Vol.2019-CVIM-216 No.10
2019/3/8

Knowledge Distillationにおける温度パラメータの適正化に関する検討

石井 遊哉[1,a)]　中野 学[1]　井下 哲夫[1]　高橋 勝彦[1]

概要：
ディープニューラルネットワークのモデル圧縮手法のひとつに，knowledge distillation と呼ばれる手法がある．教師モデルのスコアを疑似的に底上げする温度パラメータは，一般に全学習サンプルで一定の値が用いられるが，サンプルごとにスコアの最大値や最小値は異なるため，適切な温度もサンプルごとに異なると考えられる．
本稿では，教師モデルのスコアに関する関数として温度を定義することで，サンプルごとに異なる温度を設定する手法を提案する．Cifar10 および Cifar100 で実験を行い，全サンプルに対して同一の温度を用いる従来手法と認識精度の比較を行った結果，Cifar100 において最大で 0.65% の精度向上を確認した．

キーワード：knowledge distillation，機械学習，深層学習

1. はじめに

深層学習はニューラルネットワーク（DNN）を用いた機械学習の一種であり，画像認識や自然言語処理，音響解析などで広く用いられている．特に畳み込み層を持つネットワーク（CNN）は画像認識の分野で目覚ましい認識精度を達成しており，顔認証や人物検知など，実世界への応用が期待されている．

一方で，一般に大型で高精度なネットワークほどパラメータ数が多くなり，学習済みモデルの保存容量が大きくなるほか，推論時の計算コストも高くなるため，大型の計算器が使用できない場面での運用が難しいという問題がある．この問題を解決すべく，より小型なネットワークで高い認識精度を得るための研究が進められている．

そのひとつに，knowledge distillation(KD)[1] と呼ばれる手法がある．KD は転移学習の一種とみなすことができる手法で，従来の一般的な正解ラベルに加え，学習済みの高精度なネットワークモデル（教師モデル）が出力した予測スコアも正解ラベルとして学習することで，より小さいモデル（生徒モデル）の認識精度を向上させることが可能となる．

[1] で KD の基本的な枠組みが提案されて以来，様々な応用手法が提案されている．[2] や [3] は KD を物体検知用のネットワークモデルへ応用する手法を提案した．[4] ではネットワークモデルの重みを量子化する手法と組み合わせることで，より小型かつ高精度なネットワークモデルを実現している．また，生徒モデルの学習データにノイズによる欠損がある場合や [5]，生徒モデルとは異なる学習データで学習した教師モデルを用いる場合 [6][7] にも有効な手法が提案されている．このように，より高精度な DNN を実世界で運用することが KD によって実現され始めている．

上述のいずれの手法においても，以下の基本的な枠組みは共通して用いられている．先述の通り KD は，通常の学習で用いられる one-hot な正解ラベル（hard targets）のほかに，大型で高精度な教師モデルの出力も正解ラベル（soft targets）として学習を行うことで，生徒モデルの認識精度を通常の学習時よりも上昇させる手法である．例えば，入力画像を [猫，犬，車] の 3 クラスに分類するネットワークを学習する際，通常の学習では正解ラベルを [1, 0, 0] といったバイナリ値で与えており，不正解クラスの犬と車は等価に扱われている．一方，教師モデルの出力するスコアは [0.8, 0.15, 0.05] といった連続値であるため，正解クラスが猫である情報のほかに，不正解クラスの中でも車より犬の方がスコアが高い，つまり犬の方が猫に似ている，といった付加的な情報を同時に学習することが可能となる．

ところが，教師モデルは高精度であるため不正解クラスのスコアがゼロに近いことが多く，cross entropy による Loss にほとんど寄与しない．そこで温度と呼ばれるハイ

[1] NEC 中央研究所
〒 211-8666 神奈川県川崎市中原区下沼部 1753
[a)] a-ishii@jg.jp.nec.com

ⓒ 2019 Information Processing Society of Japan

パーパラメータを新たに導入することで，疑似的に不正解クラスのスコアを底上げすることを行う．一般に温度は全学習サンプルに対して同一の値が用いられるが，学習サンプルごとに教師モデルのスコアは変化するため，それに応じて温度も変更することでより効率的な学習が行えると考えられる．しかしながら，温度を教師モデルのスコアに関する関数として具体的に提案する検討は行われていない．

本稿では，教師モデルのスコアに関する sigmoid 関数によって温度を定義する手法を提案し，生徒モデルの認識精度を従来手法と比較することでその有効性を検討する．

