめに

動画像認識は，複数フレームの映像から動画内における動作などを認識するタスクである．3D CNN 等の動画像認識に対応した Convolutional Neural Network (CNN) が多く提案され，高い認識性能を実現している．動画像認識を社会実装するには認識精度だけでなく，信頼性の担保も重要である．特に，モデルの判断根拠を人間が理解可能な深層学習モデルの実現が期待されている．モデルの判断根拠に対する解釈方法の一つとして，視覚的説明がある．視覚的説明はネットワークの注視領域を可視化することで，判断根拠の解釈を可能とする．静止画像に対する物体認識における視覚的説明には Gradient-weighted Class Activation Mapping (Grad-CAM) [1]，Class Activation Mapping (CAM) [2]，Attention Branch Network (ABN) [3] 等がある．中でも ABN は，注視領域をアテンションマップとして可視化して特徴マップに重み付けすることで，高い認識精度を実現した．しかし，学習データに偏りが発生すると認識に必要な領域に注視できず誤認識が生じる．これを解決する手法として，ABN における人の知見を介したファインチューニング [4] がある．この手法は，誤認識したサンプルのアテンションマップを人手により修正し，修正したアテンションマップを用いてファインチューニングする．これにより，人が認識する際の注視領域に近いアテンションマップの獲得と高精度化を実現している．このように，画像認識分野においては人の知見を組み込むことのできる視覚的説明手法があり，その有効性が示されている．一方で，動画像認識分野を対象とした人の知見を組み込む手法はこれまでに提案されていない．

本研究は，動画像認識における視覚的説明として Spatio-Temporal Attention Branch Network (ST-ABN) と，ST-ABN に人の知見を組み込む手法を提案する．ST-ABN は，推論時の空間情報と時間情報に対する重要度を獲得し認識処理に応用することで，認識性能の向上と視覚的説明を同時に実現する．また，ST-ABN は獲得した重要度を認識に応用するアテンション機構を持つため，人の知見を組み込むことができる．本稿では，はじめに Something-Somethig データセットにより，従来の動画像認識手法と提案手法の認識精度の比較を行う．次に，Spatial attention と Temporal attention を評価することで，空間情報と時間情報を同時に考慮した視覚的説明性と人の知見の導入の有効性を示す．

2 関連研究

本章では，深層学習による動画像認識に関する手法と視覚的説明に関する手法について述べる．

2.1 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，2D CNN による手法と 3D CNN による手法，Transformer ベースの手法に分けられる．2D CNN による手法には，空間的な特徴を捉える CNN とフレーム間の動き情報を表すオプティカルフローを捉える CNN にそれぞれ入力する，2 ストリームなネットワーク構造に基づく手法 [5, 6] が多く提案されている．これらの手法では，時刻ごとに獲得した空間的な特徴とオプティカルフローを集約することで，空間情報と時間情報を同時に考慮できる．しかし，空間情報と時間情報は独立した特徴として捉えている．3D CNN による手法では，2D の畳み込み処理を時間方向に拡張することで空間方向と時間方向に対して同時に畳み込み処理を行う 3D の畳み込み処理 [7, 8] により時空間的な特徴を獲得する．3D CNN による手法は，2D CNN の手法では困難であった空間情報と時間情報との相互関係を考慮した時空間特徴を学習できる．しかし，これらの CNN ベースの手法は局所的な時空間の関係性しか考慮できず，短い動画像しか入力できない．Transformer ベースの手法は，動画中のフレーム画像をパッチに分割しそれぞれのフレーム画像の同じ位置にあるパッチから時間情報を，同じフレーム画像内のパッチから空間情報の特徴を獲得する [9, 10]．Transformer ベースの手法は，CNN ベースの手法よりも大域的な時空間の関係性を捉えた学習を高速に行うことができる．

2.2 視覚的説明

深層学習を用いた画像認識では，ネットワークが注視した領域をヒートマップで表現するアテンションマップにより判断根拠を解析する視覚的説明 [1, 2, 3] が提案されている．視覚的説明には，逆伝播時の勾配を用いてアテンションマップを獲得する手法と，ネットワークの応答値からアテンションマップを獲得する手法がある．逆伝播時の勾配を用いる Gradient-weighted Class Activation Mapping (Grad-CAM) [1] は，逆伝播時の特定のクラスに対する勾配を用いることで，アテンションマップを獲得する．Grad-CAM は，学習済みの様々なネットワークに対する視覚的説明を獲得することができる．

