動画像における時空間情報を考慮した DNN による視覚的説明に関する研究

TR19010 三津原将弘　　指導教授：藤吉弘亘

1. はじめに

画像認識分野において，深層学習の判断根拠を解析する手法の一つに視覚的説明がある．視覚的説明では，深層学習が認識する際に注視した領域をヒートマップで表現した Attention map を解析する．一方，動画像認識では静止画像とは異なり，空間情報だけでなく時間情報も考慮する必要があることから，判断根拠の解析が困難とされてきた．本研究では，視覚的説明を動画像認識に拡張し，動画像における新たな視覚的説明の手法である Spatio-Temporal Attention Branch Network (ST-ABN) 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の獲得を実現する．評価実験では，Something-Something データセットを用いて実験を行い，提案手法により認識性能の向上と空間情報と，時間情報を同時に考慮した視覚的説明が可能となることを示す．

2. 関連研究

従来手法である 3D CNN 及び視覚的説明を用いた手法について述べる．

3D Convolutional Networks　3D CNN の代表的な手法である 3D Convolutional Networks （C3D） [1] は，空間方向に対する 2D の畳み込み処理を時間方向に拡張し，3D 空間に対する畳み込み処理を行うことで時空間の特徴を獲得する．3D 空間に対して畳み込み処理を行うことで，従来の空間方向に対する畳み込み処理を行う 2D CNN では困難であった時系列情報を捉えることができる．

Attention Branch Network　視覚的説明を用いた手法に Attention Branch Network (ABN) [2] がある．ABN は，Attention map を Attention 機構により特徴マップに重み付けすることで認識性能と視覚的な説明性の向上を実現している．

3. 提案手法

本研究では，重要な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な ST-ABN を提案する．

3.1 ST-ABN の構造

ST-ABN は，図 1 に示すように Feature extractor，Spatio-Temporal (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．

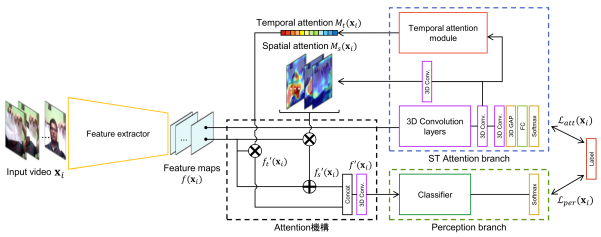

図 1：ST-ABN のネットワーク構造

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，Attention 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．ST-ABN は，式 (1) のように ST Attention branch の学習誤差 $\mathcal{L}_{att}$ と Perception branch の学習誤差 $\mathcal{L}_{per}$ を用いて学習する．クラス識別誤差は，Softmax 関数とクロスエントロピー誤差を用いて算出する．

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

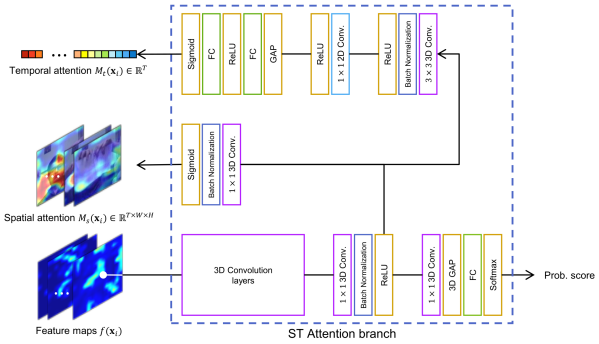

図 2：ST Attention branch の構造

3.2 Spatio-Temporal Attention branch

ST Attention branch では，図 2 に示すように Feature extractor から出力された特徴マップを用いて，空間情報と各フレームに対する重要度を獲得し，Global Average Pooling (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を獲得する．

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 枚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 Attention map 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．

