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．

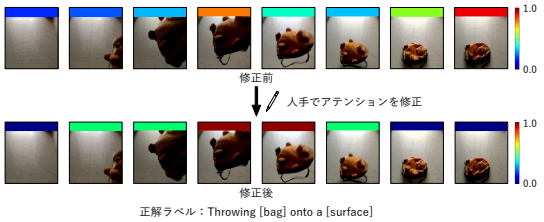

図 1：Temporal attention 修正例

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．

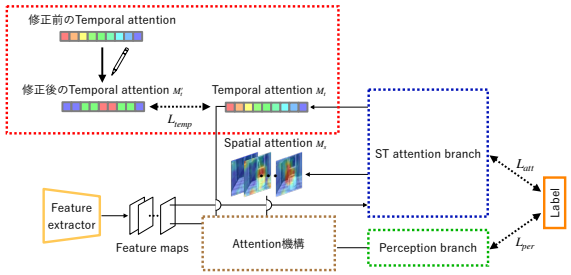

図 2：人の知見を導入した ST-ABN

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．

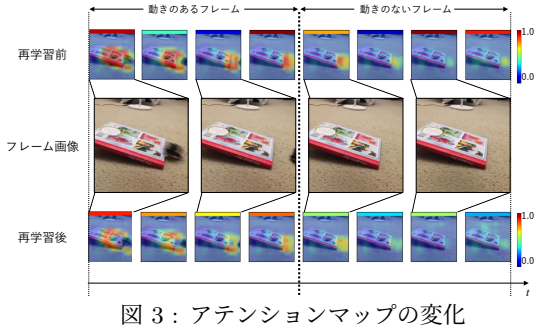

図 3：アテンションマップの変化

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.