2. knowledge distillation

一般的な knowledge distillation(KD) の学習工程は以下のとおりである．

c クラスの分類問題に対して学習済みの教師モデルを $\boldsymbol{f}^t$，入力データとそのラベルを $\{\boldsymbol{x}, \boldsymbol{y}\}$ とする．

(1) 学習済みの教師モデル $\boldsymbol{f}^t$ に学習サンプル $\boldsymbol{x}$ を入力し，(1) 式の活性化関数を用いて各クラスの確率分布 $s_k^t(\boldsymbol{x};T)$ を出力する．また T は温度と呼ばれるハイパーパラメータであり，一般に全学習サンプルで同一の値を用いる．その役割については後述する．

$$s_k^t(\boldsymbol{x};T) = \frac{\exp[f_k^t(\boldsymbol{x})/T]}{\sum_{j=1}^{c}\exp[f_j^t(\boldsymbol{x})/T]} \tag{1}$$

ここに $\boldsymbol{f}^t$ は入力データを引数に取り，(1) 式で表される softmax 関数により活性化される前の最終層の出力 (logits) を返すものとする．

(2) 同じサンプルを生徒モデル $\boldsymbol{f}^s$ に入力し，(1) 式と同様にして確率分布 $s_k^s(\boldsymbol{x};T)$ を出力する．このとき，教師モデルと同じ温度 T を用いる．

(3) hard targets による cross entropy と，soft targets による cross entropy との加重平均を Loss 関数として，生徒モデルの学習を行う．

$$E = \lambda\sum_{k=1}^{c} C(y_k, s_k^s(\boldsymbol{x};T)) + (1-\lambda)T^2\sum_{k=1}^{c} C(s_k^t(\boldsymbol{x};T), s_k^s(\boldsymbol{x};T)) \tag{2}$$

$$C(y_k, s_k^s(\boldsymbol{x};T)) = -y_k\ln(s_k^s(\boldsymbol{x};T)) \tag{3}$$

ここに λ は加重平均の重みを表す．

なお，学習後の生徒モデルを推論に用いる際は $s_k^s(\boldsymbol{x};T=1)$ として活性化を行う．

以上が一般的な KD の学習工程である．(2) 式の Loss 関数により，生徒モデルは hard targets による正解クラスか不正解クラスかのバイナリ値に加えて，soft targets による各クラスがそれぞれどの程度似ているかといった情報を学習することが可能となる．

この工程において，温度パラメータは以下のような役割を担っている．(2) 式から，soft targets は Loss 関数において各クラスに重みを置いているとみなすことができる．ところが，教師モデルが高精度であるほど soft targets は one-hot に近く，不正解クラスのスコアはゼロに近い値である．そのため KD を用いない一般の学習過程と同様に $T=1$ とした場合，不正解クラスに対する Loss も非常に小さくなり，学習にほとんど寄与しなくなってしまう．そこで温度パラメータ $T \geq 1$ を (1) 式の活性化関数に導入することで，不正解クラスに対する出力を疑似的に底上げし，全クラスにわたってなだらかに丸めることができる．

ここで，Cifar10 を学習したあるモデルが異なる 2 つの学習サンプルに対して $T=1,2,5$ の場合にそれぞれ出力する soft targets を**図 1** と**図 2** に示す．図 1 では $T=1$ では one-hot に近い分布になっており，$T=2$ でもクラス 3 などの値は依然小さく，Loss に寄与できないと考えられる．一方で，図 2 は $T=1$ でも十分なだらかであり，$T=5$ では全クラスでほとんど同一の値になってしまっている．仮に hard targets の比重が $\lambda=0$ であったとすると，学習は全クラスで同一の重みのもと (2) 式を最小化することとなり，学習後の生徒モデルの認識精度は chance rate 前後に収束してしまう．

このように，教師モデルのスコアに応じて適切な温度パラメータはサンプルごとに変化すると考えるのが妥当であるが，それに対応するための具体的な手法は提案されていない．

3. 手法

教師モデルの出力が one-hot に近いサンプルほど高い温度となるような温度関数を定義する．本稿では，教師モデルの one-hot 性を次のように評価した．教師モデルの出力するスコアのうち，最も値の高いものを S_1，次に値の高いものを S_2 とする．S_1 と S_2 の比によって $r = S_1/S_2$ を定義し，r が大きいほど one-hot 性が高いものとした．定義より，常に $r \geq 1$ である．

