1Q-01

行動検出タスクにおける人物特徴量に注目した自己教師あり学習

吉田 舜　山崎 俊彦
東京大学

1．はじめに

近年、コンピュータビジョンでは大量のデータを用いたデータドリブンな学習方法が用いられている。この大量のデータに対して人間がラベルを付与するとなると非常にコストがかかるため、ラベルを使用せずに表現を学習する自己教師あり学習（Self-Supervised Learning）の研究が盛んに行われており、動画認識の分野においてもそれは例外ではない。現に動画認識において代表的なタスクである行動分類においては、画像ベースの自己教師あり学習の手法を動画用に拡張し適用することによって認識精度の向上を達成している。

本研究では動画認識のタスクの一つである行動検出のデータセットを用いた自己教師あり学習の手法を提案する。

2．動画認識

動画認識には行動分類と行動検出と呼ばれるタスクがある。行動分類は入力動画から一つの行動ラベルを推定することを目的としている。それに対し行動検出は入力動画のうち対象となるフレーム（キーフレーム）に映る複数の人物を検出しその人物に付与された行動ラベルを推定することが目的である。行動分類では Kinetics データセット[1]、行動検出では AVA データセット[2]が用いられることが多い。

行動検出の基本的な流れを以下に示す（図 1 参照）。まずキーフレーム前後のフレームをバックボーンネットワークに通して特徴量マップを計算する。それと同時にキーフレームに対して Faster R-CNN のような物体検出器を用いて人物領域を求め、特徴量マップと人物領域から RoIPool や RoIAlign の手法を用いて人物特徴量を抽出し、その人物特徴量から行動を推定するというものである。

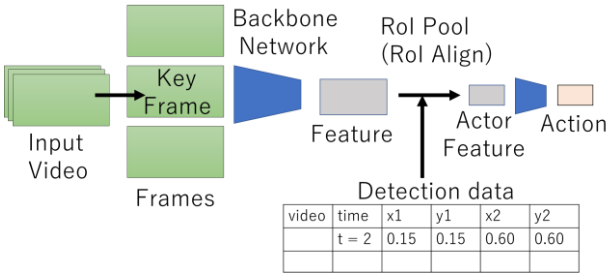

図 1　行動分類の流れ

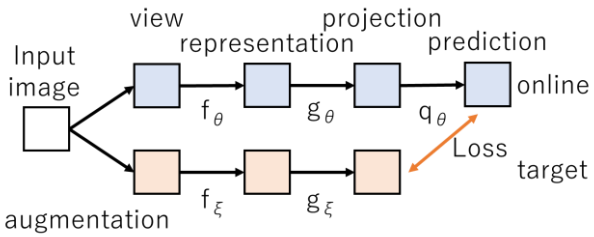

図 2 BYOL の全体図

3．自己教師あり学習

自己教師あり学習は正解ラベルなしのデータから教師データを生成することで表現の学習を行うものである。近年では BYOL[3]と呼ばれる手法が自己教師あり学習の中でも高い成績を達成している。BYOL の手法を以下に示す(図 2 参照)。

(1) online network と target network と呼ばれるネットワークを用意する。学習するのは online のみであり、target のパラメータ ξ は online のパラメータ θ の指数移動平均を用いて更新する。
(2) 一つの画像から augmentation によって異なる二つの view を生成し、それぞれのネットワークに通し、target 特徴量を予測するように online を学習する。
(3) 学習の後 f_θ を最終的なモデルとする。

動画認識における自己教師あり学習についての研究では、BYOL を含む様々な画像ベースの自己教師あり学習の手法が動画データに対しても有効であることを実証している[4]。この研究で

Self-supervised learning using action detection datasets
Shun Yoshida, Toshihiko Yamasaki, The University of Tokyo

