5ZB-08

センサベースの行動認識におけるセンサデータを用いない事前訓練

清水椋右† 長谷川達人‡
福井大学大学院† 福井大学大学院‡

1. はじめに

転移学習は，大規模データセットの活用により，認識精度向上や学習の高速化を見込める技術である．画像認識分野では，研究者らの努力により大規模データセットが整備されている[1]．行動認識分野では大規模データセット構築は困難であるため，大規模画像データセットを行動認識に転用できることが望ましい．本研究ではImageNet[1]で事前訓練された画像認識 CNN モデルのパラメータを 1 次元に圧縮し，行動認識モデルに変換する手法を提案する．

2. 提案手法

画像認識分野にて提案されている著名な CNN モデルは ImageNet などの大規模データセットで事前訓練されたモデルとして公開されていることが多い．学習済みモデルのパラメータを行動認識用モデルへと変換を行うために，本研究では畳み込み層の計算方法に着目し，2 次元畳み込み(2D-Conv)のパラメータを 1 次元畳み込み(1D-Conv)のパラメータへ変換を行う．図 1 にパラメータ圧縮の概要を示す．図 1 上部のような 2D-Conv の畳み込み演算を考える．入力データ，畳み込みのパラメータがともに3×3である場合，縦にi番目，横にj番目の入力，パラメータをそれぞれx_{ij}，w_{ij}，出力をzとすると，通常の 2D-Conv の演算は以下のように表せる．

$$z = \sum_i \sum_j x_{ij} w_{ij} \tag{1}$$

図 1 に示すように，2D-Conv のカーネルのパラメータを縦方向(i方向)に和や平均をとったものを 1D-Conv のカーネルのパラメータとして扱うことで，2D-Conv のパラメータを流用する． パラメータの和をとるものを summation (sum)，平均をとるものを average (ave)とする．

sum, ave の畳み込み演算を考える．sum, ave は

Pretraining without sensor data for sensor-based human activity recognition

†Ryosuke Shimizu, Graduate School of Engineering, University of Fukui
‡Tatsuhito Hasegawa, Graduate School of Engineering, University of Fukui

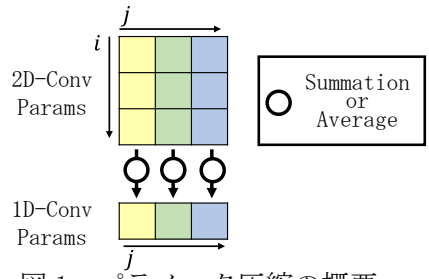

図 1：パラメータ圧縮の概要

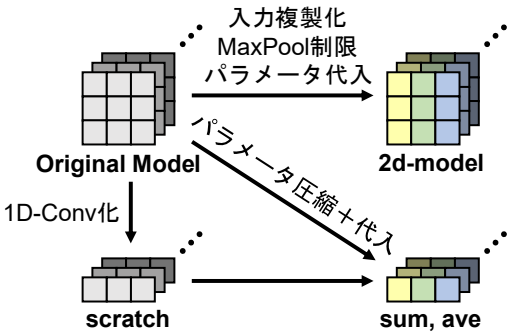

図 2：検証モデルの概要

1D-Conv であるため，入力をx_j，出力をそれぞれz_{sum}，z_{ave}とする．sum の畳み込み演算は$z_{sum} = \sum_j x_j(w_{1j} + w_{2j} + w_{3j})$であり$x_j = x_{1j} = x_{2j} = x_{3j}$の場合(センサデータを縦方向に複製し 2D-Conv に入力する場合)，式(1)と等価である．ave の畳み込み演算は$z_{ave} = \sum_j x_j \frac{w_{1j}+w_{2j}+w_{3j}}{3}$であり，$x_j = 3x_{1j} = 3x_{2j} = 3x_{3j}$の場合，式(1)と等価である．

3. 検証実験

提案手法の有効性を示すために，事前訓練前のモデルと事前訓練後のモデルで行動認識タスクを解き，比較検証を行う．今回は HASC データセットを使用し検証を行う．HASC はスマートフォンなどによって 6 種の日常行動(静止，歩行，走行，スキップ，階段上り，階段下り)が測定されたデータセットである．180 名のデータのうち訓練用として 5 名，10 名，20 名のデータを用い，検証用として 36 名のデータを用いる．最適化手法は Adam，学習率は1×10^{-4}で 350 epoch 学習を行う．垂直反転，チャネルシャッフルをデータ拡張として訓練時に適用している．評価指標には 25 試行の平均の Accuracy を用いる．

