Vol.2021-MBL-100 No.25
Vol.2021-UBI-71 No.25
Vol.2021-CDS-32 No.25
Vol.2021-ASD-21 No.25
2021/9/3

センサベースの人間行動認識における深層学習アンサンブル手法に関する考察

長谷川 達人[1,a)]　近藤 和真[1]

概要：センサを用いた人間行動認識において深層学習を用いた手法が数多く提案されている．中でも，深層学習とアンサンブル学習を併用する手法は強力な成果を発揮している．一方，アンサンブル学習を行うにはデータの分割や複数モデルを学習するなどの様々な手続きを要し，手間と計算コストがかかる．本研究では，行動認識を対象に深層学習のアンサンブル手法を解析することを通じて，単一モデルを End-to-End で訓練するだけでアンサンブルモデルと同等の推定精度を実現する手法の実現可能性を考察する．

キーワード：人間行動認識，アンサンブル学習，深層学習

Consideration of Ensemble Deep Learning Method for Sensor-based Human Activity Recognition

Tatsuhito Hasegawa[1,a)]　Kazuma Kondo[1]

Abstract: Many methods based on deep learning have been proposed for sensor-based human activity recognition. Methods that combine deep learning and ensemble learning have especially shown powerful results. On the other hand, ensemble learning requires various procedures, such as data partitioning and training multiple models, which are time-consuming and computationally expensive. In this study, we analyze ensemble methods of deep learning for activity recognition, and examine the feasibility of a method that achieves estimation accuracy equivalent to that of ensemble models by simply training a single model in an end-to-end manner.

Keywords: Human Activity Recognition, Ensemble Learning, Deep Learning

1. はじめに

スマートフォンやウェアラブルデバイスなどのスマートデバイスの普及に伴いユーザの動作をセンシングすることが容易となってきている．多くのデバイスは慣性計測装置を内蔵しており，加速度や回転を時系列センサデータとして計測可能である．センサ情報からユーザの行動を予測することを行動認識と呼び，これが実現できることでライフログ [1] などの様々な応用サービスが展開可能となる．将来の充実したサービス応用に向け，より詳細な行動情報を，より正確に認識できる技術開発が求められている．

[1] 福井大学大学院工学研究科
Graduate School of Engineering, University of Fukui
[a)] t-hase@u-fukui.ac.jp

行動認識はスマートデバイスの普及以前から活発に研究されている．Aminian ら [2] や Bao ら [3] は被験者に小型の加速度センサを着用させ，計測値から被験者の行動を推定する手法を提案した．Aminian らの手法では，計測値を一定のサンプルで切り出し，平均値と平均偏差からルールベースの手法で 4 種の行動を認識している．Bao らの手法では，計測値を同様に一定のサンプルで切り出した後に，FFT を用いて周波数に依存する特徴量を抽出し，機械学習によって 20 種の行動を認識している．井上 [4] の解説記事でも述べられているように，計測したセンサ波形から前処理を行い一定のサンプルを切り出して，特徴量を抽出し機械学習によって行動を推定するという手法は，行動認識において広く用いられている．2010 年代後半ごろから，行動

認識分野においても深層学習が用いられることが増えてきている [5]，従来は人間の経験に基づいて，生データから特徴量を設計していたが，与えられたデータセットに基づいて特徴表現自体を学習する表現学習が用いられる点が，深層学習の特色の一つである．

行動認識モデルの推定精度向上に向けて，様々な特徴抽出方法やモデル構造が提案されているが，シンプルな手法ながらも精度向上に大きく寄与する手法としてアンサンブル学習がある．アンサンブル学習は異なる複数のモデルを併用する機械学習手法の総称である．一つのデータセットから複数のブートストラップをサンプリングして複数のモデルを訓練し，各モデルの出力で多数決する Bagging[6] や，苦手なデータに対する予測を補足するように新たなモデルを訓練し多段に重ねる Boosting[7]，複数のモデルの予測値を入力として新たなモデルを訓練する Stacking[8] が有名である．深層学習モデルをアンサンブルする手法 [9] や行動認識において従来の機械学習手法と深層学習の結果をアンサンブルする手法 [10] なども提案されている．

アンサンブル学習は強力な手法ながらも，訓練時に複数のモデルを個別に訓練し，推論時に各モデルを結合して動作させる必要があるため，単一モデルを用いる場合よりも実装が煩雑になるという課題がある．そこで本研究では，行動認識の推定精度向上に向けて，訓練及び推論時の煩雑さを改善するシンプルな深層学習アンサンブル手法の開発を行う．ベンチマークデータセット HASC[11] を用いて，深層学習のアンサンブル手法を解析することを通じて，単一モデルを End-to-End で訓練するだけでアンサンブルモデルと同等の推定精度を実現する手法の実現可能性を考察する．アンサンブルモデルの持つ精度向上のメリットを維持しつつ，訓練と推論の煩雑さを軽減した学習手法を実現することを本研究の目的とする．

2. 関連研究

本章では，行動認識研究領域におけるアンサンブル学習の適用事例と，深層学習研究におけるアンサンブル手法の解析に関して調査し，本研究の立ち位置を明確にする．

2.1 行動認識におけるアンサンブル学習

行動認識におけるアンサンブル学習の適用事例として，人間が設計した特徴量を用いてアンサンブルな分類器を使用する研究が多数ある．Subasi ら [12] は行動認識において Adaboost Classifier と Random Forest を併用することが優れていると述べている．Irvine ら [13] は Neural Network (NN) をアンサンブルしてスマートハウス内の行動認識を行う手法を提案している．Nurul ら [14] は，行動認識の問題設定を活かして被験者の体型情報から類似度ごとにサブセットに分割し，推定対象者と類似したサブセットを用いて訓練したアンサンブル学習モデルで行動を予測する手法を提案している．Xu ら [15] は信号処理に基づく特徴量と FFT に基づく周波数特徴量を用いて，Cascade Ensemble Learning (CELearning) により行動認識を行う手法を提案している．CELearning は，NN を用いずに深層な表現学習を目指した Deep forest[16] 内で Cascade Forest として提案されている手法である．CELearning では複数のモデル集合を一つの層とみなし，複数モデルの出力を結合して次の層の入力とする．このとき，もとの入力も結合する点が特徴的である．入力を結合する点は ResNet[17] の Skip connection に類似しているが，Skip connection は和を取るのに対して，CELearning は特徴量方向に結合している．

