行動認識におけるSoftmax関数の温度パラメータに関する一考察

長谷川 達人[1,a)]

概要：行動認識モデルは深層学習を用いて実装されることが多くなってきているが，深層学習はモデル構造や最適化手法などハイパーパラメータが膨大であり，適切に使いこなすには熟練の技能が必要となる．本研究では，膨大なハイパーパラメータの中でも未解明な点が多い softmax 関数の温度パラメータ T と特徴マップの次元数 M に焦点を当てる．特に行動認識ではモデルサイズを調整することは少なくなく，T と M の関係の解明は重要である．深層学習モデルを出力の分散の観点から理論的に考察した結果，出力層のパラメータは M の制約を受けて最適化されており，最適な T の設定はこの制約を緩和できる可能性があると考えた．そこで本研究では，T と M の関係を理論的に考察した上で，様々な行動認識データセットやモデル構造において実験的に検証した．実験の結果，$T = 1$ の従来の設定ではモデルの最良のパフォーマンスを発揮しきれていないこと，M の増加に伴い最適な T も増加すること，最適な T においては softmax 関数の入力の分布が安定していることなどを明らかにした．

Study on Temperature Parameter of Softmax Function in Activity Recognition

Tatsuhito Hasegawa[1,a)]

1. はじめに

スマートフォンやウェアラブルデバイスに搭載されたセンサを用いて所持者の行動等を推定する行動認識は，個人のライフログや健康管理 [1]，在宅高齢者の活動認識 [2]，看護行動の自動記録 [3] など，様々な分野に活用されている．行動認識は計測したセンサ値を入力とし，推定対象の行動クラスを出力する問題としてモデル化できる．センサ値から基本統計量等を特徴量として，決定木等の機械学習アルゴリズムによりモデルを訓練する手法が多く採用されている [4], [5], [6]．近年ではセンサ値を直接入力する深層学習手法を採用する例も増えている [7], [8], [9]．深層学習は大規模な訓練データから人間の理解を超える特徴表現を獲得できる可能性から，様々な分野で高い推定精度を実現している．しかしながら，従来の機械学習手法と比較して，深層学習はモデルアーキテクチャや訓練方法等のハイパーパラメータが膨大である．デファクトスタンダードな手法は確立されているものの，複雑なハイパーパラメータは熟練者の経験的に選定されることが多く，原理的に未解明な点も多い．

本研究では，深層学習を用いた行動認識という分類タスクにおける損失関数に焦点を当てる．損失関数とはモデルの予測値と正解ラベルから算出される予測誤差の指標であり，深層学習は損失関数の勾配からネットワークを訓練する．一般的に分類タスクでは式 (1) で示す Cross-Entropy Loss が用いられる．

$$\mathcal{L}(f(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{y}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\boldsymbol{y}_{ij}\log f_j(\boldsymbol{x}_i;\boldsymbol{\theta}). \qquad (1)$$

ここで，$\boldsymbol{\theta}$ は行動認識モデル f のパラメータであり，$f_j(\boldsymbol{x}_i;\boldsymbol{\theta})$ は i 番目の入力 $\boldsymbol{x}_i$ に対する，カテゴリ j の出力を意味する．$\boldsymbol{y}_{ij}$ は i 番目の入力 $\boldsymbol{x}_i$ に対する one-hot 表現された $\{0,1\}$ の正解ラベルである．Cross-Entropy Loss は真の確率分布 $\boldsymbol{y}$ と推定した確率分布を用いるため，$f_j(\boldsymbol{x}_i;\boldsymbol{\theta})$ は確率分布の様式である必要がある．したがって，モデル f は出力直

[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

前の予測値 $\boldsymbol{z}$ に対して，式 (2) に示す softmax 関数を用いて規格化することが一般的である．

$$\mathrm{softmax}(z_i, \boldsymbol{z}) = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_{z_j \in \boldsymbol{z}} \exp\left(\frac{z_j}{T}\right)}. \qquad (2)$$

softmax 関数により，際限ない値を取る予測値 $\boldsymbol{z}$ を，各々 0 から 1 までの値を持ち，合計が 1 となるように規格化することで，モデル f の出力を各クラスの予測確率であるとみなす．ここで，T は温度パラメータと呼ばれる softmax 関数のハイパーパラメータである．$T = 1$ が標準的な softmax 関数であり，分類問題で深層学習モデルを訓練する際には多くの場合 $T = 1$ の softmax 関数が用いられる．

ここで，温度パラメータ T の役割に着目する．**図 1**(a) は，とあるモデルの出力値 z_i であり，標準的な $T = 1$ の softmax 関数に通すと図 1(b) のように合計が 1 になるよう規格化される．T を大きくすると $T = 10$ の例のようにカテゴリ間の差が緩やかになり，小さくすると $T = 0.1$ のように差が極端になるような挙動を示す．温度パラメータの応用事例として，Hinton らの提案した大規模なモデルの知識を小規模なモデルに学習させる知識蒸留（Knowledge Distillation）[10] がある．$T = 1$ の通常の softmax 関数ではなく，T を大きくし緩やかな分布から損失を計算することで蒸留時の効率的な訓練を実現している．他にも，$T = 1$ の softmax 関数は必ずしも予測クラスの出力確信度の確率分布を意味しない点を指摘し，温度パラメータを調整することで確信度を調整する手法 Temperature Scaling [11] も提案されている．このように，温度パラメータの応用事例は存在するものの，一般的な分類問題では $T = 1$ の標準的な softmax 関数が用いられることが大半であり，T の変化に対する学習のダイナミクスは未解明な点が多い．

