大規模計算機なしで戦う深層学習研究

長谷川 達人（福井大学）

1 はじめに

2012年に開催された画像認識コンペ[1]でAlexNetが2位に大差で勝利したことを皮切りに，深層学習が盛んに研究されるようになった．従来は入力データから分野の知識に基づいて設計された特徴量を計算し，SVMやRandom Forest等の機械学習アルゴリズムで特徴量から出力ラベルの変換を近似していた．深層学習は特徴量設計自体をモデル内で実現するEnd-to-Endな学習手法であり，入力と出力ラベルだけを与えれば特徴抽出も含めて自動で変換を近似する．このように入力から特徴表現を如何にして獲得するかを探求することが深層学習研究の醍醐味とも言える．

2020年頃まではConvolutional Neural Network (CNN)のモデル構造を工夫することで優れた特徴表現の獲得を目指す研究が盛んに行われていた．AlexNetから，VGG，Inception-v3，ResNet，SE-Net，そしてEfficientNetで一つの区切りとなった．ここ1，2年の動向としては，自然言語処理分野で大躍進を遂げたTransformerを画像認識（Vision Transformer; ViT）や音声認識（Conformer）に転用する研究が行われていたり，Multi Layer Perceptron（MLP）の可能性が見直されていたりする（MLP-Mixer[1]等）．

モデル構造の設計が進む傍らで，様々な問題設定の解決を図る手法も登場している．現実世界の問題設定ではラベルありデータよりもラベルなしデータの方が大量に収集しやすく，これを併用して精度向上を図る半教師あり学習が盛んに研究されている．ラベルありデータで訓練したモデルで，ラベルなしデータに擬似ラベルを付与し，疑似ラベルを含めた全データでモデルを改めて訓練するといったPseudo Labelをベースに，Dual Student，やNoisy Student，近年ではMeta Pseudo Labels等が提案されている．類似の研究分野として，Contrastive Learningによる自己教師あり学習も盛んに研究されている（SimCLR等）．

ここで，これらの訓練に要するコストに着目する．Mikamiら[2]によるとResNet論文[3]ではTesla P100の8並列機を用いてImageNetの訓練に29時間かかっている．Tesla P100は1基およそ100万円であり8並列の時点で既に筆者のような地方大学の一研究者では土俵に上がる難易度が高い．新しいモデルだとImageNet-21kの訓練にかかる時間はViTで230，MLP-Mixerで410 TPUv3-core-daysである．現在クラウド環境でTPUが使えるとはいえ410日×24時間×1ドル/TPUv3-hourで概算100万円である．

本稿では，筆者が専門とする行動認識分野における研究と最新の深層学習手法を一部紹介する．筆者のように大規模なデータセットや膨大な計算機環境を容易に準備できない研究者が，深層学習研究で如何に立ち位置を獲得しているのかを述べ，今後の深層学習研究及び，深層学習を用いた分野横断研究の発展に寄与することを目的とする．

[1]ImageNet Large Scale Visual Recognition Challenge（ILSVRC）：http://www.image-net.org/challenges/LSVRC/

2 行動認識と深層学習

行動認識はスマートフォンやウェアラブルデバイスに搭載された様々なセンサを用いて人間の動作を計測し，着用者がどのような行動を行っているのかを認識する技術である．サンプリング周波数100Hz程度で計測された3軸センサ波形を主に扱うため画像や音声と比べデータサイズが小さい．行動認識分野でも深層学習の応用は盛んに研究されており，特に行動認識独自の課題解決を図る研究が多い．

Ahmadら[4]は，慣性センサから観測される波形データをRecurrence Plot等を用いて様々な形式の画像に変換し，ResNetで画像からの特徴抽出を図っている．CNNは近傍情報の畳み込み処理を多段に組み合わせることで画像から特徴表現を獲得するが，これが波形データに対して上手く動作するとは限らない．したがって，彼らの研究は波形データから如何にして優れた特徴表現を獲得するのかを探求する研究であると言える．Baiら[5]も時間方向やモダリティ方向にどのようにデータを結合するかを複合的に検証し，アンサンブルする表現学習手法を提案している．

行動認識は個人ごとに行動波形が大きく異なることがあり個人適応手法も研究されている．Gongら[6]は，メタ学習手法のMAMLを行動認識問題に落とし込んだMetaSenseを提案している．Stisenら[7]は，計測機器ごとにセンサ特性が異なるという問題提起を行い，クラスタリングにより精度低下を改善する手法を提案している．筆者の研究[8]では計測者間でサンプリング周波数が異なる可能性を指摘し，敵対的訓練により精度低下を改善する手法を提案している．他にも，異なるモダリティ間で知識転移を行う手法や，訓練時に未知の行動を予測する手法など，様々な問題提起と解決策が提案されている．このように，行動認識分野においては分野独自の課題に対して深層学習を如何に工夫するのかという点で研究が進められていることが多い．

3 MLP-Mixerを用いた行動認識

行動認識研究に対して最新の深層学習技術の転用を図る一事例として，2021年5月にGoogleの研究チームにより提案されたMLP-Mixer[1]を行動認識に適用する．画像認識におけるMLP-Mixerは入力画像をn × nのパッチに分