深層学習モデルをアンサンブルする手法も提案されている．Zhu ら [10] は人間が設計した特徴量を用いた分類器を複数構築した上で，Convolutional Neural Network (CNN) を用いた生データからの表現学習を併用し，最終的な出力を Weighted Voting により決定する手法を提案している．Semwal ら [18] は CNN に Long short-term memory (LSTM) を接続したモデルを 4 つ並列で用いて得られた出力を結合し，更に全結合層を通して最終的な出力を求める手法を提案している．深層学習を用いない手法では複数モデルの出力から多数決や重み付き和，もしくはルールベースの手法により最終的な出力を決定する手法が多かった．深層学習を用いる場合，Semwal らのように獲得した特徴表現を結合し新たな全結合層に接続する手法が採用できる．

2.2 深層学習アンサンブル手法の解析

アンサンブル学習はモデルの多様性により推定精度向上に寄与することが知られているが，一方で，未解明な点も多く残っている．Wasay らの研究 [19] では，画像認識における CNN を対象に，比較的少ないパラメータを持つ複数モデルのアンサンブルと，同等のパラメータ数を持つ単一モデルではどちらが優れているのかを実験的に解析している．複数の画像認識ベンチマークで比較した結果，アンサンブル学習が有効に働くにはある程度のパラメータ数を有する必要があることや，データセットが複雑であるほどアンサンブル学習が効果的であることなどを明らかにした．

Allen-Zhu ら [20] の研究では，深層学習におけるアンサンブル学習と知識蒸留 [21] がテストデータに対する推定精度を向上させることは Mystery である[*1]として考察を行っている．CIFAR-100 を用いた実験では個別の 10 モデルのテスト精度は 81.51%だが，各モデルの出力を重み付け平均するアンサンブルによりテスト精度が 84.69%に向上する．一方，各モデルの和を最適化して全てのモデルをまとめて訓練する手法ではこのような精度向上が起こらずテス

[*1] Microsoft Research Blog: Three mysteries in deep learning: Ensemble, knowledge distillation, and self-distillation https://www.microsoft.com/en-us/research/blog/three-mysteries-in-deep-learning-ensemble-knowledge-distillation-and-self-distillation/

Vol.2021-MBL-100 No.25
Vol.2021-UBI-71 No.25
Vol.2021-CDS-32 No.25
Vol.2021-ASD-21 No.25
2021/9/3

ト精度が 81.83%となる．このように，深層学習アンサンブルではモデルを集約して訓練してもアンサンブルの恩恵は得られず（すなわち，モデルの大規模化による影響ではなく），個別に学習することが多様性の獲得に寄与していることを示している．また，大雑把に捉えれば自己蒸留（Self-Distillation）はアンサンブル学習と知識蒸留を組み合わせたものであると述べ，単一モデルにおいてアンサンブルと同等の精度向上を目指す仕組みの実現可能性を示している．

2.3 本研究の立ち位置

以上を踏まえ，本研究ではアンサンブル学習の持つメリットを維持しつつ，アンサンブル学習の訓練と推論の煩雑さを軽減した学習手法の実現を目指す．特に，行動認識という画像認識とは異なるドメイン設定において，深層学習アンサンブル手法を改めて解析し，煩雑さ軽減に向けた新たな知見を明らかにした点が本研究の貢献である．

3. 行動認識における深層学習アンサンブル手法の解析

本章では，行動認識における深層学習アンサンブル手法を分解して考察し，どの要素が予測精度の向上に特に寄与するのかを明らかにする．アンサンブル学習の精度向上のためには，(1) 複数モデルをどのようにしてアンサンブルするのか，(2) 各モデルにどのようにして多様性をもたせるのかの 2 点に関して議論が必要となる．本章では (1) のアンサンブル手法に関して議論を行い，次章で (2) の多様性のもたせ方を含めた議論を行う．

3.1 実験設定

議論を行うに当たって，アンサンブル手法以外の要素を統一するため，本実験では先行研究 [22] において有効性が示されている VGG 構造を持つ同一の CNN モデル（**図 1**）を 5 つアンサンブルする．各モデルに多様性を獲得させるため，データセットを人ごとに分割し組み合わせを変えて複数のサブセットを構築し，モデルごとに訓練データを変えるようにした．なお，実験の間でサブセットは固定とする．

3.1.1 データセット

本研究では HASC データセット [11] を用いてスマートフォンの加速度センサデータから行動認識を行う．HASC は基本行動 6 種類（停止，歩行，走行，スキップ，階段上り，階段下り）のラベルがついた加速度，ジャイロ等のセンサデータが提供されている．2011 から 2013 年までの BasicActivity よりサンプリング周波数が 100Hz のデータを抽出し，加速度センサデータのみを用いた．前処理として，各計測ファイルから前後 5 秒を除去し，sliding-window 方式で 256 サンプルの window を stride=256 で抽出した．

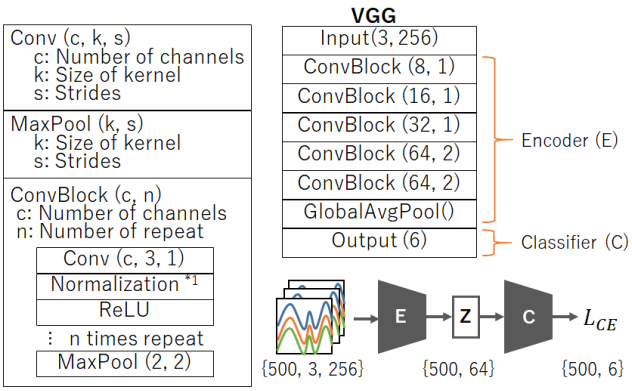

図 1 実験で使用した VGG ベースのモデル構造．Normalization 層は実験によって変更があるため本文で説明する．

Fig. 1 Model architecture based on VGG that we used in experiments. Because the Normalization layer is changed in each experiment, we describe in the main text.