本研究では，行動認識を対象に，温度パラメータの変化がモデルの訓練に与える影響を調査する．特徴抽出器となる Convolutional Neural Network (CNN) の特徴マップのサイズと温度パラメータの関係に着目し，汎化性能を向上させる最適な温度パラメータを決定する手法を模索する．特に，行動認識分野では特定のデファクトスタンダードなモデル構造が確立されておらず，様々なタスクにおいてモデル構造やモデルサイズをチューニングした上で用いる．したがって，モデルサイズと最適な温度パラメータの関係を明らかにすることにより，今後の行動認識モデル探索時のパラメータ設定の一助となることを目指す．

本研究の貢献は以下の 4 点である．

- 汎化性能向上に最適な温度パラメータ T は，特徴マップの次元数 M の影響を受けることを理論的に示した．
- 行動認識タスクを対象に網羅的な実験を行い，$T = 1$ の従来の訓練環境では必ずしも深層学習モデルの最良のパフォーマンスを発揮できないことを実験により示した．

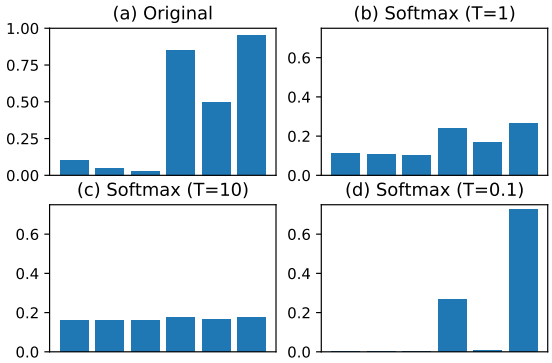

図 1 モデルの出力 z_i を様々な温度をもつ softmax 関数に入力した際の出力例

Fig. 1 Output examples when the model output z_i is inputted into the softmax function having various temperature.

- 対象ドメインやモデル構造によらず，最適な T は M の影響を受けることを大規模な比較実験により示した．
- 訓練済みモデルの，重み w, b や出力 z/T の分布を確認した結果，T を変えることによって w の分散を調整し M の影響を低減できることや，b や z/T の期待値をモニタリングすることで，最適な T を算出できる可能性を示した．

2. 温度パラメータ

2.1 関連研究

対象ドメインは異なるが，温度パラメータに関する議論が一部で行われている．He らは，強化学習において，反復法により最適な温度パラメータを求める手法を提案している [12]．softmax 関数の前後で平均情報量の損失を最小化しつつ，出力確率分布の多様性を最大化するような損失関数を独自に定義し，温度パラメータを最適化している．他にも，画像認識分野ではあるが，Agarwala らは訓練時の逆温度 $\beta = 1/T$ の設定が学習ダイナミクスや汎化性能に与える影響を調査している [13]．最適な β は多くの場合 1 ではなく，$\beta \in [10^{-2}, 10^{1}]$ の範囲でチューニングすることが望ましいとしている．一方，最適な β はモデル構造に依存し探索的に決定されると述べられている．これらの研究のように，温度パラメータはタスクやモデル構造に応じて最適な値が存在することが示唆されているものの，議論はまだ発展途上にある状況である．

2.2 モデル出力と softmax 関数への入力

一例として，3ch の加速度センサデータを入力し，6 種類の行動を識別する行動認識モデルは**図 2** のような構造となる．E は CNN で構成された Encoder であり，図の例では長さ 256，3ch のセンサデータ $\boldsymbol{x} \in \mathbb{R}^{?\times 3\times 256}$ を入力し，M 次元の特徴マップ $\boldsymbol{f} \in \mathbb{R}^{?\times M}$ を出力する．E は

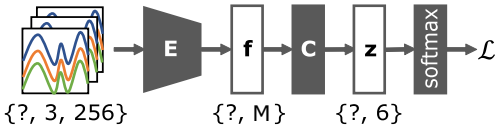

図 2 一般的な行動認識 CNN モデル
Fig. 2 General CNN model of human activity recognition.

VGG[14] や ResNet[15] 等の CNN であり，M はハイパーパラメータである．C は全結合ニューラルネットワークで構成された Classifier であり，図の例では $\boldsymbol{f}$ を 6 クラスの出力 $\boldsymbol{z} \in \mathbb{R}^{?\times 6}$ への写像の役割を果たす．その後，softmax 関数を通して Cross-Entropy Loss が算出される．

ミニバッチ内のとある 1 データを考えると，softmax 関数への入力となる $z_i \in \boldsymbol{z}$ は $f_j \in \boldsymbol{f}$ の全結合として式 (3) で算出される．

$$z_i = \sum_{j=1}^{M} w_{i,j} f_j + b_i. \tag{3}$$

ここで，$w_{i,j}, b_i$ はそれぞれ C の重みとバイアスであり，Xavier[16] や He[17] の手法に基づいて所定の分布に従う乱数で初期化される．E の出力 f_j も $w_{i,j}$ も互いに独立であり，それぞれで同じ確率分布に従うと仮定すると，z_i の期待値と分散は以下のように表現できる．

$$\mathbb{E}[z_i] = M\mathbb{E}[w_i]\mathbb{E}[f] + \mathbb{E}[b]. \tag{4}$$
$$\mathbb{V}[z_i] = M\mathbb{V}[w_i f] + \mathbb{V}[b] \tag{5}$$