ネットワークの応答値を用いる Class Activation Mapping (CAM) [2] では，畳み込み層から得られる特徴マップからアテンションマップを獲得する．アテンションマップを獲得する際は，畳み込み層で得られたチャネル数分の特徴マップと各チャネルに対応する全結合層の結合重みを用いることで，各クラスにおけるアテンションマップを獲得する．しかし，CAM は畳み込み層と全結合層の間の Global Average Pooling (GAP) により空間情報が欠落するため，認識精度の低下を誘発しやすい．CAM の問題を解決した手法として，Attention Branch Network (ABN) [3] が提案されている．ABN は，アテンションマップをアテンション機構へ応用することで，より視覚的な説明性の高いアテンションマップの出力と認識精度の向上を同時に実現している．

深層学習で用いられるネットワークは，クラスラベルのみを用いて学習しているため適切な領域に注視できないことがある．一方，人間は経験の中で認識に必要な情報を既に持っており認識時に有益な領域を正確に判断することが可能である．ABN のようなアテンション機構を用いたモデルは認識時にアテンションマップを入力に重み付けするため，不適切な注視領域が出力されると誤認識を誘発する．これを解決する手法として，ABN における人の知見を介したファインチューニング [4] がある．この手法は，ABN が誤認識したサンプルのアテンションマップを人手により修正し，修正したアテンションマップを用いてファインチューニングすることで，人が認識する際の注視領域に近いアテンションマップの獲得と ABN の高精度化を実現している．しかし，動画像認識において人の知見を導入できる手法は存在しない．

3 提案手法

本研究では，動画像認識において重要となる空間情報と時間情報を同時に考慮する視覚的説明が可能な Spatio-Temporal Attention Branch Network (ST-ABN) を提案する．提案する ST-ABN はアテンション機構を用いているため，三津原らの手法 [4] と同様にアテンションを手動で修正することでモデルに人の知見を組み込むことができる．

3.1 ST-ABN

ST-ABN は，図 1 に示すように Feature extractor, ST Attention branch，Perception branch の 3 つのモジュールで構成する．Feature extractor は，複数の畳み込み層で構成されており，入力から特徴マップを獲得する．ST Attention branch は，空間情報に対する重要度を示す Spatial attention と時間情報に対する重要度を示す Temporal attention を獲得する．Perception branch は，アテンション機構により Spatial attention と Temporal attention を重み付けした特徴マップを入力し，各クラスの確率を出力する．

3.1.1 Spatio-Temporal Attention branch

Spatio-Temporal Attention branch (ST Attention branch) では，Feature extractor から出力された特徴マップを用いて空間情報と時間情報に対する重要度を獲得し，GAP を介してクラス識別を行う．Feature extractor から出力された特徴マップは，複数の畳み込み層を経てクラス数分のチャネルを持つ 1×1 の畳み込み層によりクラス数分の特徴マップとなる．

Spatial attention は，このクラス数分の特徴マップを用いて生成する．クラス数分の特徴マップは，1×1 の畳み込み層により 1 つの特徴マップに集約する．その後，フレームごとに Sigmoid 関数で 0 から 1 の範囲に正規化したアテンションマップを獲得する．

Temporal attention も Spatial attention と同様にクラス数分の特徴マップを用いて生成する．はじめに，クラス数分の特徴マップを 1×1 の畳み込み層によりチャネル方向に対して次元を圧縮する．その後，フレーム数分のチャネル数を持つ畳み込み層と GAP を介して，各特徴マップの空間方向に対する平均値を求める．最後に，全結合層，ReLU 及び Sigmoid 関数を介して，各フレームの重要度を獲得する．

3.1.2 アテンション機構

入力動画を $\mathbf{x}_i$ としたときの Spatial attention $M_s(\mathbf{x}_i)$ は，式 (1) より特徴マップ $f(\mathbf{x}_i)$ に重み付けし，重み付け前の特徴マップを加算する．これにより，特徴マップの消失を抑制し，アテンションマップを効率的に認識に反映させることができる．