次に，温度関数の具体的な形状を設計する．一般に温度は $T \geq 1$ とされており，one-hot 性が最も低い $r=1$ で $T=1$ をとるほか，one-hot 性が十分低い $r \simeq 1$ においても，緩やかな変化率で $T \simeq 1$ をとれるような設計である必要がある．また，$T \to \infty$ において全クラスのスコアが chance rate に収束するため，任意の r に対して温度は有限の値を取る必要がある．以上の理由により，温度関数は以下の 3 点を満たす必要がある．(1) 単調増加関数である．(2)r が十分小さいときに有限の値に収束する．(3)r が十分大きいときに有限の値に収束する．上記の 3 点を満たす関数として，本稿では (4) 式の sigmoid 関数を提案する．

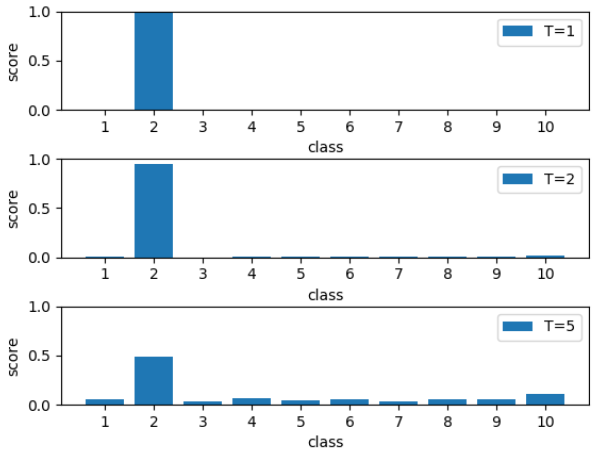

図 1 学習済みのあるモデルが Cifar10 のある学習サンプルに対して出力するスコア．活性化時の温度はそれぞれ $T = 1$（上），$T = 2$（中），$T = 5$（下）．

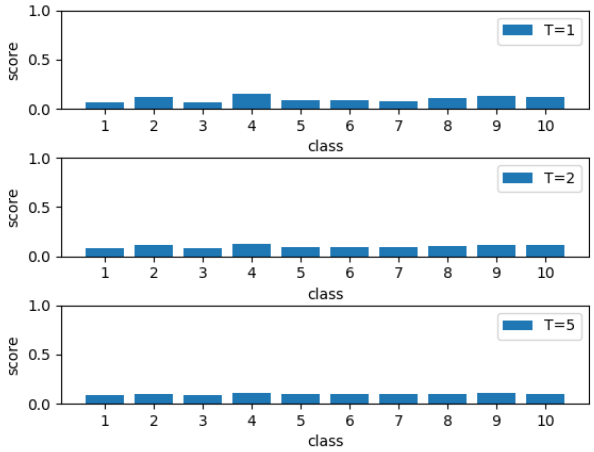

図 2 学習済みのあるモデルが図 1 とは別の学習サンプルに対して出力するスコア．活性化時の温度はそれぞれ $T = 1$（上），$T = 2$（中），$T = 5$（下）．

$$T(r) = \frac{a}{1 + \exp[c(-r + r_0)]} + b \tag{4}$$

a, b, c, r_0 が関数の詳細な形状を決定するパラメータである．**表 1** に，本稿で用いたパラメータの組を示す．a，b の値は直接設定せず，代わりに通るべき 2 点 $(r_0, T(r_0))$ と $(1, T(1))$ を定め，連立方程式を解くことによって求めた．簡単のために，表 1 にはこれら 2 点の値を表記している．

r_0 は全サンプルで求めた r の平均値をもとに，$T(r_0)$ は後述のグリッドサーチによる従来手法において最高精度となった温度をもとに定めた．なお，$T(\infty)$ は以上のパラメータで自動的に決定される値であるが，参考のために附記する．また表中にあるように，それぞれのパラメータの組に対し便宜的な名前を付ける．func1 から func4 のグラフを**図 3** に，func1 と func2 の部分を拡大したグラフを**図 4** にそれぞれ示す．