前処理後のデータから，ランダムに抽出した 10 名分のデータを訓練用データセット（D_{train}），別の 10 名分のデータを検証用データセット（D_{val}），更に別の 50 名分のデータをテスト用データセット（D_{test}）として分割した．また，アンサンブル用に今回は被験者単位で D_{train} を 5 サブセットに分割した（$S_1, S_2, ...S_5$）．すなわち，今回はアンサンブル数 $n = 5$ とした．各サブセットには交差検証のように 8 名分のデータが含まれるように設定し，実験内で被験者は固定とした．言い換えると，本研究では Bagging に近い形でアンサンブル学習を行う．

3.1.2 モデル構造

実験で用いる VGG モデル（図 1）は 8 層のシンプルな VGG 構造を採用した．アンサンブル学習による影響を調査するため，層数は 8 層に減少させ，かつ，特徴抽出器 E で獲得した特徴マップを Global Average Pooling によりチャネルごとに平均し，直接出力層につなぐ形としている．バッチサイズは 500 としている．したがって，入力は {500, 3, 256}，特徴マップ z は {500, 64}，出力は {500, 6} となる．図 1 中の Normalization は本章の実験においては一般的な Batch Normalization を採用している．その他，予備実験においてチューニングした結果，VGG の第一層のフィルタ数は 8，最適化手法は Adam を学習率 0.001 で 100epochs 訓練することとした．

3.2 実験結果

単一の深層学習モデル及びシンプルな深層学習アンサンブルモデルの挙動を解析することで，アンサンブル学習のどの要素が精度向上に寄与するのかを考察する．実験で使用したモデルは**図 2** に示すように単一の深層学習モデルを D_{train} 全体，もしくは各サブセット（$S_1, S_2, ...S_n$）を用いて訓練する（それぞれ，Single10，Single8 と呼ぶ）．これらはベースライン兼，以降のアンサンブルモデルのための事

Vol.2021-MBL-100 No.25
Vol.2021-UBI-71 No.25
Vol.2021-CDS-32 No.25
Vol.2021-ASD-21 No.25
2021/9/3

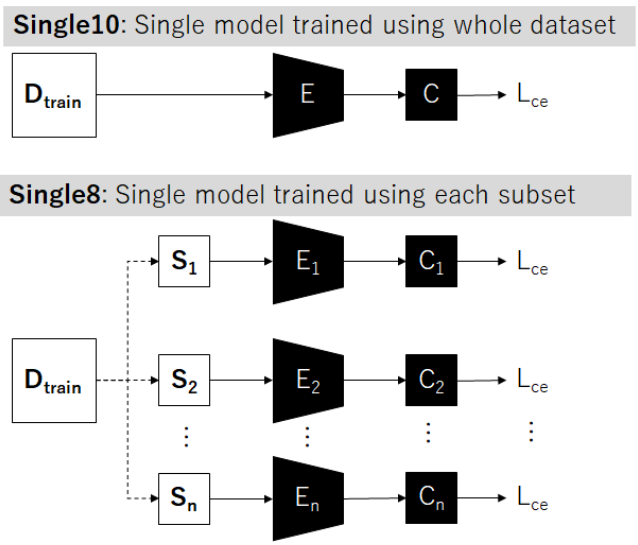

図 2 アンサンブル前の個別のベースモデルの訓練概要．Single10 は D_{train} 全体を用いて単一モデルを訓練し，Single8 は $S_1, S_2, ...S_n$ を用いて各単一モデルを訓練する．

Fig. 2 Training outline of each model before ensembling. Single10 is trained using whole of D_{train}, and each Single8 model is trained using each subset $(S_1, S_2, ...S_n)$.

前訓練モデルである．Single8 をベースラインとして使用する際には 5 モデルのうち最高精度のものを採用する．この事前訓練済みモデルを使用して，**図 3** に示す 3 種類のアンサンブルモデルを準備する．Vote は各サブセットで個別に訓練したモデルの出力から，Bagging のように多数決で最終出力を決定する手法である．E-Ens は各サブセットで訓練した Encoder の出力を結合し，新たな分類器 C に接続する手法である．ここで，新しい分類器 C は D_{train} 全体を用いて訓練されるため，エンコーダのみがアンサンブル性を獲得するモデルとなる．C-Ens は Single10 の Encoder を用いて，新たな分類器 $(C_1, C_2, ...C_n)$ に分岐接続する手法である．ここで，新しい分類器は各サブセットで追加訓練されるため，分類器のみがアンサンブル性を獲得するモデルとなる．

表 1 に乱数シードを変えて 50 試行した検証結果を示す．個別モデルの精度は訓練データの総数や Dropout 層の有無によらず検証精度 81.5%，テスト精度 84.5%程度に収束する結果となった．検証精度よりもテスト精度のほうが高いのは選択された被験者の偏りによるものである．検証データは 10 人，テストデータは 50 人で構成されていることから，検証データのほうが推定が難しい被験者の影響を顕著に受けるためである．

続いて各アンサンブル手法の精度を見ると，概ね Vote≥E-Ens>C-Ens≃Single10 となっていることがわかる．この結果から，行動認識においてもアンサンブル学習により推定精度向上が確認された．また，アンサンブル学習により多様性を獲得するのは Classifier ではなく Encoder 側の方が