$$f'_s(\mathbf{x}_i) = (1 + M_s(\mathbf{x}_i)) \cdot f(\mathbf{x}_i) \tag{1}$$

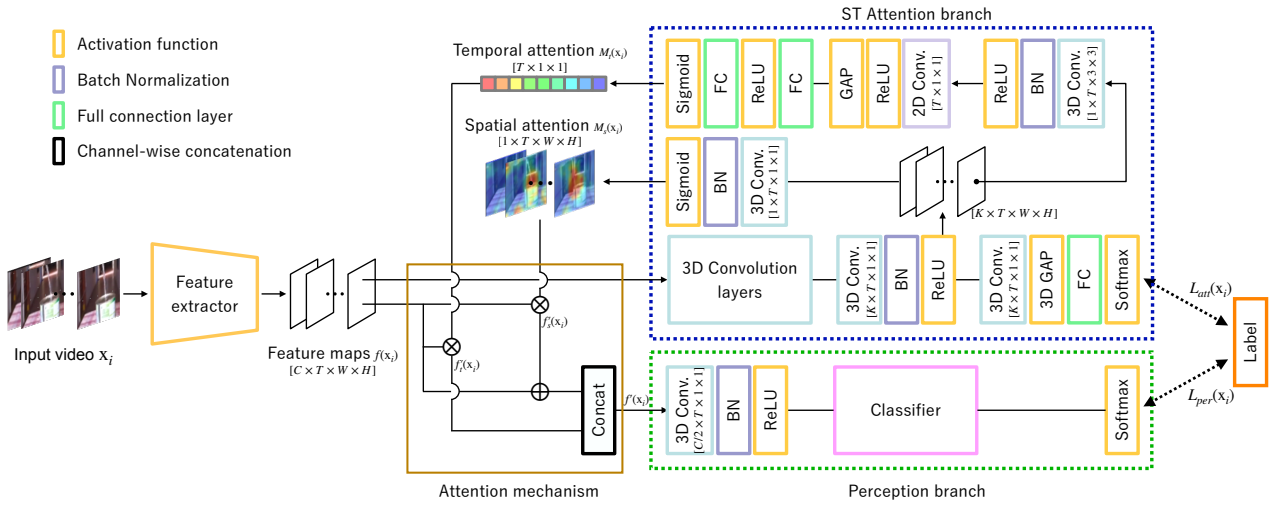

図 1　ST-ABN のネットワーク構造

ここで，$f_s'(\mathbf{x}_i)$ は Spatial attention を重み付けした特徴マップを示す．Temporal attention $M_t(\mathbf{x}_i)$ は，式 (2) より特徴マップに乗算することで重み付けを行う．

$$f_t'(\mathbf{x}_i) = M_t(\mathbf{x}_i) \cdot f(\mathbf{x}_i) \tag{2}$$

ここで，$f_t'(\mathbf{x}_i)$ は Temporal attention を重み付けした特徴マップを示す．Spatial attention と Temporal attention をそれぞれ重み付けした特徴マップは，式 (3) によりチャネル方向に結合する．

$$f'(\mathbf{x}_i) = \mathrm{concat}[f_s'(\mathbf{x}_i), f_t'(\mathbf{x}_i)] \tag{3}$$

この結合した特徴マップを Perception branch に入力し，最終的な認識結果を出力する．これにより，重要な時空間特徴に着目した学習が可能となる．

3.1.3　ST-ABN の学習

ST-ABN の学習誤差 $\mathcal{L}(\mathbf{x}_i)$ は，式 (4) のように ST Attention branch の学習誤差と Perception branch の学習誤差の単純な加算により求める．

$$\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i) \tag{4}$$

ここで，$\mathcal{L}_{att}(\mathbf{x}_i)$ は ST Attention branch の学習誤差，$\mathcal{L}_{per}(\mathbf{x}_i)$ は Perception branch の学習誤差である．各ブランチの学習誤差は Softmax 関数とクロスエントロピー誤差を用いて算出する．また，ST-ABN の Feature extractor は ST Attention branch と Perception branch の勾配を受け取ることで最適化される．

3.1.4　実装の詳細

本章では ST-ABN の実装について説明する．ST-ABN はベースネットワークを Feature extractor と Perception branch に分割し，Feature extractor と Perception branch の間に ST Attention branch を追加することで構築される．そのため，C3D や 3D ResNet などさまざまなネットワークモデルに容易に導入することができる．本研究では，ベースネットワークとして ResNet [11] を時間方法に拡張した 3D ResNet を slowfast network の slow pathway [12] に基づいて用いる．入力の空間次元は 224×224，データサイズは $C \times T \times W \times H$ である．また，ST Attention branch と Perception branch に対して 0.5 のドロップアウト [13] を適用し，オーバーフィッティングを抑制している．

3.2　ST-ABN への人の知見の導入

ST Attention branch で獲得したアテンションも他の視覚的説明手法と同様に，不適切な注視領域を示していることがある．しかし ST-ABN は，認識時にアテンションマップを特徴マップに重み付けし認識に用いるアテンション機構があるため人の知見を組み込むことができる．

3.2.1　Temproal attention の修正

図 2 に示すように 3 つの Step によって ST-ABN へ人の知見を導入し，認識により有益なアテンションを獲得可能にする．

Step1　ST-ABN が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．

Step3　修正した Temproal attention を真値としてファインチューニングする．

Temporal attention の修正例を図 3 に示す．ここで，Temporal attention は 0 ∼ 1 の範囲で正規化し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