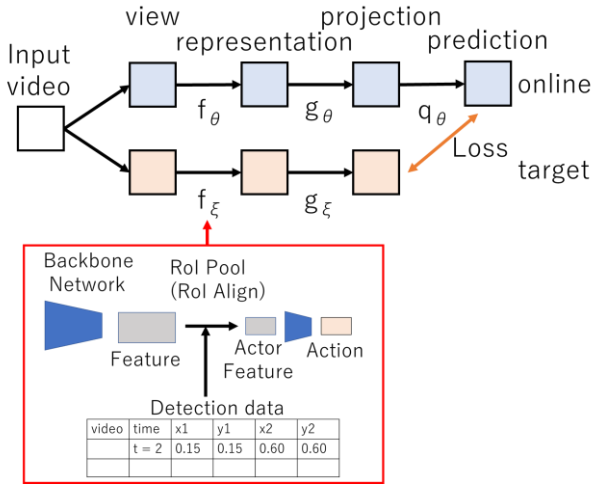

図 3 提案手法のモデル

は Kinetics データセットをラベルなしで事前学習を行った後に AVA データセットで fine-tune することで、ランダムに初期化した状態から AVA データセットを学習するよりも高い精度を達成している。

4．提案手法

本研究では AVA データセットのような複数の人物が様々な行動をしている長時間の動画からラベルなしでより良い表現方法を事前学習することが目標である。既存手法では入力動画をバックボーンネットワークに通して得られた特徴量をそのまま自己教師あり学習で使用していたのに対し、本研究では BYOL をベースに f_θ の部分に人物特徴量を抽出するネットワークを用いた自己教師あり学習モデルを提案する(図 3 参照)。

5．実験と考察

AVA データセットは映画の 15 分から 30 分までのデータにラベルを付与したものであり、本研究ではこの映画の 0 分から 15 分までのラベルなしデータを用いた事前学習を行うことによる認識精度の変化、および提案手法の性能の確認の確認を行った。具体的にはランダムに初期化した状況から AVA データセットを学習した場合と、入力動画をバックボーンネットワークに通して得られた特徴量をそのまま自己教師あり学習に用いて事前学習を行った後に AVA データセットを学習した場合と、提案手法のモデルのように人物特徴量を用いて事前学習を行った後に AVA データセットを学習した場合で比較を行った。

本実験では PySlowfast[5]のコードを使用し、

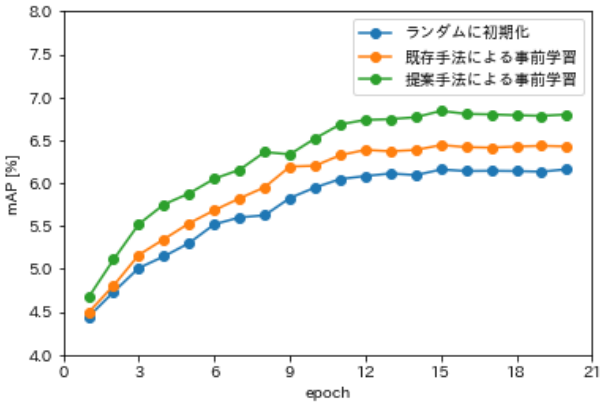

図 4 手法間の比較

バックボーンとなるネットワークは Slow を用い、事前学習と AVA の学習はそれぞれ 20 エポック行い、人物検出データは Detectron2 [6]の faster R-CNN を用いて作成した。また、評価指標は行動検出において用いられている指標である mean Average Precision (mAP)を使用した。

実験結果を以下に示す(図 4 参照)。20 エポック学習した時点で mAP はそれぞれ、ランダムに初期化した場合が 6.16[%]、既存手法による事前学習を用いた場合が 6.43[%]、提案手法による事前学習を用いた場合が 6.80[%]となった。既存手法と比較して提案手法における認識精度が上昇していることが読み取れる。このことから AVA データセットのような複数の人物が様々な行動をしている長時間の動画から事前学習する場合では人物特徴量に注目した自己教師あり学習を行うことで、より良い表現方法を学習できると考えられる。

6．まとめ

本稿では行動検出のデータセットにおける人物特徴量を用いた自己教師あり学習の手法を提案し、既存手法と比較してより認識精度が上昇するという結果が得られた。今後は異なるバックボーンネットワークを用いた場合や異なる自己教師あり学習のフレームワークを用いた場合の性能評価を実施する予定である。

参考文献

[1] Will Kay, et al., The kinetics human action video dataset. arXiv:1705.06950, 2017
[2] C. Gu, et al., AVA: A video dataset of spatio-temporally localized atomic visual actions. CVPR, 2018
[3] J. B. Grill, et al., Bootstrap your own latent: A new approach to self-supervised learning. NeurIPS, 2020
[4] C. Feichtenhofer, et al., A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. CVPR, 2021
[5] https://github.com/facebookresearch/SlowFast
[6] https://github.com/facebookresearch/detectron2