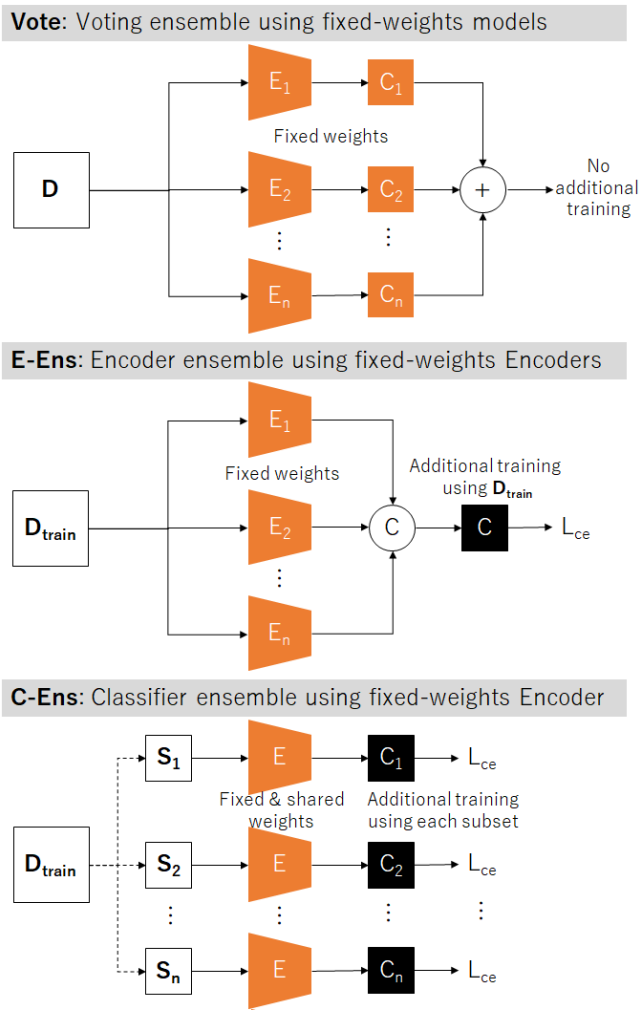

図 3 訓練済みモデルを用いたアンサンブルモデル．Vote はサブセットで訓練済みのモデルの出力を加算することで最終的な出力を決定する．E-Ens はサブセットで訓練済みの各 Encoder の出力を結合し新たな Classifier を D_{train} 全体で訓練する．C-Ens は D_{train} で訓練済みの Encoder を新たな Classifier に分岐させ各サブセットで訓練する．

Fig. 3 Ensemble models using pre-trained models. Vote determines final output by adding each model output pre-trained using subsets. In E-Ens, a new attached classifier C with fixed-weights encoders is additionally trained using D_{train}. In C-Ens, new attached classifiers with a fixed-weights encoder is additionally trained using each subset.

重要であることが示唆された．

4. 深層学習アンサンブル手法の単一モデル化に関する検討

ここまでの実験では，深層学習アンサンブルが行動認識においても有効に働くこと，また，Backbone となる Encoder に多様性をもたせることの重要性を明らかにした．これを踏まえ本章では，深層学習アンサンブルの訓練及び推論時の煩雑さを解消する手法について検討を行う．

Vol.2021-MBL-100 No.25
Vol.2021-UBI-71 No.25
Vol.2021-CDS-32 No.25
Vol.2021-ASD-21 No.25
2021/9/3

表 1 乱数シードを変更し基本モデルで 50 試行した検証結果 [%]
Table 1 Verification results of 50 trials with the basic model by changing the random number seed [%].

	Val acc	Val std	Test acc	Test std
Single10	81.5	1.3	84.7	1.0
Single8 (best)	81.4	1.4	84.8	0.7
Vote	**83.3**	0.9	**86.1**	0.6
E-Ens	82.9	**0.6**	85.9	**0.3**
C-Ens	81.1	1.4	84.8	0.6

4.1 深層学習アンサンブル手法の煩雑さ

はじめに，深層学習アンサンブル手法の煩雑さとは何かを定義する．図 2，図 3 の Single10 と Vote を比較すると，訓練時と推論時でそれぞれ異なる煩雑さがあることがわかる．訓練時に着目すると，Vote は Single10 に加えて，(1) 複数のサブセットを用意し，(2)n 回の順伝播で n 個の出力を算出し，(3) 各サブセットで計算した Loss を用いて n 回の逆伝播によりモデルを更新すると言った流れになる．実装を考慮すると，Loss が複数あることから各パスごとに一つのモデル（E+C）を準備する必要がある．推論時に着目すると，(1) 入力を複数のパスに分岐させ，最終結果を加算して出力する必要がある．推論時にはモデルを集約して一本化することは可能であるが，複雑に分岐するモデルは転移学習に向いていないため，同一のパラメータ数であればシングルパスのモデルにできる方が管理の都合上望ましい．すなわち深層学習アンサンブル手法の煩雑さは，複数回モデルを訓練しなければならないという訓練上の煩雑さと，複数モデルを管理しなければならないという運用上の煩雑さが内包されている．

訓練時の煩雑さを解決するためには，入出力をそれぞれ 1 つに集約したモデルが実装できることで解決できる．すなわち，入力は D_{train} で統一され，(1) モデル内部で D_{train} を動的にサブセットに分割する．(2) 出力は複数個別に実施せず加算等でひとまとめにして出力する．これにより，単一のデータセットを入力し，単一の Loss でモデル全体を訓練できるようになり，Single10 等の単一モデルと同等の訓練手順が実施できる．

推論時の煩雑さを解決するためには，得られた訓練済みモデルを 1 つのシングルパスのモデルに変換できることが望ましい．Allen-Zhu ら [20] の研究では，これを知識蒸留により実現する手法が検証されているが，知識蒸留を行う場合，別途 Student モデルを準備し，Teacher-Student モデル間の出力が同一になるように別途訓練を行う必要がある．すなわち推論時の煩雑さを改善するために訓練が煩雑になってしまう．したがって，訓練で得られた複数モデルの重みをうまく流用することでこの解決を図ることが望ましい．

4.2 入出力の集約とアンサンブルモデル

訓練時の煩雑さ解決に向け，モデルの入出力をそれぞれ 1 つに集約することを考える．ここで，各個別モデルに対してサブセットを用いずに，全てのモデルを D_{train} で訓練するという制約を課すことによって，図 3 の Vote や E-Ens のような単一入力，単一出力のモデルを構成することが可能となる（ただし，各 E，C は事前訓練しないものとする）．これをそれぞれ Vote(scratch)，E-Ens(scratch) と呼ぶことにする．

入力された D_{train} をモデル内で動的にサブセット化する仕組みとして，Instance Masking Layer (IML) を新たに提案する．IML は**図 4** に示すように，通常の入力（D_{train}）に加えて，Mask を入力する必要がある．Mask はバッチサイズ × アンサンブル数 n の書式で 0, 1 の値が格納されているものとする．IML はこの 2 入力を受け取り，slice または zero mask によってサブセットを生成する．slice は Mask が 1 となるインスタンスのみを残してスライシングする手法であり，zero mask は 0 となるインスタンスの値をすべて 0 とする手法である．slice はサブセットを完全に再現できる利点がある一方で，各サブセットのバッチサイズが異なってしまうという欠点がある．すなわち，各モデルの最終的な出力に対して加算や結合によって Loss を一つに集約することができない．一方で，zero mask はサブセットを近似的に再現するが，バッチサイズを変えないため最終的な出力に対して加算や結合によって Loss を一つに集約できる．