3.3 Attention 機構

Spatial attention $M_s(\mathbf{x}_i)$ は，式 (2) 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，Attention map を効率的に認識に反映させることができる．

$$f_s'(\mathbf{x}_i) = (1 + M_s(\mathbf{x}_i) \cdot f(\mathbf{x}_i)) \tag{2}$$

Temporal attention $M_t(\mathbf{x}_i)$ は，式 (3) より特徴マップに乗算することで重み付けを行う．

$$f_t'(\mathbf{x}_i) = M_t(\mathbf{x}_i) \cdot f(\mathbf{x}_i) \tag{3}$$

Spatial attention と Temporal attention でそれぞれ重み付けした特徴マップは，式 (4) よりチャネル方向に結合する．その後，結合した特徴マップを畳み込み層に入力して統合する．

$$f'(\mathbf{x}_i) = \mathrm{conv}\left(f_s'(\mathbf{x}_i) \odot f_t'(\mathbf{x}_i)\right) \tag{4}$$

この統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．

4. 評価実験

評価実験では，Something-Something データセット [3] の V1 と V2 を用いてベースラインと提案手法との認識精度を比較する．Something-Something データセットは，大規模な動作認識用のデータセットであり，人が日用品を扱う 174 種類の基本的な動作を認識する．提案手法は，ベースラインである 3D ResNet-50 をベースに構築する．入力

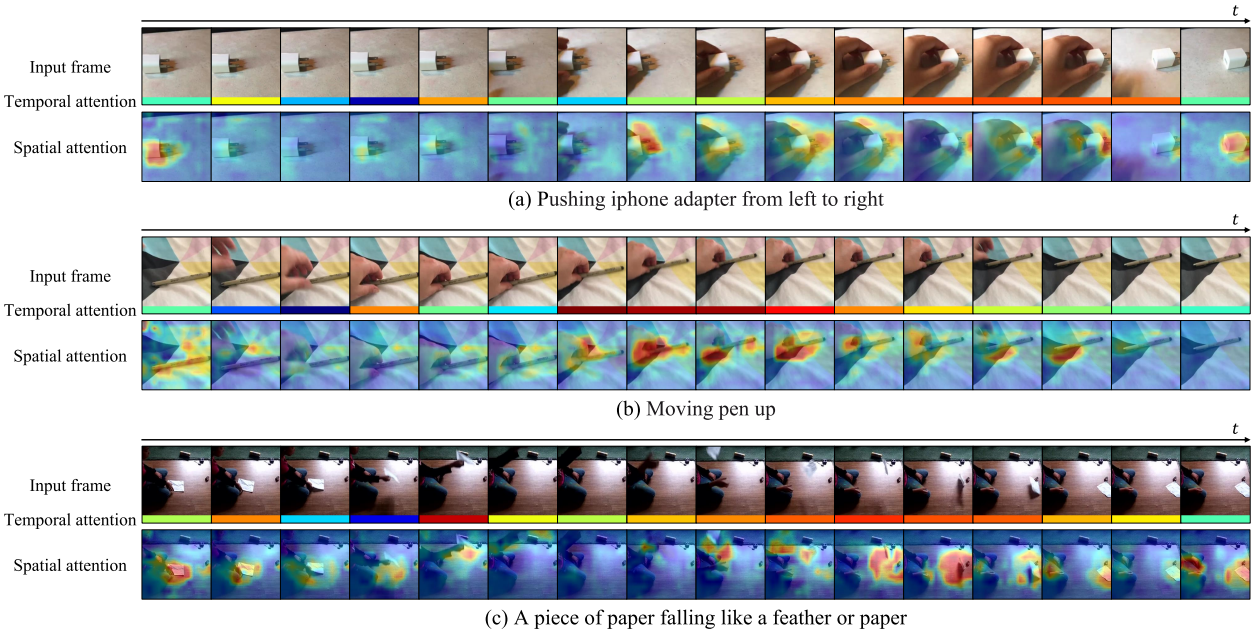

図 3：Spatial attention と Temporal attention の可視化結果

表 1：ベースラインと提案手法との認識精度の比較結果 [%]