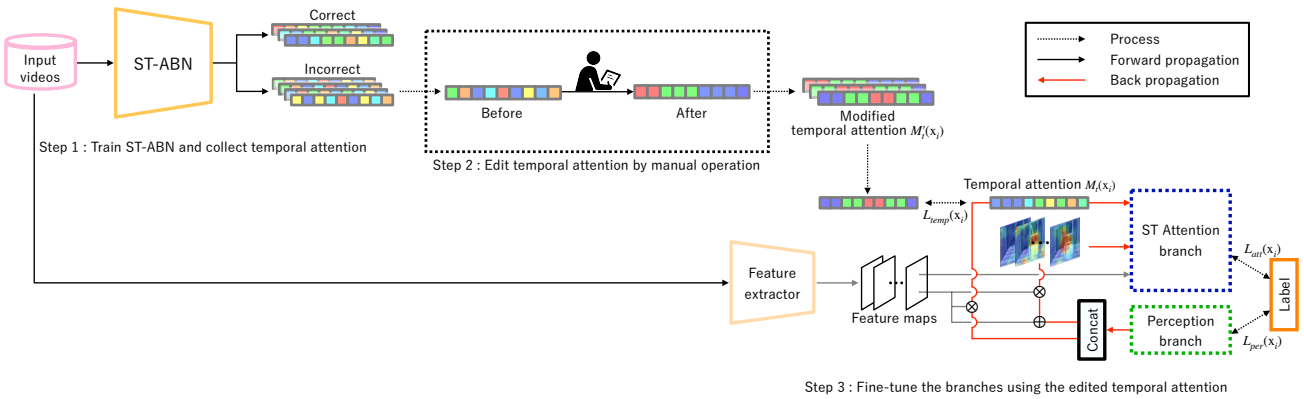

図 2　ST-ABN への人の知見導入の流れ

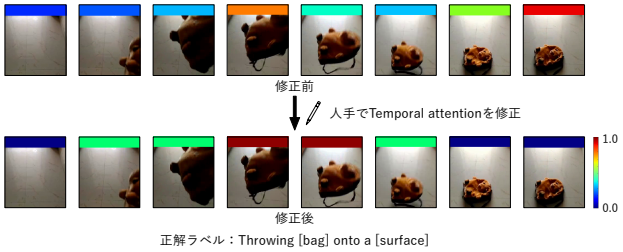

図 3　Temporal attention 修正例

は，認識に不要なフレームにアテンションスコア 0.0 を，認識に必要な動きのあるフレームにはアテンションスコア 0.5 を，認識時に特に重要なフレームにはアテンションスコア 1.0 を付与する．

3.2.2　ST-ABN のファインチューニング

人手によって修正した Temporal attention を用いて ST-ABN をファインチューニングすることで，ST-ABN に人の知見を組み込む．ファインチューニング時には，式 (5) に示すように，ST-ABN の学習誤差 $\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i)$ に $\mathcal{L}_{temp}(\mathbf{x}_i)$ を追加する．

$$\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i) + \mathcal{L}_{temp}(\mathbf{x}_i) \quad (5)$$

$\mathcal{L}_{temp}(\mathbf{x}_i)$ の追加により，ST-ABN から出力される Temporal attention は人手によって修正されたアテンションに近くなる．そのため，ST-ABN は人の知見を考慮した Temporal attention を出力できるようにファインチューニングされる．Temporal attention の学習誤差 $\mathcal{L}_{temp}(\mathbf{x}_i)$ は，ネットワークから出力される Temporal attention $M_t(\mathbf{x}_i)$ と，修正した Temporal attention $M'_t(\mathbf{x}_i)$ の平均二乗誤差を用いて算出する．Temporal attention の学習誤差 $\mathcal{L}_{temp}(\mathbf{x}_i)$ の算出式を式 (6) に示す．

$$L_{temp}(\mathbf{x}_i) = \gamma_t \frac{1}{n} \sum_{j=1}^{n} \left(\{M'_t(\mathbf{x}_i)\}_j - \{M_t(\mathbf{x}_i)\}_j \right)^2 \quad (6)$$

ここで，n は入力フレーム数，$\{M_t(\mathbf{x}_i)\}_j$ はネットワークから出力される j 番目のフレームの Temporal attention，$\{M'_t(\mathbf{x}_i)\}_j$ は j 番目のフレームの修正した Temporal attention を示す．また，Temporal attention の学習誤差は $\mathcal{L}_{att}(\mathbf{x}_i)$ や $\mathcal{L}_{per}(\mathbf{x}_i)$ と比べて誤差の値が小さいため，γ_t を乗算し学習誤差の大きさを調整する．ST-ABN をファインチューニングする際には，ST attention branch と Perception branch のみを対象に学習を行い，入力から特徴マップを抽出する Feature extractor のパラメータは更新しない．