IML を用いて入力に多様性を持たせつつ，単一の入出力でモデルを訓練するモデルとして，IML-Vote と IML-Ens を実装する．IML-Vote は図 3 の Vote に対して先頭に IML(zero mask) を挿入する手法である．IML(zero mask) はサブセット間でバッチサイズを変更しないため，最終的な出力を加算によりマージできる．IML-Ens は図 2 の Single8 の先頭に IML(slice) を挿入する手法である．IML(slice) はスライスによりバッチサイズを変更するため，IML が通常のアンサンブル学習と同等の学習を実現できているのかを確認するために実装する．なお，Loss は各出力に対して計算するが，最終的にすべての Loss を加算してモデル全体を訓練するため Optimizer は 1 つで良い．

また，zero mask を行うことと，将来的に単一モデルにマージすることを想定し，図 1 のベースモデルの以下の点を変更して以降の実験を行った．

- Normalization を Layer Normalization (LN) に変更
- LN で学習可能パラメータを不使用に変更
- 全 Convolution, Linear 層でバイアス項を不使用に変更

この変更によって，zero mask されたインスタンスはモデルを経由しても出力が全て 0 になる．なお，Normalization 層の変更に伴い学習の収束に時間を要する挙動が確認されたため，訓練エポック数を 300 に増やした．

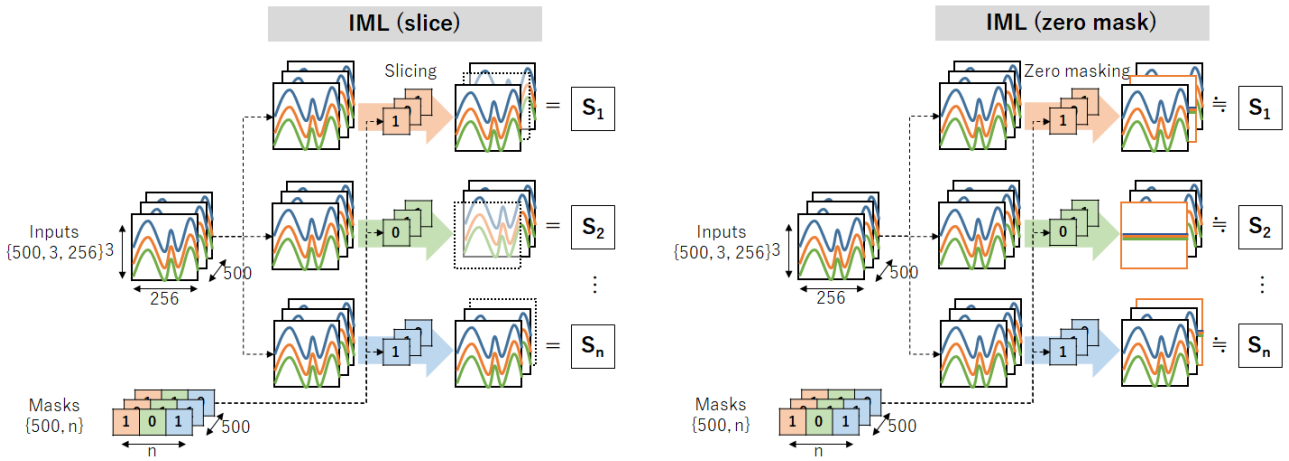

図 4 Instance Masking Layer (IML) の処理概要図
Fig. 4 Process outline of Instance Masking Layer (IML).

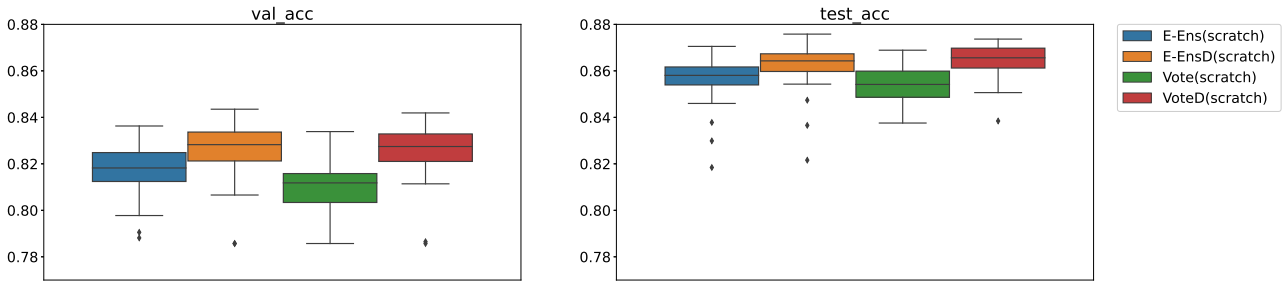

図 5 単一入出力アンサンブル深層学習モデルの事前検証結果（50 試行） [%]
Fig. 5 Verification results for single input and single output ensemble models (50 trials) [%].

4.3 単一入出力モデルの事前検証

単一入出力としたアンサンブル深層学習手法 Vote(scratch)，E-Ens(scratch) の事前検証の結果，アンサンブルしない場合と比較して精度向上が僅かであった．原因を考察した結果，アンサンブル深層学習を実装する際の実装上の注意点が明らかとなったため説明する．

3 章の実験における図 3 の Vote は，追加訓練を必要としないためシンプルに各出力を加算するだけで良かった．一方で，Vote(scratch) や IML-Vote のようにスクラッチから学習する際には，各出力に $\frac{1}{n}$ を乗算する必要がある．なぜならば，通常の訓練は出力に対して式 (1) に示す Softmax 関数で正規化された上で，カテゴリカルクロスエントロピー誤差を算出し，モデルの最適化を行う．

$$Softmax(y_i, \boldsymbol{y}) = \frac{e^{\frac{y_i}{T}}}{\sum_{y \in \boldsymbol{y}} e^{\frac{y}{T}}} \tag{1}$$