解し，それぞれ全結合層を経て Mixer-Layer に入力される．Mixer-Layer では各パッチに対して Layer normalization（LN）を行い，転置した上でシンプルな 2 層の MLP に入力され，その後更に転置，LN を経て再び 2 層の MLP に入力される．この Mixer-Layer が従来の CNN の ConvBlock 相当の働きをするため，Mixer-Layer はハイパーパラメータに基づいて多段に積み重ねられる．最終的に出力層を追加することで CNN 相当の働きをする MLP が実現できる．

本研究では，行動認識のベンチマークデータセット HASC を用いて MLP-Mixer の有効性検証実験を行った．MLP-Mixer は大規模な訓練データを要すると言われているが，HASC から可能な限り多くのデータとして 130 名の 6 行動（停止，歩行，走行，スキップ，階段上り，下り）約 20,000 件を訓練データに 30 名分を検証データとした．データは 256 サンプルで時系列分割された x，y，z 軸の加速度センサ値であるため 3 × 256 の書式となる．MLP-Mixer は画像を n × n のパッチに分割していたが，行動認識では波形データのため n 個のパッチに分割することとした．

MLP-Mixer のハイパーパラメータ探索結果を図 1 に示す．横軸は対数表記であり，depth は Mixer-Layer の繰り返し数，dim はパッチ後の埋め込み次元数，patch_size はパッチの大きさ，token_dim は Token-Mixier 時の中間層のユニット数，channel_dim は Channel-Mixer 時の中間層のユニット数である．それぞれ depth が 6，dim が 32，patch_size が 32，token_dim が 128，channel_dim が 512 をデフォルト値として，単一のパラメータのみチューニングした．結果として，depth が 2，dim が 64，patch_size が 32，token_dim が 32，channel_dim が 512 付近が良いことがわかる．

チューニング結果を踏まえ，複数手法間で精度比較を 10 試行した結果を図 2 に示す．比較対象として，MLP-Mixer の転置処理をなくした MLP-Patch，更にパッチ処理をなくしたシンプルな 8 層 MLP，8 層の VGG 構造の CNN を VGG として採用した．比較の結果，パッチ処理による精度向上は見込めないが，転置を含むことで MLP-Mixer は 87.8%の平均推定精度を達成した．中央値で見ると MLP-Mixer は 87.6%，VGG は 89.7%となり CNN に及ばないという結果となったが，試行間のばらつきは軽減できることが確認された．以上より，近年 MLP の活躍が見直されつつあるものの精度面では改善が必要なことが示唆された．

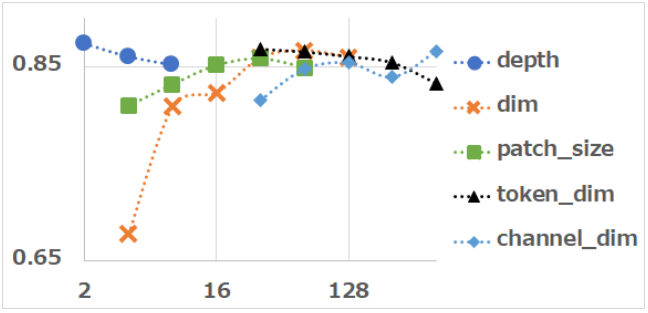

図 1: MLP-Mixer のハイパーパラメータチューニング

4 おわりに

本稿では近年の深層学習研究と行動認識分野における深層学習研究を紹介した．その上で，最新モデルである MLP-Mixer を行動認識に適用する事例を紹介した．本事例のように，様々な最新研究の技術を転用していくだけでなく，各分野に適用させる工夫を検討していくことが重要であると考える．また，分野固有の課題を最新技術の応用により改善することも望まれる．深層学習には多大な計算資源が重要ではあるが，これがなくても様々な分野で立ち位置を獲得する方法はある，と信じたい．

参考文献

[1] Tolstikhin, I. and et al.: MLP-Mixer: An all-MLP Architecture for Vision (2021).

[2] Mikami, H. and et al.: Massively Distributed SGD: ImageNet/ResNet-50 Training in a Flash (2018).

[3] He, K. and et al.: Deep Residual Learning for Image Recognition, *Proc. of the CVPR*, pp. 770–778 (2016).

[4] Ahmad, Z. and Khan, N.: Inertial Sensor Data to Image Encoding for Human Action Recognition, *IEEE Sensors Journal*, Vol. 21, No. 9, pp. 10978–10988 (2021).

[5] Bai, L. and et al.: Adversarial Multi-View Networks for Activity Recognition, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 4, No. 2 (2020).

[6] Gong, T. and et al.: Adapting to Unknown Conditions in Learning-based Mobile Sensing, *IEEE Transactions on Mobile Computing*, pp. 1–16 (2021).

[7] Stisen, A. and et al.: Smart Devices Are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition, *Proc. of the SenSys* (2015).

[8] Hasegawa, T.: Smartphone Sensor-Based Human Activity Recognition Robust to Different Sampling Rates, *IEEE Sensors Journal*, Vol. 21, No. 5, pp. 6930–6941 (2021).

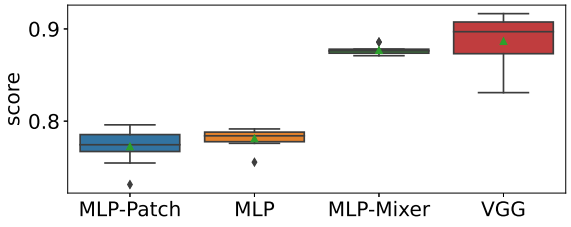

図 2: 各モデルの精度比較