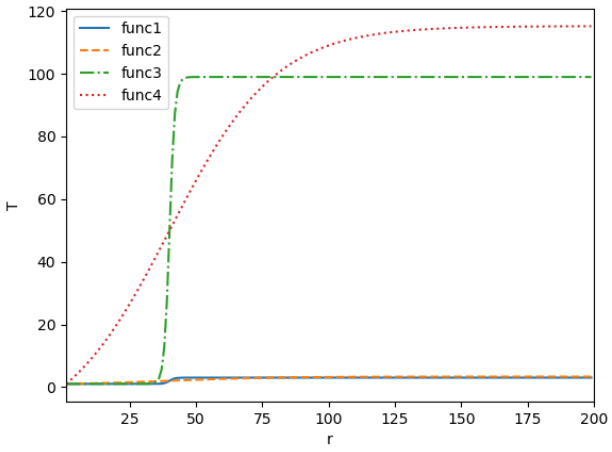

図 3 func1 から func4 のグラフ形状

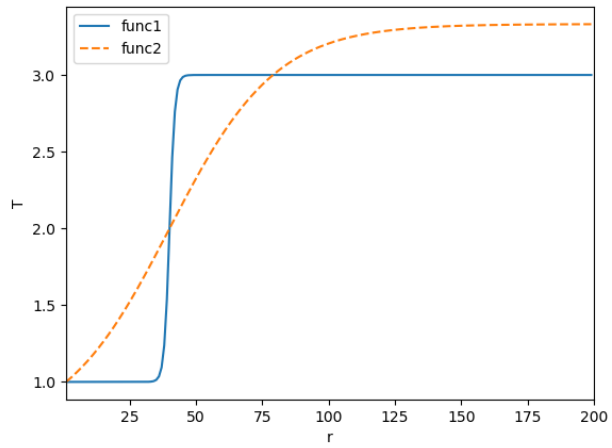

図 4 func1 と func2 のグラフ形状

4. 実験

教師モデルを ResNet50[8]，生徒モデルを MobileNet[9]（width multiplier $\alpha = 1, 0.5$ の 2 種）とし，Cifar10 および Cifar100 でそれぞれ実験を行った．全ての実験において，hard targets と soft targets の比は $\lambda = 0.5$ とした．

本手法の有効性を評価する上での比較対象を，全サンプルで一定の温度を用いる一般的な KD とし，次の 2 パターンを用意した．一つは $T = 1, 2, 3, 4, 5, 10, 50, 100$ によるグリッドサーチで，一般に KD を行う際はこれらのうち最も精度の良いものを選ぶ．もう一つは提案手法でサンプルごとに算出された温度の平均値を全サンプルに対して用いる従来手法で，提案手法とより近い温度設定となることから，温度がサンプルごとに変化することの有効性はこちらがより公平に評価できる．

Cifar10 での実験結果を 4.1 節で，Cifar100 での実験結果を 4.2 節で述べる．

ⓒ 2019 Information Processing Society of Japan

表 1　実験に用いた sigmoid 関数の特徴

	r_0	c	$T(1)$	$T(r_0)$	$T(\infty)$
func1	40	1	1	2	3
func2	40	0.05	1	2	3.33
func3	40	1	1	50	99
func4	40	0.05	1	50	115

表 2　Cifar10 における認識精度

Model	T	Width multiplier	
		$\alpha = 1$	$\alpha = 0.5$
ResNet50 (Teacher)	–	81.78	81.78
MobileNet (Baseline)	–	70.38	78.67
MobileNet (Fixed T)	1	77.45	81.73
	2	**77.46**	81.32
	3	76.93	80.13
	4	76.86	81.46
	5	76.54	82.80
	10	76.67	77.53
	50	76.48	**83.07**
	100	76.77	82.83
MobileNet (Variable T)	$1 \leq T \leq 3$	74.32	80.77
	$1 \leq T \leq 3.3$	74.35	81.59
	$1 \leq T \leq 99$	74.62	80.87
	$1 \leq T \leq 115$	74.40	81.53
MobileNet (Fixed T_{mean})	1.65	74.50	77.64
	1.79	74.42	80.81
	32.7	73.89	78.61
	39.6	73.95	81.68

4.1　Cifar10

従来手法と提案手法による実験結果をそれぞれ**表 2** に示す[*1]．Teacher は教師モデルの，Balseline は蒸留を用いない一般の学習による生徒モデルの認識精度をそれぞれ表し，Fixed T がグリッドサーチによる従来手法，Variavle T が提案手法，Fixed T_{mean} が提案手法から得られた温度の平均値による従来手法に対応する．

グリッドサーチによる従来手法では，$\alpha = 1$ では $T = 2$ で 77.46%，$\alpha = 0.5$ では $T = 50$ で 83.07% がそれぞれ最高精度となった．