4　評価実験

評価実験では，ベースラインと提案手法による認識精度を比較する．また，Spatial attention と Temporal attention をそれぞれ可視化することで提案手法における判断根拠を示す．

4.1　実験概要

評価実験では，データセットとして動作認識タスクのベンチマークである Something-Something v.2 [26] を使用する．Something-Something v.2 は，人が日用品を扱う基本的な 174 種類の動作で構成されており，学習データには 168,913 本，検証データには 24,777 本，評価データには 27,157 本の動画が含まれる．動画の長さは 2 秒から 6 秒である．また，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 クラスを Temporal attention の修正対象とする．本実験では，修正対象クラスのうち評価時に誤認識した 2396 本の学習サンプルに対する Temporal attention を 74 人で手動修正したものを修正後のアテンションとして使用する．ST-ABN は，SlowFast Networks [12] のベースネットワークである 3D ResNet をバックボーンとして構築する．事前学習モデルとして，ImageNet で学習された 2D ResNet を使用する．学習時には，エポック数は 150，ミニバッチサイズは 64，初期学習率は 0.01 に設定し，75 エポックと 125 エポックで，学習率を 1/10 倍する．

表 1　従来手法と ST-ABN の認識精度比較 [%]

Method	Backbone	Frames	Top-1	Top-5
TSN [14]	BN-Inception	8	27.8	57.6
TRN Multiscale [15]	BN-Inception	8	48.8	77.6
TRN Two-stream [15]	BN-Inception	16	55.5	83.1
CPNet [16]	ResNet-34	24	57.7	84.0
TSM [17]	ResNet-50	8	59.1	85.6
TSM [17]	ResNet-50	16	63.4	88.5
STM [18]	ResNet-50	8	62.3	88.8
STM [18]	ResNet-50	16	64.2	89.8
GST [19]	ResNet-50	16	62.6	87.9
ABM [20]	ResNet-50	16×3	61.3	–
DFB-Net [21]	ResNet-152	16	57.7	84.0
bLVNet-TAM [22]	bLResNet-101	32×2	65.2	90.3
SmallBig$_{En}$ [23]	ResNet-50	24×2×3	64.5	89.1
PEM [24]	ResNet-50	16×2	65.0	–
Zhou *et al.* [25]	3D DenseNet-121	16	62.9	88.0
3D ResNet-50 (Our baseline)	–	32	51.4	80.1
ST-ABN (Ours)	3D ResNet-50	32	**58.6**	**85.5**
3D ResNet-50 (Our baseline)	–	32×2	63.8	89.2
ST-ABN (Ours)	3D ResNet-50	32×2	**64.1**	**89.6**
3D ResNet-101 (Our baseline)	–	32	57.7	82.8
ST-ABN (Ours)	3D ResNet-101	32	**58.0**	**83.2**
3D ResNet-101 (Our baseline)	–	32×2	65.3	90.1
ST-ABN (Ours)	3D ResNet-101	32×2	**65.8**	**90.4**

表 2　人の知見の組み込みによる精度比較 [%]

	修正対象	その他	全体
ST-ABN	20.50	59.76	58.62
ST-ABN + 人の知見	**26.32**	**61.68**	**60.65**

最適化手法は Momentum SGD を使用し，momentum は 0.9，weight decay は 0.0005 に設定する．また，ファインチューニング時には，初期学習率は 0.0001，学習誤差を調整する係数 γ_t は 10 とする．その他の条件は学習時と同様である．

4.2　従来手法との認識精度の比較

表 1 に，Something-Something v.2 における Top-1 accuracy と Top-5 accuracy のベースライン，従来手法及び提案手法との比較結果を示す．表 1 より，ST-ABN は CNN がベースの従来手法と同等以上の精度が獲得できることが確認できる．また，ベースラインのネットワークである 3D ResNet に提案手法を導入することで，認識精度が向上することも確認できる．さらに，フレーム数を増やした場合においても ST-ABN がベースラインよりも高い認識精度を獲得した．このことから，ベースネットワークに ST Attention branch を追加し時間情報と空間情報の重要度を認識に応用することは，認識精度向上に有効であるといえる．

4.3　人の知見導入前後の認識精度の比較

人の知見導入による認識精度の変化を表 2 に示す．このとき ST-ABN は入力フレーム数を 32, ベースネットワークを 3D ResNet-50 としたモデルを使用する．また，修正対象クラスは学習用データと評価用データにおける ST-ABN の認識精度が共に 50 %未満の 8 クラスとする．