ここで式 (1) の T は温度パラメータであり，$T = 1$ で一般的な Softmax 関数として，$T < 1$ で高い値を強調するように，$T > 1$ で低い値を強調するように働く．アンサンブルモデルでは出力を n 個加算して出力するため，各出力が n 倍された状態と近似できる．各 y が n 倍された状態は，すなわち温度パラメータ $T = 1/n$ と同義であり，高い出力値を強調するようにした Softmax 関数となる．実験的にこれは $T = 1$ のものよりも推定精度を低下させる傾向を確認したことから，本研究では $T = n$ として，すなわち，出力に $\frac{1}{n}$ を乗ずることとした．なお IML-Vote の際には各モデルの出力を加算した際に zero mask によって n が変動するため，mask から n を算出する必要がある．

E-Ens(scratch) のように各 Backbone の特徴マップを直接結合し，単一の Classifier を経て Loss を計算するようなケースでは上記現象は発生しないと思われた．しかしながら，事前検証ではアンサンブルによる精度向上が僅かであった結果を受け，温度パラメータとは別に結合された特徴マップ自体に $\frac{1}{n}$ を乗じたモデルの検証を行った．

検証結果を**図 5** に示す．ここで，温度パラメータや特徴マップへのアンサンブル数 n に関する対応を行わなかったモデルをそれぞれ Vote(scratch)，E-Ens(scratch) と呼び，対応を行ったモデルを Divided Model としてそれぞれ VoteD(scratch)，E-EnsD(scratch) と呼ぶ．図より対応を行わなかったモデルはアンサンブル学習しなかった場合の検証精度（81.5 ± 1.3%）とほぼ変わらなかったが，n の対応を行うことで 1%程度の精度向上を果たした（82.6 ± 1.2%）．以上を踏まえ，アンサンブル深層学習を行う際には各モデルをマージする前に $\frac{1}{n}$ を乗ずることが重要であることが明らかとなった．また，単一入出力モデルにお

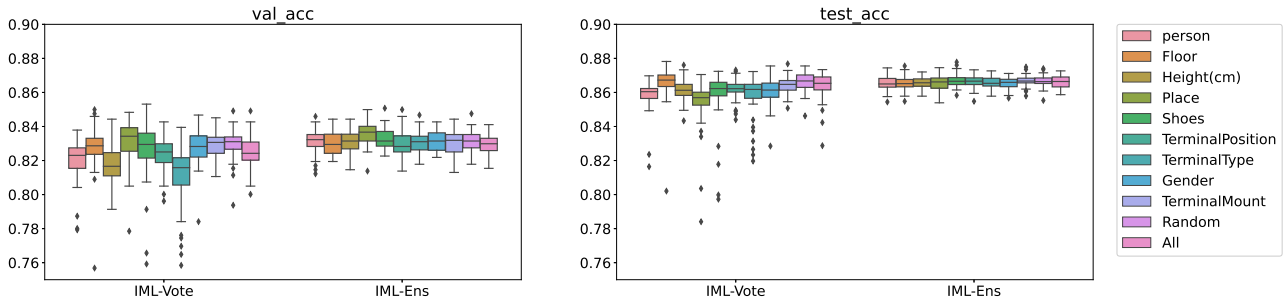

図 6 Mask 生成手法別の検証結果（50 試行） [%]
Fig. 6 Verification results for each mask generation method (50 trials) [%].

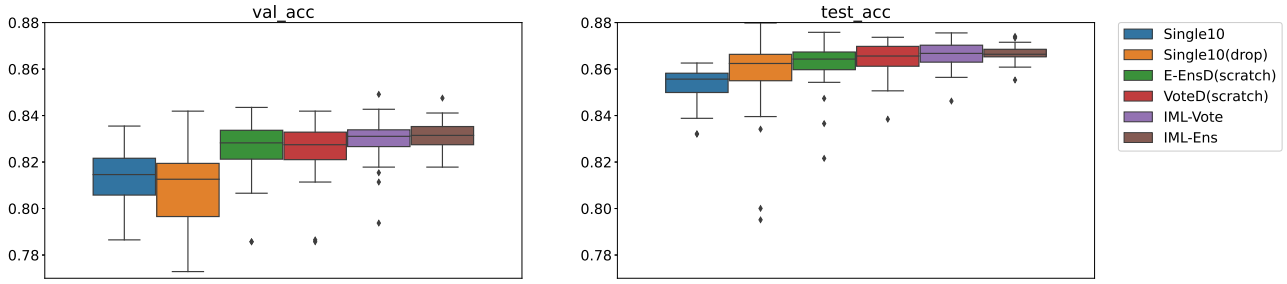

図 7 各アンサンブルモデルの検証結果（50 試行） [%]
Fig. 7 Verification results for each ensemble model (50 trials) [%].

いては Classifier を 1 つに集約してもしなくても同等の精度に収束することも明らかとなった．興味深い結果として，温度パラメータで説明のつく Vote(scratch) のみでなく，E-Ens(scratch) に関しても同様の改善が得られた点がある．この原因については追って調査することとする．

4.4 IML モデルの事前検証

IML を用いたアンサンブルモデルの評価を行うに当たり，Mask の作成方法（すなわち，3.1.1 節で述べたサブセットの分割方法）に関する事前検証を行った．乱数シードを変えて 50 試行を行った結果を**図 6** に示す．各 Mask は HASC に含まれていたメタデータを参照し，交差検証を行うように適宜設定したものである．例えばメタデータの Shoes は行動時の靴ラベルであり，このラベルを基準に交差検証するように Mask を設定した．Height のような量的変数は一定の間隔で区切って質的変数に変換している．Random は全インスタンスからランダムに交差検証するように Mask を設定したものであり，All は Mask を全て 1 としたものである．すなわち，All の場合 IML-Vote は Vote(scratch) と同義であり，IML-Ens は図 3 の Vote で行った各モデルを事前訓練した上でアンサンブルする手法と同義である．図 6 をみると，Mask の作成方法によって多少精度が上下するが基本的には誤差の範囲内の変動であると思われる．ただし，全体を通して All よりも Random 等で Mask を作成した方がやや精度が向上するようにも見える．