ここで，torchvision の EfficientNet 実装[*1]を見ると，畳み込み層は He の初期化を，全結合層は $w_{i,j} \sim U(-1/\sqrt{M}, 1/\sqrt{M})$，$b_i = 0$ で初期化している．このとき，式 (4)，(5) はそれぞれ

$$\mathbb{E}[z_i] = 0. \tag{6}$$
$$\mathbb{V}[z_i] = M(\mathbb{V}[w_i]\mathbb{V}[f] + \mu_{w_i}^2 \mathbb{V}[f] + \mu_f^2 \mathbb{V}[w_i]) + 0 \tag{7}$$
$$= M\{\mathbb{V}[w_i](\mathbb{E}[f^2] - \mu_f^2) + 0 + \mu_f^2 \mathbb{V}[w_i]\} \tag{8}$$
$$= M\mathbb{V}[w_i]\mathbb{E}[f^2] \tag{9}$$
$$= \frac{1}{3}\mathbb{E}[f^2]. \tag{10}$$

となり，M の影響を受けない．

一方，ネットワークの訓練が進むにつれて $w_{i,j}$ の分布は保証されないため，式 (4)，(5) より，$\mathbb{E}[z_i], \mathbb{V}[z_i]$ はいずれも M の一次関数と言える．特徴抽出器 E の末尾で ReLU 等の活性化関数を用いている場合 $\mathbb{E}[f] > 0$ となるため，$\mathbb{E}[w_i] = 0$ を仮定したとしても $\mathbb{V}[z_i]$ は式 (9) のように M の関数となる．すなわち訓練可能なパラメータ w_i は係数 M の制約の上で最適化されている状況であると言える．

[*1] github [efficientnet.py] https://github.com/pytorch/vision/blob/main/torchvision/models/efficientnet.py

2.3 一般的なモデルの特徴マップ

画像認識分野では様々なモデル構造が提案されており，それぞれにおいてデフォルトの特徴マップの次元数が設定されている．torchvision の models[*2]を参照すると，特徴マップの次元数 M は，ResNet50 や Inception-v3 で 2048 次元，DenseNet121 で 1024 次元，MNASNet や EfficientNet で 1280 次元と様々である．

我々の先行研究 [18] において画像認識モデルを行動認識モデルに変換して網羅的に検証は行っているものの，行動認識分野ではデファクトスタンダードなモデル構造はまだ確立されていない．特に画像に対してセンサデータは比較的低次元の入力を扱うことから，モデルのサイズは縮小するように調整されることが経験的に分かっている．

以上のように，画像認識モデル間でも M は一定ではなく，更に行動認識では M を含めたモデルサイズを調整して用いる．上述の仮説である，最適な温度パラメータ T は M の影響を受けるとすると，T と M の関係を明らかにできれば，様々な M に頑健な softmax 関数を実現できる可能性がある．

3. 特徴マップのサイズと温度パラメータ

3.1 実験設定

本研究では，行動認識を対象に M と T の関係を実験により考察する．以降の実験では様々な条件下における結果を考察するが，デフォルトの設定として，データセットは HASC[19] を，モデル構造は VGG[14] を採用する．

HASC[19] は，スマートフォンなどのデバイスを用いて日常行動 6 種類（停止，歩行，走行，スキップ，階段上り，下り）を計測したデータセットである．本実験では，HASC の BasicActivity から 2013 年までのコーパスより，100Hz で計測された加速度センサデータを用いた．人に対する汎化性能を評価するため，被験者を基準にデータセットを分割し，訓練用に 10 名，テスト用に 50 名をランダムに選択した．計測データは前後 2 秒をトリミングし，window size=256，stride=256 のスライディングウインドウ方式でデータを整形した．データ拡張として，チャネルのシャッフルと軸の反転をランダムに行っている．VGG[14] は画像認識コンペ ILSVRC の 2014 年準優勝モデルであり，畳み込み層の連結で構成可能なシンプルなモデル構造である．出力層付近の T の影響を考察するために，シンプルな構造を持つ VGG を採用した．なお，標準的な VGG は 16 層であるが，図 2 のように出力部を GAP と 1 層の全結合層に置き換えた上で，8 層の畳み込みと 1 層の全結合を持つ浅層な VGG 構造を採用している．以降用いるモデルはすべて，E を He の初期化 [17] で，C を $U(-1/\sqrt{M}, 1/\sqrt{M})$ で初期化することとする．畳み込みではバイアス項は用い

[*2] github [torchvision.models] https://github.com/pytorch/vision/tree/main/torchvision/models