表 2 より，Temporal attention を修正したクラスは 5.8 pt 向上したことから，人の知見を組み込むことは認識精度の向上に有効である．また，Temproal attention を修正していないクラスも 1.9 pt 向上したことから，一部のクラスの Temporal attention を修正することで修正対象クラスと似た動作をする，その他クラスの認識精度も向上すると考えられる．

4.4　視覚的説明性の定性的な評価

ST-ABN と人の知見を組み込んだ ST-ABN の Spatial attention と Temporal attention をそれぞれ可視化し，定性的な評価を行う．

4.4.1　ST-ABN の可視化

図 4 に Something-Something v.2 における Spatial attention と Temporal attention の可視化例を示す．図 4 の上から下に向かって入力フレーム，Temporal at-

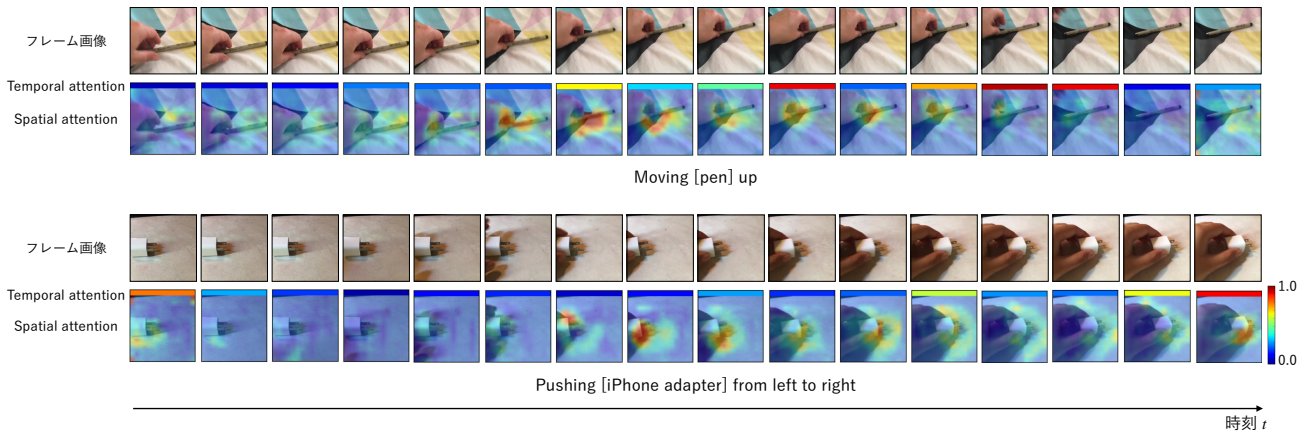

図 4　ST-ABN における Spatial attention と Temporal attention の可視化

tention，Spatial attention，入力動画のラベルを示す．時間情報の重要度を表す Temporal attention は，各フレームに対して出力された重要度をヒートマップの色に変換し，カラーバーに表示することで可視化する．空間情報の重要度を表す Spatial attention は，各フレームに対して出力されたアテンションマップをフレーム画像に重ねたヒートマップとして可視化する．Spatial attention と Temporal attention は共に重要度が高いほど赤く，重要度が低くなるにつれ青色に近づく．

図 4 に示すように，Temporal attention は動きを含むフレームの重要度が高くなっていることが確認できる．Spatial attention は，動作特有の特徴的な空間領域を捉えつつ，動きを含むフレームに対して強く注視していることが確認できる．これらの結果から，ST-ABN は動画像認識において重要となる空間情報と時間情報を同時に考慮した視覚的説明が可能であることがわかる．

4.4.2　人の知見を組み込んだ ST-ABN の可視化

図 5 に ST-ABN への人の知見導入前後の Spatial attention と Temporal attention の可視化結果を示す．図 5 の上から，ST-ABN の可視化結果，入力フレーム，人の知見導入後の ST-ABN の可視化結果，入力動画のラベルを示している．図 5 より，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くなっている．それに対して人の知見を組み込んだ ST-ABN は，カラーバーの色が徐々に変化しており，動きの有無を正確に認識できている．このことから ST-ABN に人の知見を組み込むことで，より適切なアテンションの獲得ができるようになったことがわかる．また Spatial attention において，人の知見を組み込む前の ST-ABN は移動物体を注視しているのに対し，人の知見組み込み後の ST-ABN はフレーム前後で変化した領域を注視している．Something-Something データセットを用いた動作認識では，同じクラスの動画でも扱っている物体が異なるため，物体ではなく動作によって生じた領域のみに注視することは，本来の Spatial attention を獲得できたといえる．

表 3　Spatial attention 反転による比較 [%]