4.5 実験結果

事前検証の結果を踏まえ，IML の Mask を Random とした結果と，Mask を使用しない単一入出力モデルとの比較を行った結果を**図 7** に示す．図より興味深い点がいくつか確認できる．1 点目は Instance Mask による精度向上が極僅かであるという点である．IML を挿入した 2 手法は Mask によりサブセットに擬似的な多様性をもたせるアンサンブル学習手法であるが，図 6 の結果を踏まえてもサブセットを変えるメリットは小さい可能性が示唆された．2 点目は Instance Mask を用いた手法がわずかに試行間のばらつきを抑えるという点である．**表 2** に各モデルの平均精度と標準偏差を示す．IML-Ens は多出力によって一般的なアンサンブル深層学習手法と変わらないことから，アンサンブルによるばらつきの抑制効果があったと考えられる．一方，VoteD(scratch) と IML-Vote の違いは Instance Mask により各 Backbone に多様性を持たせたかという点である．これが僅かに精度向上とばらつきの抑制に貢献したと考えられる．3 点目は表 1 の Vote と表 2 の各結果の差が僅かである点である．まず，Vote と IML-Ens は実質的に同義の手法であり精度差もほぼない．興味深い点は Vote と VoteD(scratch) の精度差が検証データで+0.7%，テストデータで-0.4%である．すなわち，Allen-Zhu ら [20] の研究で述べられていた「各モデルの和を最適化して全てのモデルをまとめて訓練する手法ではこのような精度向上が発生しない」という説明に反する結果となった．この点については，行動認識分野独自の結果なのかを今後深く検証する必要がある．

表 2 各アンサンブルモデルの検証結果（50 試行）[%]
Table 2 Verification results for each ensemble model (50 trials) [%].

	Val acc	Val std	Test acc	Test std
Single10	81.4	1.1	85.3	0.7
Single10(drop)	80.7	2.1	85.8	1.6
E-EnsD(scratch)	82.6	1.2	86.3	0.9
VoteD(scratch)	82.6	1.1	86.5	0.7
IML-Vote	83.0	0.9	86.6	0.5
IML-Ens	**83.1**	**0.6**	**86.7**	**0.3**

なお，zero mask によって学習中にインスタンスを 0 にマスクする手法は，よく知られた正則化手法である Dropout[23] と類似しているが異なる手法である．図 4 に示したように，IML はバッチサイズの方向に対してマスク処理を行うためインスタンスを drop する動作となる．一方，Dropout は図 4 でいう Window（256 samples）や channel（3 ch）方向に対してランダムにマスク処理を行う手法であり，かつ epoch ごとに drop するユニットも異なる．結果として，図 7 に E の直後に Dropout 層を追加したモデルの検証結果も載せているが，シンプルな単一モデルと大差ない結果となっている．

4.6 訓練済みモデルのシングルパス化

推論時の煩雑さを軽減するために，訓練済みモデルをシングルパスの単一モデルにする手法を考える．有名なモデル圧縮手法として，前述の知識蒸留 [21] がある．知識蒸留は大規模な単一モデルまたは，アンサンブルモデル（Teacher）を小規模な単一モデル（Student）に置き換える手法である．通常，モデルの訓練には one-hot ベクトル表現された教師ラベルとカテゴリカルクロスエントロピー誤差を用いた最適化が行われる．知識蒸留では Teacher モデルの出力値を正解ラベルとして Student モデルを訓練することで one-hot ベクトル表現よりもリッチな情報を教師として与える手法である．一方で，知識蒸留では Student モデルを追加訓練する必要があるという手間が残る．

訓練済みモデルの重みを流用してモデルを合成することを考えるとき，図 3 に示した Vote を E-Ens にするように，複数の Classifier を単一の Classifier にすることは容易に実現できる．両者は入力及び出力の書式が同一であるため，Vote の各 C の重みを連結し E-Ens の C にコピーすることで同等の動作をする単一の Classifier となる．一方で Encoder については畳み込み演算を行っていることからモデルの合成には一工夫が必要である．通常の畳み込み演算は，入力のチャネルそれぞれに対してカーネルを定義し，カーネルを畳み込んだ和を次の層の入力とする．したがって，複数の個別に学習された E をマージするときにはシンプルに連結するだけでなく，不要な入力チャネルに対してゼロマスクを行う，もしくは，Group Convolution のようにチャネル方向に制限のある畳み込みそうで置き換える必要がある．

他にも，1x1 Convolution 層を同等の役割を持つ 3x3 Convolution 層に re-parameterization する手法で，VGG の精度改善を行った RepVGG[24] がある．また，異なる形状と役割を持つ 2 つの CNN モデルを，1 つのマルチタスクモデルに変換する手法 [25] も提案されている．

本稿では，Vote の各 C をマージすることで E-Ens のような単一の Classifier を持つアンサンブル構造において同等の挙動を示すことを実験的に確認済みである．一方，畳み込みを含む Encoder をマージする手法に関しては現在未検証であるため，今後の課題としたい．

5. おわりに

本研究では，センサを用いた人間行動認識における深層学習アンサンブル手法の有効性検証と解析を行った．実験の結果，以下のような様々な知見を明らかにした．

- 行動認識においても深層学習アンサンブル手法は有効に働くこと
- 深層学習アンサンブルは Classifier の多様性よりも Encoder の多様性が重要であること
- 入力に多様性を与えなければ，単一入力単一出力のモデル構造を実現できるが，各モデルの特徴マップを結合する場合においても $\frac{1}{n}$ を乗ずることが重要であること
- 入力に多様性を与える場合，Instance Masking Layer により実現ができ，Random な Instance Mask により，同一入力のモデルの精度を上回り，試行間のばらつきを抑制する可能性を示したこと

本研究により，シンプルな精度向上を求めるのであれば，同一モデルに対して同一入力を渡して出力の重み付き和を取るラッパーモデルを用いるだけで良いことが明らかとなった．一方，今後の課題として，Instance Mask 以外の手法で各モデルに多様性をもたせる手法に対する検証実験を行う必要がある．本実験ではメタデータを用いてデータセットから複数の基準でサブセットを生成し比較を行った．アンサンブルモデルに多様性を与えるには，他にも，モデル構造自体を多様にすることや，データ拡張を多様にすること等の手法があるため，これらの検証を行いたい．他にも，単一モデルでは同一パラメータ数であっても何故アンサンブルによる精度向上効果を得られないのかという点について深く考察していく．また，訓練済みモデルを単一のシングルパスモデルに圧縮する手法に関しても実装と評価を行っていく予定である．

謝辞 本研究の一部は，JSPS 科学研究費助成事業若手研究 (19K20420) の助成によるものである．ここに謝意を表す．

参考文献

[1] Rawassizadeh, R., Tomitsch, M., Wac, K. and Tjoa, A. M.: UbiqLog: a generic mobile phone-based life-log framework, *Pers. Ubiquit. Comput.*, Vol. 17, pp. 621––637 (online), DOI: 10.1007/s00779-012-0511-8 (2013).