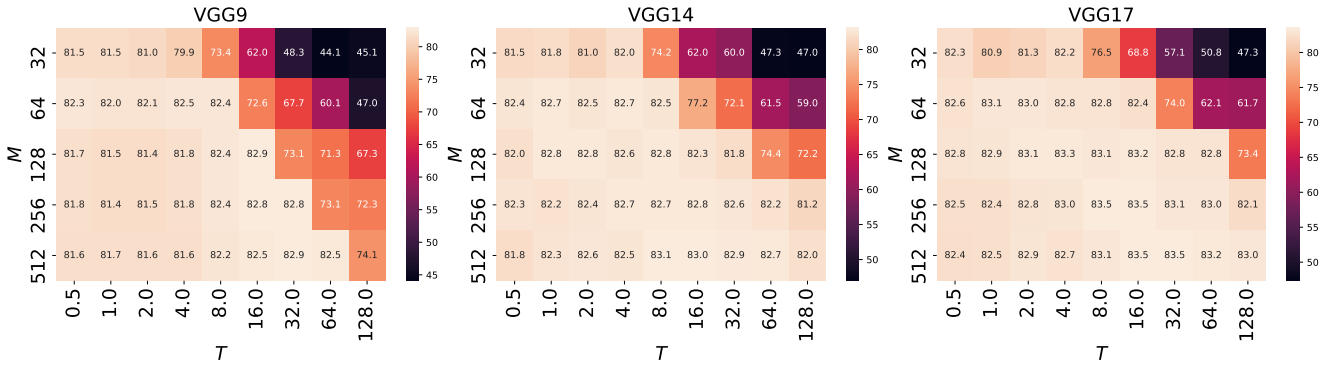

図 3　HASC with VGG における T と M ごとの推定精度（15 試行中央値）
Fig. 3　Accuracies in HASC with VGG for each T and M (median values of 15 trials).

ず，Normalization は Batch Normalization を訓練可能パラメータなしで採用した．

今回，M を変化させる方法として，モデル全体のフィルタ数を増減させる手法を採用した．例えば，標準的な VGG であれば 5 Blocks のフィルタ数がそれぞれ $[64, 128, 256, 512, 512]$ となっている．これをいずれの Block も $1/n$ することで，全体のフィルタ数の変化率は変えずに，M の大きさを制御することとした．

その他，訓練の手続きは以下の通りで固定している．

- 最適化手法は Adam とする．
- 学習率は 0.001 とする．
- エポック数は 500 とする．
- バッチサイズは 1000 とする．

ただし，計算機資源の都合上，複数の端末で分散的に実験を行っていることから必ずしも同一端末上で実施した結果ではない．場合によっては Multi GPU 環境で実験を行っていることもある．しかし，同一の比較実験内に限定すると同一端末内で動作しているものとする．なお，評価指標は乱数シードを変えて複数回試行した際の Accuracy の中央値とする．深層学習モデルは初期値の乱数によっては時折収束がうまくいかない場合が起こりうるため，平均値よりも異常値の影響に強い中央値で考察することとした．

3.2　推定精度への影響

3.2.1　HASC with VGG

VGG と HASC データセットを用いた際の，T と M の推定精度を**図 3** に示す．VGG9，14，17 はそれぞれ層数を示している（C を変更しているため原著論文から 2 層減っている）．考察を以下に述べる．

- 従来の実験手法である $T = 1$ を見ると，層が深くなるほど推定精度が高くなる傾向はあるが，M=64 程度で頭打ちな傾向がある．すなわち，$T = 1$ では M の大きいモデルのパフォーマンスを発揮しきれていない．
- 最良の精度を見ると，M が大きくなるほど，層が深くなるほど推定精度が高くなる傾向がある．
- 各 M で最良精度の T^* に着目すると，$T^* = aM + b$ の関係がありそうであるが，一概に $\{a, b\}$ を決定することは難しそうである．モデルの層数が増加するに伴って，係数 $\{a\}$ が大きくなっているようにも見える．
- 推定精度は $T <= T^*$ にかけて緩やかに上昇し，$T > T^*$ で急激に減少する傾向がある．

3.2.2　モデル構造やデータセットに対する頑健性

本研究の仮説の頑健性を確認するべく，前節の結果が，VGG 構造と HASC データセットを用いた環境に限定的な現象ではないことを実験により調査する．使用した行動認識のベンチマークデータセットは，UniMiB SHAR（UniMiB）[20]，WISDM[4]，UCI-HAR（UCI5）[21] である．それぞれ，被験者を基準にデータセットを分割し，UniMiB で $[16, 9]$ 名，WISDM で $[25, 6]$ 名，UCI5 で $[5, 9]$ 名を訓練用，テスト用に用いた．UCI のみ全データを利用せず，訓練データが 5 名と少ない場合を再現させている．モデルは代表的な構造として ResNet[15]，PyramidNet[22]，SE-Net VGG ver（SEVGG）[23] を採用した．試行回数が膨大（6 パターン $\times 5$ 種の $M \times 9$ 種の $T \times$ 各 10 試行のため総計 2700 試行）であるため，モデル構造は比較的浅層かつ軽量なものを採用した．

図 4 上段は VGG9 でモデル構造を統一し，様々なデータセットで比較している．図 4 下段は HASC でデータセットを統一し，様々なモデル構造で比較している．いずれも全体の傾向はデータセットやモデル構造によらず前章の結果と同様であることがうかがえる．一方で，層数が増えたときと同様に，係数 $\{a, b\}$ もデータセットやモデル構造に応じて変化しているようすがうかがえる．以上の結果より，係数 $\{a, b\}$ こそ変われど，モデル構造やデータセットによらず最適な温度パラメータ T^* は $T^* = aM + b$ によって算出できる可能性が理論的及び実験的に示唆された．