(a) ST-ABN

	Top-1	Top-5
反転前	58.62	85.45
反転後	27.12	52.68

(b) ST-ABN + 人の知見

	Top-1	Top-5
反転前	60.65	86.93
反転後	20.09	43.35

4.5　Spatial attention への影響調査

Temporal attention 修正による Spatial attention への影響を定量的に評価する．評価方法として ST Attention branch で出力される Spatial attention を式 (7) で反転させる．その後反転した Spatial attention を特徴マップに重み付けし推論に用いることによって，認識精度がどれだけ低下するかを調査し，認識に有効な空間情報の重要度が得られていることを確認する．

$$M_{s\ \mathrm{inverse}}(\mathbf{x}_i) = 1 - M_s(\mathbf{x}_i) \tag{7}$$

ここで，$M_s(\mathbf{x}_i)$ は ST Attention branch で出力された Spatial attention，$M_{s\ \mathrm{inverse}}(\mathbf{x}_i)$ は反転後の Spatial attention を示す．

表 3 に Spatial attention の反転による認識精度の比較結果を示す．表 3 より，人の知見組み込み前の ST-ABN は Spatial attention の反転によって Top-1 Accuracy が 31.5 pt，Top-5 Accuracy が 32.8 pt 低下した．それに対し，人の知見組み込み後の ST-ABN は Top-1 Accuracy が 40.6 pt，Top-5 Accuracy が 43.6 pt 低下したため，人の知見組み込み前と比較して認識精度の低下率が大きい．このことから，Temporal attention を修正することによって Spatial attention にも良い影響を及ぼすといえる．

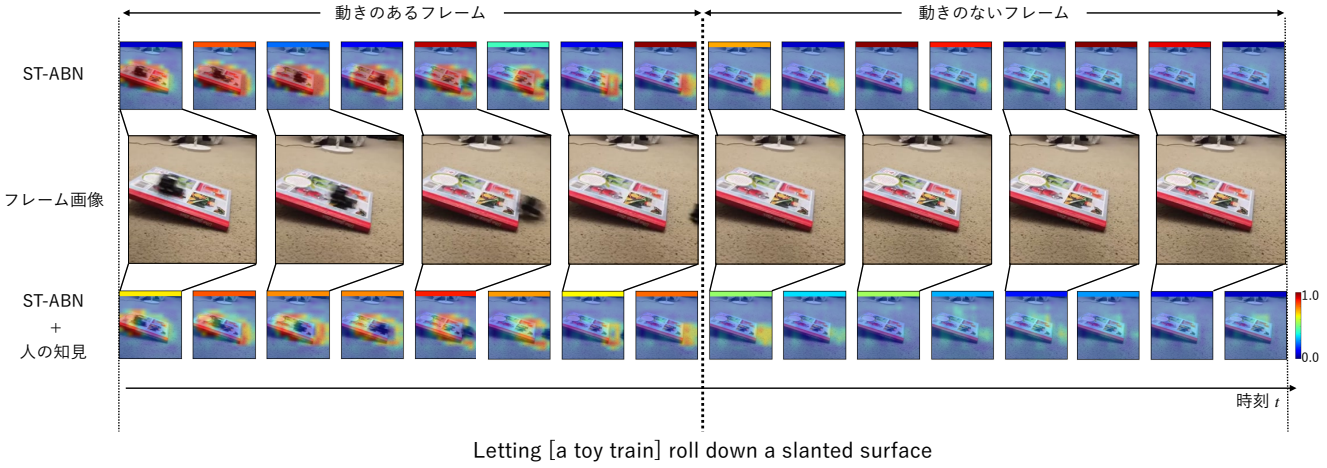

図 5　人の知見導入前後の Spatial attention と Temporal attention の可視化

5　おわりに

本研究では，空間情報と時間情報の両方に対する視覚的説明を行う動画像認識モデルである ST-ABN を提案し，人の知見を組み込んだ．ST-ABN は 3D CNN をベースとしたモデルの推論時に時空間情報の重要度を獲得し，これをアテンション機構に応用することで視覚的説明と認識精度の向上を達成した．また，ST-ABN へ人の知見を組み込むことにより ST-ABN の認識精度向上とアテンションマップの改善を確認した．さらに Spatial attention の定量的な評価により，Temporal attention を介して ST-ABN に人の知見を組み込むことで Spatial attention もより適切な注視領域の獲得が可能となることを確認した．

今後は，Spatial attention を介して ST-ABN に人の知見を組み込むことによる ST-ABN のさらなる高精度化を図るとともに，Transformer ベースのモデルへの人の知見の組み込み方法の検討を行う．

謝辞

本研究は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務 (JPNP20006) の支援を受けたものである．

参考文献

[1] S. Ramprasaath, R., C. Michael, D. Abhishek, V. Ramakrishna, P. Devi and B. Dhruv: “Grad-CAM: Visual explanations from deep networks via gradient-based localization”, 2017 IEEE International Conference on Computer Vision, pp. 618–626 (2017).