Model	Frames	Something-Something V1		Something-Something V2	
		Top-1	Top-5	Top-1	Top-5
3D ResNet-50 (ベースライン)	32	45.2	74.5	55.6	81.6
3D ResNet-50 + 提案手法	32	**45.9**	**75.4**	**56.6**	**82.0**
3D ResNet-50 (ベースライン)	32+32	52.9	81.5	63.8	89.2
3D ResNet-50 + 提案手法	32+32	**53.3**	**82.0**	**64.1**	**89.6**

するフレーム数は，32 フレームと 64 フレームを入力した場合で認識精度を比較する．64 フレームを入力する場合は，32 フレームの動画を 2 つ入力し，2 つの動画に対するスコアを平均することで最終的な認識精度を算出する．

4.1 ベースラインとの認識精度の比較

表 1 にベースラインと提案手法との認識精度の比較結果を示す．表 1 に示すように提案手法がベースラインよりも高い認識精度を獲得した．さらにフレーム数を増やした場合においても提案手法がベースラインよりも高い認識精度を獲得した．

4.2 Attention の定性的な評価

図 3 に Something-Something データセット V2 における Spatial attention と Temporal attention の可視化結果を示す．図 3 の上から下に向かって入力フレーム，Temporal attention，Spatial attention，入力動画のクラスを示している．Temporal attention では，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することで可視化しており，重要度が高いほど赤く，重要度が低いほど青くなる．図 3 に示すように，動きを含むフレームの重要度が高くなることが確認できる．

Spatial attention では，各フレームに対して出力された Attention map をヒートマップとして可視化する．Spatial attention は，動作特有の特徴的な空間領域を捉えつつ，動的な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，提案手法により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．

4.3 Attention の有効性の評価

Spatial attention と Temporal attention の有効性を定量的に評価する．評価方法として Spatial attention と Temporal attention を反転させて推論を行う．そして，反転する場合と反転しない場合で認識精度がどれだけ低下するかを調査し，認識に有益な空間情報と時間情報に対する重要度が得られていることを確認する．

表 2 に Something-Something データセットにおける Spatial attention と Temporal attention の反転による認識精度の比較結果を示す．表 2 に示すように，Spatial attention のみの反転と Temporal attention のみの反転により認識精度が低下することが確認できる．さらに，Spatial attention と Temporal attention を反転させることで大幅に認識精度が低下することから認識に有益な重要度が得られていることが確認できる．

表 2：Attention の反転による認識精度の比較結果 [%]

Attention		Something. V1		Something. V2	
Spatial	Temporal	Top-1	Top-5	Top-1	Top-5
		45.9	75.4	56.6	82.0
✓		36.2	66.8	19.1	39.7
	✓	28.3	57.4	45.5	75.6
✓	✓	**22.7**	**50.0**	**12.4**	**29.0**

5. おわりに

本研究では，ST-ABN を提案し，動画像認識における空間情報と時間情報を同時に考慮した視覚的説明を実現した．ST-ABN は，3D CNN がベースであるモデルの推論時における空間情報と時間情報に対する重要度を獲得し，Attention 機構に応用することで視覚的な説明性と認識性能の向上が可能となる．今後は，提案手法を分類以外の他の動画像認識タスクへ応用することを検討する．

参考文献

[1] D. Tran, *et al.*, “Learning Spatiotemporal Features with 3D Convolutional Networks”, ICCV, 2015.

[2] H. Fukui, *et al.*, “Attention Branch Network: Learning of Attention Mechanism for Visual Explanation”, CVPR, 2019.

[3] R. Goyal, *et al.*, “The“Something Something” Video Database for Learning and Evaluating Visual Common Sense”, ICCV, 2017.

研究業績

[1] 三津原将弘 等, “Attention map を介した人の知見の組み込み”, 第 22 回 画像の認識・理解シンポジウム, 2019.

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