3.3　訓練済みモデルの出力分布

ここまで M と T の変化に伴う最終的な汎化性能への影響を考察してきたが，これらの結果がなぜ発生したのかと

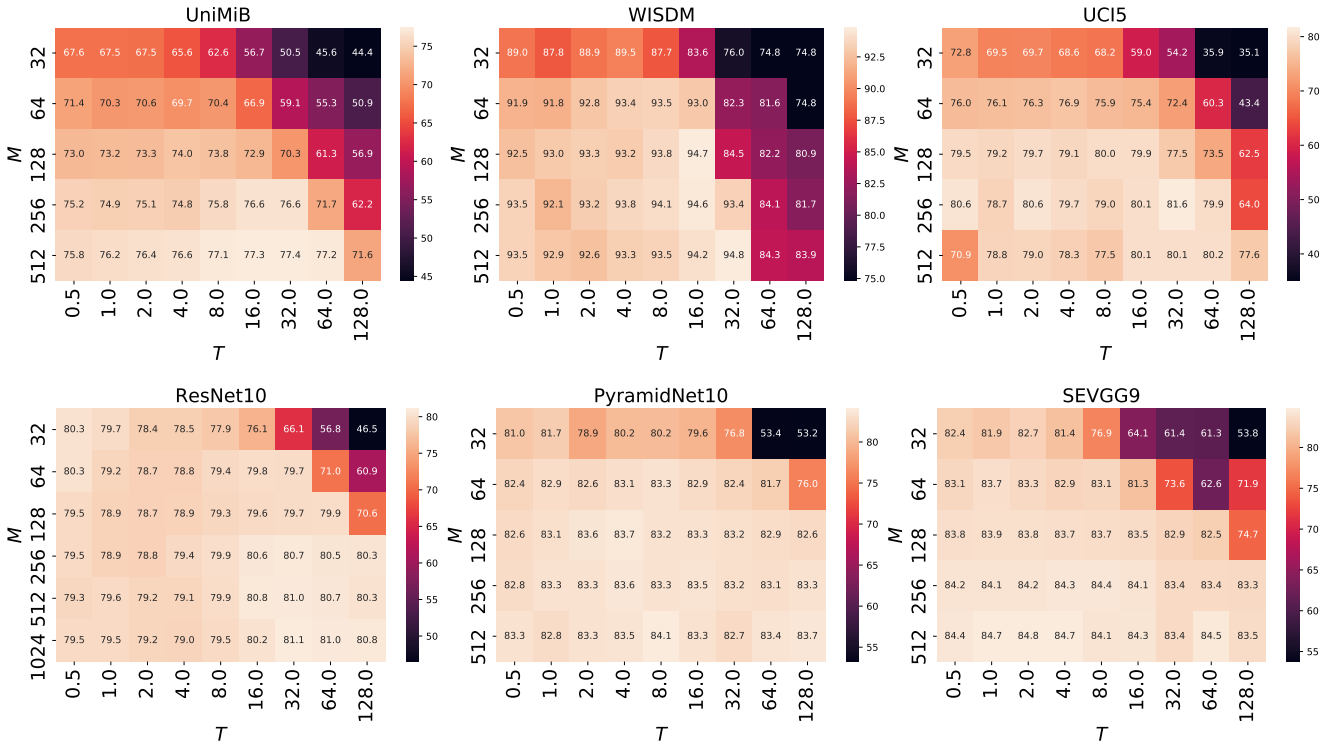

図 4 様々なデータセットとモデル構造の T と M ごとの推定精度（10 試行中央値）．上段は VGG9 で，下段は HASC で統一している．

Fig. 4 Accuracies in various datasets and model architectures for each T and M (median values of 10 trials). Each row's environment is unified (upper row: VGG model and lower row: HASC dataset).

いう点に言及する．WISDM データセットと VGG9 構造を用いたときの，様々な条件下における，$f, z/T, w$ の分布を**図 5** に示す．上から順に $T = 1$ の場合，$T = 16$ の場合，$M = 512$ の場合，$T = T^*$ と最適な温度パラメータが設定できた場合である．

はじめに上二行（$T = \{1, 16\}$）に着目する．いずれも特徴マップ f の分布は M によらず大きな変化がない様子がうかがえ，モデルの幅の広さによって f の分布は変わらないことが確認された．一方，f の分布が変わらないにも関わらず，softmax 関数の入力となる z/T の分布も大きく変動がないことから，w が M の増加による影響を低減するような働きが見られる．すなわち，M が大きくなるに連れて w の分散が小さくなっている．したがって，M が増加するほど僅かな損失の変化に機敏になるといえる．

次に三行目（$M = 512$）に着目する．まず，f の分布が T によって変動している様子がうかがえる．本稿では出力層付近の学習のダイナミクスに焦点を当て議論しているが，少なからず特徴抽出器 E にも T は影響を与えていると言える．更に，z/T に着目すると，M を変化させたときと比べて T の変化に対して分布が大きく異なっている事がわかる（x 軸のスケールが異なる点に注意されたい）．当然ながら T が大きくなるに連れ z/T の分散が小さくなっている．図 4 の結果とあわせると，$T \geq 64$ から推定精度が大幅に低下していることから，z/T の分散が小さくなりすぎると推定精度が低下する可能性が示唆された．