これに対し，提案手法では $\alpha = 1$ で 74.62%，$\alpha = 0.5$ で 81.59% がそれぞれ最高精度となり，いずれもグリッドサーチを下回る結果となった．各温度関数において算出された温度頻度分布を**図 5** に示す．

一方，提案手法で得られた温度の平均値による従来手法と比較すると，$\alpha = 1$ では func3 と func4 で，$\alpha = 0.5$ では func1，func2，func3 で提案手法が従来手法を上回ることが確認できた．

[*1] MobileNet は一般に $\alpha = 1$ の方が $\alpha = 0.5$ より高精度であるが，本実験では逆の結果となっている．$\alpha = 1$ では ImageNet の pretrained モデルを最終層のみ fine-tuning した一方で，$\alpha = 0.5$ ではスクラッチで全層の学習を行った影響であると考えられる．

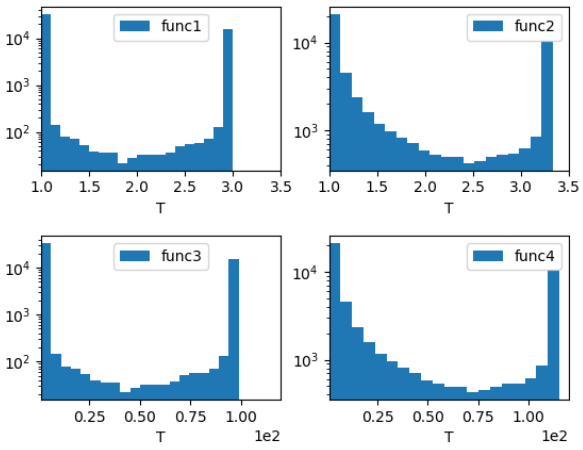

図 5　Cifar10 において各温度関数で設定された温度の頻度分布

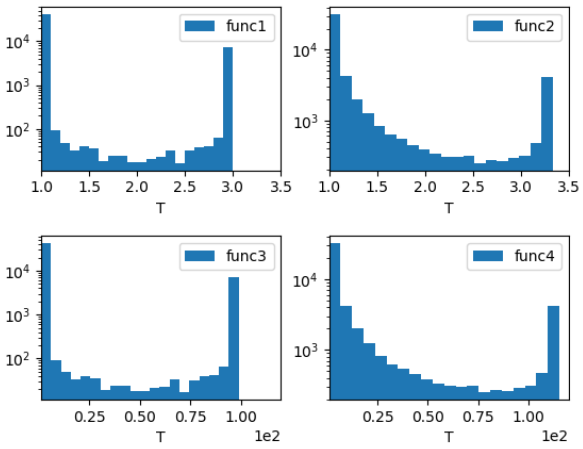

図 6　Cifar100 において各温度関数で設定された温度の頻度分布

4.2　Cifar100

従来手法と提案手法による実験結果をそれぞれ**表 3** に示す．Cifar10 での場合と同様に，$\alpha = 1$ では $T = 2$ で 56.83%，$\alpha = 0.5$ では $T = 50$ で 54.56% とそれぞれ最高精度となった．

これに対し，提案手法では $\alpha = 1$ においては func1 から func4 のすべてにおいて，$\alpha = 0.5$ では func4 でグリッドサーチを上回る認識精度となり，それぞれ最高で 57.48%，56.31% を得た．表中ではグリッドサーチでの最高精度を上回った結果を太字で示している．各温度関数において算出された温度の頻度分布を**図 6** に示す．

また，提案手法で得られた温度の平均値による従来手法においても，$\alpha = 1$ および $\alpha = 0.5$ の全ての温度関数において提案手法が従来手法を上回っており，提案手法の有効性が確認できた．

ⓒ 2019 Information Processing Society of Japan

表 3　Cifar100 における認識精度

Model	T	Width multiplier	
		$\alpha = 1$	$\alpha = 0.5$
ResNet50 (Teacher)	–	60.29	60.29
MobileNet (Baseline)	–	54.56	36.61
MobileNet (Fixed T)	1	55.94	50.10
	2	**56.83**	47.13
	3	54.89	51.60
	4	54.83	52.73
	5	54.96	45.65
	10	56.31	47.04
	50	55.63	**54.56**
	100	56.29	51.33
MobileNet (Variable T)	$1 \leq T \leq 3$	**57.30**	52.87
	$1 \leq T \leq 3.3$	**57.31**	54.35
	$1 \leq T \leq 99$	**57.48**	53.79
	$1 \leq T \leq 115$	**57.10**	**56.31**
MobileNet (Fixed T_{mean})	1.30	55.62	50.54
	1.39	55.34	49.36
	41.7	55.92	41.67
	20.1	56.51	48.97