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

今回は著名かつシンプルなモデルである VGG16[2]，ResNet50[3]を対象として検証を行う．特徴抽出器の影響を調査するため VGG16，ResNet50 ともに特徴抽出器から得られた出力を Global Average Pooling を用いて 1 次元の特徴マップに変換したのち，1 層の全結合層によってクラス分類を行うようにしている． 検証に使用するモデルの概要を図 2 に示す．比較対象として，各モデルの 2D-Conv を 1D-Conv に変換し，訓練済みパラメータを流用しない scratch，入力であるセンサデータを縦方向に複製し 2 次元モデルで学習を行う 2d-model の 2 つのモデルを作成した．scratch 以外のモデルは ImageNet で事前訓練済みのパラメータを用いている．2d-model は単純にセンサデータを縦方向に複製して横長の画像のように扱うのではなく，各畳み込み層の入力を縦方向に複製し，1 次元の特徴を出力するようにしている．また，Max Pooling を横方向のみに適用するように制限を掛けている．これにより 2d-model は計算上では実質 sum と同等になる．

4. 結果・考察

VGG16，ResNet50 の結果をそれぞれ表 1，表 2 に示す．表 1，2 より，すべての条件において 2d-model, sum, ave は scratch の精度を上回っていることがわかる．したがって，センサベースの行動認識において ImageNet での事前訓練が有効であると考えられる．センサベースの行動認識はユーザや計測部位，計測機器によるドメインの差が大きく，事前訓練するデータセットによっては精度が悪化することが知られている[4]．検証によって，画像とセンサデータというドメインに大幅な差異がある場合においても事前訓練が有効であることが判明した．これは ImageNet がデータ数，クラス数共に充実しており，事前訓練済みのモデルは様々な特徴を獲得しやすいパラメータを有しているためだと考えられる．

また，訓練に使用する人数が少ないほど事前訓練によって精度向上していることがわかる．これより，訓練データ数が少数の場合に特に有効に働くと考えられる．2d-model と sum, ave を比較すると，ほとんどの条件において 2d-model が最高精度であることがわかる．2d-model は他のモデルと比較して，VGG16 では約 3 倍，ResNet50 では約 1.5 倍の数のパラメータを有しており，パラメータ数による精度向上であると考えられる．

sum と ave を比較した際に，ave のほうが高精度であることがわかる．理論上 sum は 2 次元モデルのパラメータを 1 次元へとロスなく圧縮できていると考えられる．CNN は層を重ねることで局所的な特徴から大局的な特徴を獲得していく．sum は計算上 2d-model と同等のため，畳み込み層のつながりを保持したまま 1 次元に圧縮可能であり，ave は畳み込み層のつながりが崩れるため，sum のほうが高精度になると予測できる．しかし ave のほうが sum より高精度であるため，この予測は異なると考えられる．今回の検証では sum と ave の精度の違いの原因を考察することができなかったため，議論の余地がある．

表 1：VGG16 の結果

	Encoder Params	Encoder FLOPs	# of Train persons 5	10	20
scratch	**4.92M**	**144.85M**	55.05	65.02	76.81
2d-model	14.72M	434.55M	**64.10**	**71.64**	**82.00**
sum	**4.92M**	**144.85M**	61.57	70.72	79.60
ave	**4.92M**	**144.85M**	**62.65**	**72.78**	**80.89**

表 2：ResNet50 の結果

	Encoder Params	Encoder FLOPs	# of Train persons 5	10	20
scratch	**15.96M**	**199.93M**	45.05	58.56	72.28
2d-model	23.53M	293.76M	**61.28**	**68.77**	**79.36**
sum	**15.96M**	**199.93M**	58.71	66.65	77.46
ave	**15.96M**	**199.93M**	**59.58**	**67.48**	**79.16**

5. まとめ

本研究では大規模画像データセットである ImageNet で事前訓練した 2 次元モデル VGG16, ResNet50 のパラメータを 1 次元モデル用に圧縮することで，センサデータを用いない事前訓練手法の提案を行った．HASC データセットを用いた検証により，画像データセットでの事前訓練によって行動認識精度向上を見込めることが判明した．また，行動認識を行う訓練用データの人数が少数であるときに特に有効であることが判明した．行動認識分野は，画像認識分野と比較して大規模なデータセットを構築することが困難であるため，本手法は実運用を考えると特に有効であると考えられる．今後は sum と ave 間の精度の違いの解明，他のデータセットに対しても有効であるのかの検証を行っていく．

参考文献

[1] Deng, J., Dong, W., Socher, R., et al.: ImageNet: A large-scale hierarchical image database, *CVPR*, pp. 248-255 (2009)
[2] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *ICLR* (2015).
[3] He, K., Zhang, X., Ren, S., et al.: Deep Residual Learning for Image Recognition, *CVPR*, pp. 770-778 (2016).
[4] Gjoreski, M., Kalabakov, S., Luštrek, M., et al.: Cross-dataset deep transfer learning for activity recognition, *UbiComp/ISWC*, pp. 714–718 (2019).

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.