最後に，四行目（$T = T^*$）に着目する．これは図 4 の結果をもとに T^* が既知であると仮定した際の分布である（それぞれ $\boldsymbol{T}^* = \{4, 8, 16, 16, 32\}$）．特徴的な点は z/T の分布がいずれもほぼ同一で，$-5, 5$ の二点を平均とする二峰性の正規分布に見える点である．二峰性の理由は softmax ないしは sigmoid 関数でクラス分類を行うことに起因している．したがって，正解クラスのときに $z/T \sim \mathcal{N}(5, \gamma)$，不正解クラスのときに $z/T \sim \mathcal{N}(-5, \gamma')$ 付近に近づくような T が最適な温度パラメータ T^* である可能性が示唆された．

3.4 訓練済みモデルの出力分布と汎化性能

前節では特定の 1 試行における各パラメータの分布を考察したが，本節では各パラメータの分布から代表値を算出し，それぞれと汎化性能との関連を考察する．**図 6** に WISDM と VGG を用いた 10 試行における各パラメータの分布の代表値を示す．ここで，λ_f は f を指数分布と仮定した際の λ である．その他，$w, b, z/T$ の期待値と分散をプロットしている．図 6 より，各値が M と T によって変動しているが，特に $\mathbb{E}[b], \mathbb{V}[b]$，$\mathbb{E}[z/T]$ がテスト精度と同

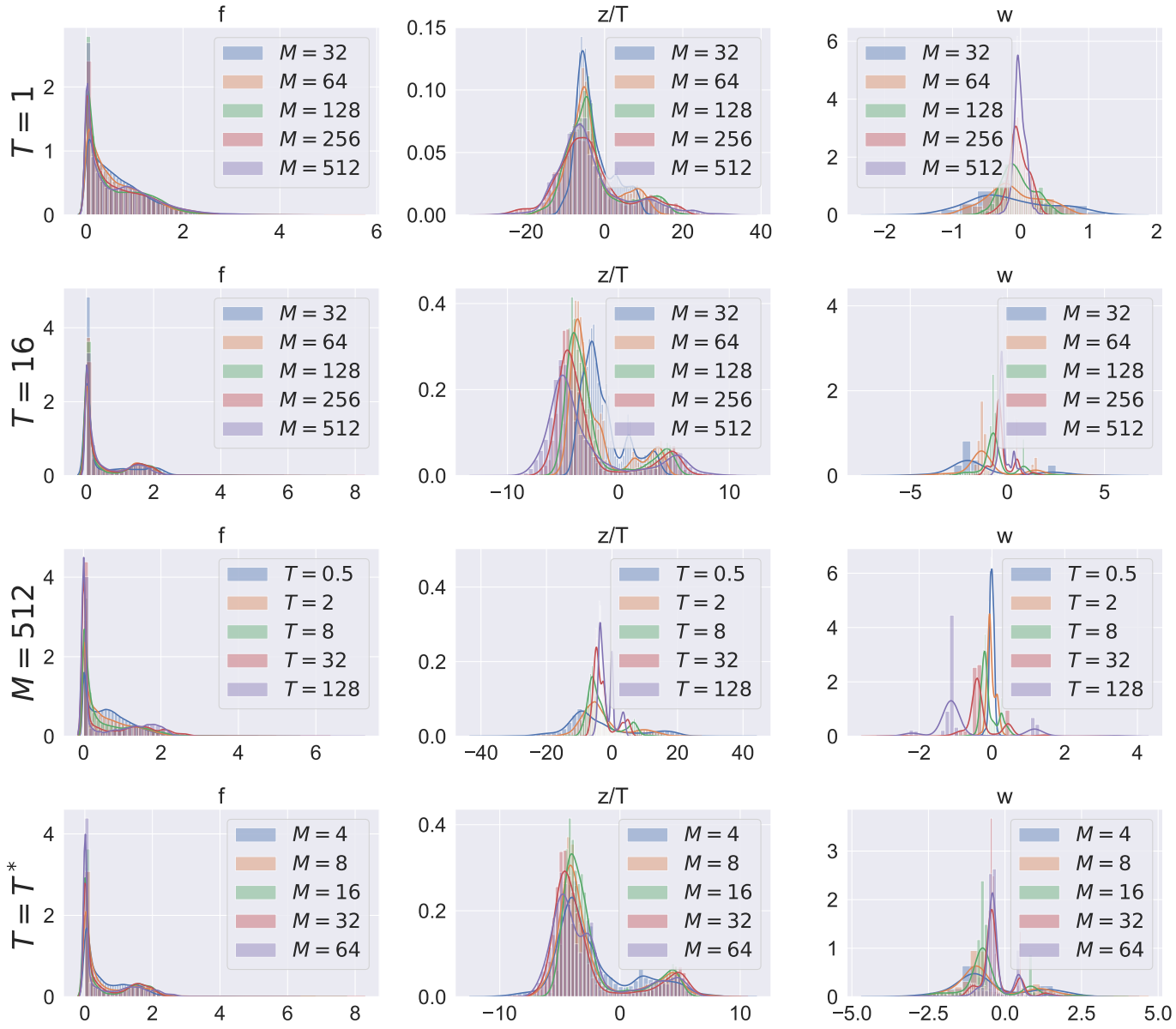

図 5 WISDM と VGG9 における出力とモデルパラメータの分布の比較

Fig. 5 Distribution comparison of outputs and model parameters in WISDM with VGG9.