4.3　考察

提案手法で得られた温度の平均値による従来手法と比較すると，Cifar10 と Cifar100 のいずれにおいても提案手法がより良い認識精度を与えることから，サンプルごとに温度を変化させる手法は有効であることが確認できた．

一方で，Cifar100 では提案手法がグリッドサーチを上回ったものの，Cifar10 では下回る結果となった．Cifar100 において，グリッドサーチでは低温側と高温側で比較的高精度に，中間で低精度となる傾向が見られる．これは高温側では one-hot なサンプルを，低温側では $T = 1$ で十分なだらかなサンプルをそれぞれ効率的に学習できた一方で，中間ではどちらに対しても中途半端な温度となっていることが一因と考えられる．これに対し，提案手法ではいずれのサンプルでも効率的に学習ができたため，従来手法よりも良い認識精度が得られた．一方 Cifar10 の従来手法では温度による精度の違いが比較的小さく，本実験で定義した温度関数とパラメータでは効果的に温度を変えることができなかったものと考えられる．図 5 と図 6 を比較しても，両データセットで選択された温度に大勢の違いはないことから，適切な温度関数やパラメータはデータセットごとに変化することが考えられる．

5.　まとめ

本稿では，教師モデルのスコアに基づき knowledge distillation の温度パラメータを sigmoid 関数で変化させる手法を提案し，その有効性を評価した．提案手法で算出された温度の平均値による従来手法と比較すると認識精度の向上が確認でき，温度をサンプルごとに変化させることは有効であることが確認できた．一方グリッドサーチによる従来手法との比較では，Cifar100 で従来手法を最大 0.65% 上回ったものの，Cifar10 では全ての実験で下回る結果となり，データセットや教師モデル，生徒モデルによって適切なパラメータは変化することが示唆された．

教師モデルの one-hot 性の評価方法や温度関数の定義方法とそのパラメータの選び方など，検討すべき課題は多く残されている．ハイパーパラメータの選び方については，Nelder-Mead 法による探索 [10][11] などが挙げられる．また，本手法は正解ラベルにおける各クラスの重みを適正化するという意味において，hard targets と soft targets の重み λ とも密接に関りがあると考えられ，T と λ の両方を変化させる手法も今後の検討課題のひとつに挙げられる．

参考文献

[1] Hinton, G., Vinyals, O. and Dean, J.: Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531* (2015).
[2] Chen, G., Choi, W., Yu, X., Han, T. and Chandraker, M.: Learning efficient object detection models with knowledge distillation, *Advances in Neural Information Processing Systems*, pp. 742–751 (2017).
[3] Zhou, G., Dulloor, S., Andersen, D. G. and Kaminsky, M.: EDF: Ensemble, Distill, and Fuse for Easy Video Labeling, *arXiv preprint arXiv:1812.03626* (2018).
[4] Polino, A., Pascanu, R. and Alistarh, D.: Model compression via distillation and quantization, *arXiv preprint arXiv:1802.05668* (2018).
[5] Lopez-Paz, D., Bottou, L., Schölkopf, B. and Vapnik, V.: Unifying distillation and privileged information, *arXiv preprint arXiv:1511.03643* (2015).
[6] Lopes, R. G., Fenu, S. and Starner, T.: Data-Free Knowledge Distillation for Deep Neural Networks, *arXiv preprint arXiv:1710.07535* (2017).
[7] Kulkarni, M., Patil, K. and Karande, S.: Knowledge distillation using unlabeled mismatched images, *arXiv preprint arXiv:1703.07131* (2017).
[8] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).
[9] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint arXiv:1704.04861* (2017).
[10] Nelder, J. A. and Mead, R.: A simplex method for function minimization, *The computer journal*, Vol. 7, No. 4, pp. 308–313 (1965).
[11] 渡邊修平，尾崎嘉彦，大西正輝：Nelder-Mead 法の並列化による識別器のハイパパラメータチューニングの高速化，研究報告コンピュータビジョンとイメージメディア (CVIM)， Vol. 2019, No. 8, pp. 1–6 (2019).

ⓒ 2019 Information Processing Society of Japan