[2] Aminian, K., Robert, P., Buchser, E., Rutschmann, B., Hayoz, D. and Depairon, M.: Physical activity monitoring based on accelerometry: validation and comparison with video observation, *Med. Biol. Eng. Comput.*, Vol. 37, pp. 304—308 (online), DOI: 10.1007/BF02513304 (1999).

[3] Bao, L. and Intille, S. S.: Activity Recognition from User-Annotated Acceleration Data, *Pervasive Computing. Pervasive 2004. Lecture Notes in Computer Science*, Vol. 3001, pp. 1–17 (online), DOI: 10.1007/978-3-540-24646-6_1 (2004).

[4] 井上創造：ウェアラブルセンサを用いたヒューマンセンシング，知能と情報，Vol. 28, No. 6, pp. 170–186（オンライン），DOI: 10.3156/jsoft.28.6_170 (2016).

[5] Frédéric Li, Kimiaki Shirahama, M. A. N. L. K. and Grzegorzek, M.: Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors, *Sensors*, Vol. 18, No. 679, pp. 1–22 (2018).

[6] Breiman, L.: Bagging Predictors, *Machine Learning*, Vol. 24, pp. 123–140 (online), DOI: 10.1023/A:1018054314350 (1996).

[7] Freund, Y. and Schapire, R. E.: Experiments with a new boosting algorithm, *Proc. of the ICML 1996*, pp. 148—-156 (1996).

[8] Wolpert, D. H.: Stacked generalization, *Neural Networks*, Vol. 5, No. 2, pp. 241–259 (online), DOI: 10.1016/S0893-6080(05)80023-1 (1992).

[9] Chen, Y., Wang, Y., Gu, Y., He, X., Ghamisi, P. and Jia, X.: Deep Learning Ensemble for Hyperspectral Image Classification, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 12, No. 6, pp. 1882–1897 (online), DOI: 10.1109/JSTARS.2019.2915259 (2019).

[10] Zhu, R., Xiao, Z., Li, Y., Yang, M., Tan, Y., Zhou, L., Lin, S. and Wen, H.: Efficient Human Activity Recognition Solving the Confusing Activities Via Deep Ensemble Learning, *IEEE Access*, Vol. 7, pp. 75490–75499 (online), DOI: 10.1109/ACCESS.2019.2922104 (2019).

[11] Kawaguchi, N., Ogawa, N., Iwasaki, Y., Kaji, K., Terada, T., Murao, K., Inoue, S., Kawahara, Y., Sumi, Y. and Nishio, N.: HASC Challenge: Gathering Large Scale Human Activity Corpus for the Real-World Activity Understandings, *In Proc. of the AH 2011* (2011).

[12] Subasi, A., Dammas, D. H., Alghamdi, R. D., Makawi, R. A., Albiety, E. A., Brahimi, T. and Sarirete, A.: Sensor Based Human Activity Recognition Using Adaboost Ensemble Classifier, *Procedia Computer Science*, Vol. 140, pp. 104–111 (online), DOI: 10.1016/j.procs.2018.10.298 (2018).

[13] Irvine, N., Nugent, C., Zhang, S., Wang, H. and NG, W. W. Y.: Neural Network Ensembles for Sensor-Based Human Activity Recognition Within Smart Environments, *Sensors*, Vol. 20, No. 1, pp. 1–26 (2020).

[14] Choudhury, N. A., Moulik, S. and Roy, D. S.: Physique-based Human Activity Recognition using Ensemble Learning and Smartphone Sensors, *IEEE Sensors Journal*, Vol. Early Access, pp. 1–10 (online), DOI: 10.1109/JSEN.2021.3077563 (2021).

[15] Xu, S., Tang, Q., Jin, L. and Pan, Z.: A Cascade Ensemble Learning Model for Human Activity Recognition with Smartphones, *Sensors*, Vol. 19, No. 10, pp. 1–17 (online), DOI: 10.3390/s19102307 (2019).

[16] Zhou, Z.-H. and Feng, J.: Deep Forest: Towards An Alternative to Deep Neural Networks, *Proc. of the IJCAI 2017*, pp. 3553–3559 (online), DOI: 10.24963/ijcai.2017/497 (2017).

[17] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proc. of the CVPR 2016*, pp. 770–778 (online), DOI: 10.1109/CVPR.2016.90 (2016).

[18] Vijay Bhaskar Semwal, A. G. . P. L.: An optimized hybrid deep learning model using ensemble learning approach for human walking activities recognition, *The Journal of Supercomputing*, pp. 1–25 (online), DOI: 10.1007/s11227-021-03768-7 (2021).

[19] Wasay, A. and Idreos, S.: More or Less: When and How to Build Convolutional Neural Network Ensembles, *Proc. of the ICLR 2021* (2021).

[20] Allen-Zhu, Z. and Li, Y.: Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning, *arXiv*, Vol. 2012.09816v2, pp. 1–70 (online), available from ⟨https://arxiv.org/abs/2012.09816v2⟩ (2021).

[21] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *NIPS Deep Learning and Representation Learning Workshop*, (online), available from ⟨http://arxiv.org/abs/1503.02531⟩ (2015).

[22] Hasegawa, T. and Koshino, M.: Representation Learning by Convolutional Neural Network for Smartphone Sensor Based Activity Recognition, *In Proc. of the CIIS 2019* (2019).

[23] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, Vol. 15, No. 56, pp. 1929–1958 (2014).

[24] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G. and Sun, J.: Repvgg: Making vgg-style convnets great again, *Proc. of the CVPR 2021*, pp. 13733–13742 (2021).

[25] Chou, Y.-M., Chan, Y.-M., Lee, J.-H., Chiu, C.-Y. and Chen, C.-S.: Unifying and Merging Well-trained Deep Neural Networks for Inference Stage, pp. 2049 – 2056 (online), available from ⟨https://www.ijcai.org/Proceedings/2018/0283.pdf⟩ (2018).