じような分布を示している様子がうかがえる．特に，汎化性能が急激に低下するタイミングで，分類器 C のバイアス項 b が平均 0，分散 0 付近から大きく変動している．

更に，テスト精度と各パラメータの分布の代表値の相関係数を**表 1** に示す．結果，$\mathbb{E}[w], \mathbb{V}[w], \mathbb{E}[b], \mathbb{V}[b], \mathbb{E}[z/T]$ において正負の強い相関 $|R| > 0.6$ を確認している．したがって，これらの代表値を用いて T を最適化することで，最適な温度パラメータを動的に算出できる可能性がある．

4. おわりに

本研究では，深層学習による行動認識モデルの学習ダイナミクスを明らかにすることを目的に，softmax 関数の温度パラメータ T と，特徴マップの次元数 M に焦点を当てて議論を行った．行動認識分野では特にデファクトスタンダードなモデルが確立されておらず，M を含めてモデル構造を探索的に決定することが多いため，M と T の関係を明らかにすることが重要である．モデルの出力 z または分類器の重み w の確率分布は M の一次関数に従うという仮説を立て，温度パラメータ T を変更することでこれを改善できると考えた．行動認識ベンチマークデータセットを用いて，様々なモデル構造で T と M の関係を網羅的に調査する実験を行った．実験の結果，$T = 1$ の環境では特に M の大きいモデルの最善のパフォーマンスを発揮できていないことや，M の増大に伴い T を増加させることで，汎化

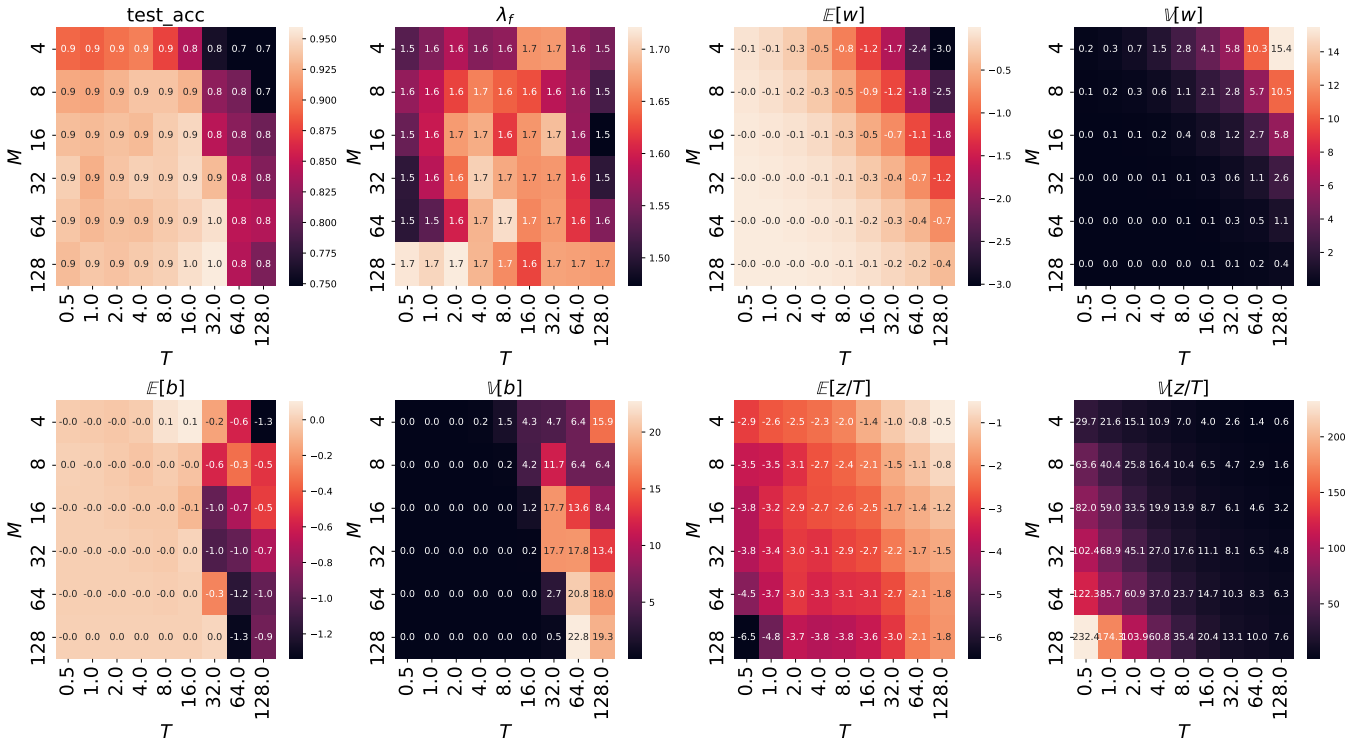

図 6 WISDM with VGG における T と M ごとの各種パラメータ（10 試行中央値）

Fig. 6 Various parameters in WISDM with VGG for each T and M (median values of 10 trials).

表 1 テスト精度と各パラメータ分布の代表値の相関係数

Table 1 Correlation coefficient between test accuracy and representative value of each parameter distribution.