[2] 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, pp. 2921–2929 (2016).

[3] H. Fukui, T. Hirakawa, T. Yamashita and H. Fujiyoshi: “Attention branch network: Learning of attention mechanism for visual explanation”, 2019 IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705–10714 (2019).

[4] M. Mitsuhara, H. Fukui, Y. Sakashita, T. Ogata, T. Hirakawa, T. Yamashita and H. Fujiyoshi: “Embedding human knowledge into deep neural network via attention map”, arXiv preprint arXiv:1905.03540 (2019).

[5] C. Feichtenhofer, A. Pinz and A. Zisserman: “Convolutional two-stream network fusion for video action recognition”, 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016).

[6] K. Simonyan and A. Zisserman: “Two-stream convolutional networks for action recognition in videos”, Advances in Neural Information Processing Systems, pp. 568–576 (2014).

[7] S. Ji, W. Xu, M. Yang and K. Yu: “3d convolutional neural networks for human action recognition”, IEEE transactions on pattern analysis and machine intelligence, **35**, 1, pp. 221–231 (2012).

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri: “Learning spatiotemporal features with 3d convolutional networks”, 2015 IEEE International Conference on Computer Vision, pp. 4489–4497 (2015).

[9] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić and C. Schmid: “Vivit: A video vision transformer”, Proceedings of the IEEE/CVF

international conference on computer vision, pp. 6836–6846 (2021).

[10] G. Bertasius, H. Wang and L. Torresani: "Is space-time attention all you need for video understanding?", ICML, Vol. 2, p. 4 (2021).

[11] 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, pp. 770–778 (2016).

[12] C. Feichtenhofer, H. Fan, J. Malik and K. He: "Slowfast networks for video recognition", 2019 IEEE International Conference on Computer Vision, pp. 6202–6211 (2019).

[13] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov: "Improving neural networks by preventing co-adaptation of feature detectors", arXiv preprint arXiv:1207.0580 (2012).

[14] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. Van Gool: "Temporal segment networks: Towards good practices for deep action recognition", European Conference on Computer VisionSpringer, pp. 20–36 (2016).

[15] B. Zhou, A. Andonian, A. Oliva and A. Torralba: "Temporal relational reasoning in videos", European Conference on Computer Vision, pp. 803–818 (2018).

[16] X. Liu, J.-Y. Lee and H. Jin: "Learning video representations from correspondence proposals", 2019 IEEE Conference on Computer Vision and Pattern Recognition, pp. 4273–4281 (2019).

[17] J. Lin, C. Gan and S. Han: "Tsm: Temporal shift module for efficient video understanding", 2019 IEEE International Conference on Computer Vision, pp. 7083–7093 (2019).

[18] B. Jiang, M. Wang, W. Gan, W. Wu and J. Yan: "Stm: Spatiotemporal and motion encoding for action recognition", 2019 IEEE International Conference on Computer Vision, pp. 2000–2009 (2019).

[19] C. Luo and A. L. Yuille: "Grouped spatial-temporal aggregation for efficient action recognition", 2019 IEEE International Conference on Computer Vision, pp. 5512–5521 (2019).

[20] X. Zhu, C. Xu, L. Hui, C. Lu and D. Tao: "Approximated bilinear modules for temporal modeling", 2019 IEEE International Conference on Computer Vision, pp. 3494–3503 (2019).

[21] B. Martinez, D. Modolo, Y. Xiong and J. Tighe: "Action recognition with spatial-temporal discriminative filter banks", 2019 IEEE International Conference on Computer Vision, pp. 5482–5491 (2019).

[22] Q. Fan, C.-F. R. Chen, H. Kuehne, M. Pistoia and D. Cox: "More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation", 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).

[23] X. Li, Y. Wang, Z. Zhou and Y. Qiao: "Smallbignet: Integrating core and contextual views for video classification", 2020 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1092–1101 (2020).

[24] J. Weng, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, X. Jiang and J. Yuan: "Temporal distinct representation learning for action recognition", European Conference on Computer VisionSpringer, pp. 363–378 (2020).

[25] Y. Zhou, X. Sun, C. Luo, Z.-J. Zha and W. Zeng: "Spatiotemporal fusion in 3d cnns: A probabilistic view", 2020 IEEE Conference on Computer Vision and Pattern Recognition, pp. 9829–9838 (2020).

[26] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al.: "The" something something" video database for learning and evaluating visual common sense.", 2017 IEEE International Conference on Computer Vision, Vol. 1, p. 5 (2017).