	Corr.
λ_f	0.230
$\mathbb{E}[w]$	0.832
$\mathbb{V}[w]$	-0.767
$\mathbb{E}[b]$	0.605
$\mathbb{V}[b]$	-0.625
$\mathbb{E}[z/T]$	-0.709
$\mathbb{V}[z/T]$	0.365

性能が向上することを明らかにした．また，訓練後の出力やパラメータの分布を可視化することで本現象の原因を考察した．結果として，M の増大に伴い w の分散が小さくなることで z/T を一定に保つよう訓練が進んでいるが必ずしも最適な z/T を表現できてはいないことを確認した．更に，z/T の分布が特定の分布に従うように T を調整することで最適な温度パラメータが設定できる可能性が示唆された．今後の課題として，最適な T^* を訓練中に動的に決定する方法を模索するとともに，出力層のみでなく特徴抽出器に関しても T による影響を考察していく．

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] Lee, M.-W., Khan, A. M. and Kim, T.-S.: A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation, *Personal and Ubiquitous Computing*, Vol. 15, No. 8, pp. 887–898 (online), DOI: 10.1007/s00779-011-0403-3 (2011).

[2] Xu, H., Pan, Y., Li, J., Nie, L. and Xu, X.: Activity Recognition Method for Home-Based Elderly Care Service Based on Random Forest and Activity Similarity, *IEEE Access*, Vol. 7, pp. 16217–16225 (online), DOI: 10.1109/ACCESS.2019.2894184 (2019).

[3] 桑原教彰，春生野間，鉄谷信二，紀博萩田，潔　小暮，洋　伊関：ウェアラブルセンサによる看護業務の自動行動計測手法，情報処理学会論文誌，Vol. 44, No. 11, pp. 2638–2648 (2003).

[4] Kwapisz, J. R., Weiss, G. M. and Moore, S. A.: Activity Recognition Using Cell Phone Accelerometers, *SIGKDD Explor. Newsl.*, Vol. 12, No. 2, pp. 74–82 (online), DOI: 10.1145/1964897.1964918 (2011).

[5] Shoaib, M., Bosch, S., Incel, O. D., Scholten, H. and Havinga, P. J. M.: Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors, *Sensors*, Vol. 16, No. 4 (online), DOI: 10.3390/s16040426 (2016).

[6] Voicu, R.-A., Dobre, C., Bajenaru, L. and Ciobanu, R.-I.: Human Physical Activity Recognition Using Smartphone Sensors, *Sensors*, Vol. 19, No. 3 (online), DOI: 10.3390/s19030458 (2019).

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

[8] Mehmood, K., Imran, H. A. and Latif, U.: HARDenseNet: A 1D DenseNet Inspired Convolutional Neural Network for Human Activity Recognition with Inertial Sensors, *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pp. 1–6 (online), DOI: 10.1109/INMIC50486.2020.9318067 (2020).

[9] Ronald, M., Poulose, A. and Han, D. S.: iSPLInception: An Inception-ResNet Deep Learning Architecture for Human Activity Recognition, *IEEE Access*, Vol. 9, pp. 68985–69001 (online), DOI: 10.1109/ACCESS.2021.3078184 (2021).

[10] 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).

[11] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q.: On Calibration of Modern Neural Networks, *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, pp. 1321–1330 (2017).

[12] He, Y.-L., Zhang, X.-L., Ao, W. and Huang, J. Z.: Determining the optimal temperature parameter for Softmax function in reinforcement learning, *Applied Soft Computing*, Vol. 70, pp. 80–85 (online), DOI: https://doi.org/10.1016/j.asoc.2018.05.012 (2018).

[13] Agarwala, A., Pennington, J., Dauphin, Y. and Schoenholz, S.: Temperature check: theory and practice for training models with softmax-cross-entropy losses (2020).

[14] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *Proceedings of the International Conference on Learning Representations*, pp. 1–14 (2015).

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

[16] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Teh, Y. W. and Titterington, M., eds.), Proceedings of Machine Learning Research, Vol. 9, Chia Laguna Resort, Sardinia, Italy, PMLR, pp. 249–256 (online), available from ⟨https://proceedings.mlr.press/v9/glorot10a.html⟩ (2010).

[17] He, K., Zhang, X., Ren, S. and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015).

[18] Zhongkai, Z., Kobayashi, S., Kondo, K., Hasegawa, T. and Koshino, M.: A Comparative Study: Toward an Effective Convolutional Neural Network Architecture for Sensor-Based Human Activity Recognition, *IEEE Access*, Vol. 10, pp. 20547–20558 (online), DOI: 10.1109/ACCESS.2022.3152530 (2022).

[19] 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 2nd Augmented Human International Conference* (2011).

[20] D. Micucci, M. M. and Napoletano, P.: UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones, *Apld. Sci.,*, Vol. 7, No. 10 (online), DOI: 10.3390/app7101101 (2017).

[21] Anguita, D., Ghio, A., Oneto, L., Parra, X. and Reyes-Ortiz, J. L.: A Public Domain Dataset for Human Activity Recognition Using Smartphones, *In Proceedings of the 21st European Symposium on Artificial Neural Networks (ESANN)*, pp. 437–442 (2013).

[22] Han, D., Kim, J. and Kim, J.: Deep Pyramidal Residual Networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6307–6315 (online), DOI: 10.1109/CVPR.2017.668 (2017).

[23] Hu, J., Shen, L. and Sun, G.: Squeeze-and-Excitation Networks, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).