Vol.2017-ARC-228 No.3
2017/11/7

画像認識における効率的な転移学習のための学習モデル選択手法の検討

上野 洋典[1,a)]　東 耕平[1]　近藤 正章[1]

概要：畳み込みニューラルネットワーク（CNN）に代表される深層学習は，一般物体認識の分野において目覚ましい成果を上げ注目されている．CNN の学習には大量のラベル付きデータが必要となるが，実環境での応用を考えると，少量の訓練データしか利用できない場合も多い．その際，ImageNet に代表される大規模物体認識データセットによってあらかじめ学習されたパラメータを初期値として，適用先のデータセットでそのモデルの再学習をおこなう転移学習（Fine-tuning）が用いられることが一般的である．本稿では，転移学習の際に元になるモデルと，ターゲットとなるデータセットとの親和性をモデルの類似度として定量化し，効率的な転移学習を行うためのモデル選択指標を得ることを検討する．評価の結果，提案した類似度指標によりターゲットタスクと類似度が高いと評価されたモデルを元にして fine-tuning を行うことで，高い認識性能が少ない学習回数で得られることがわかった．

1. はじめに

深層ニューラルネットワーク (Deep Neural Network: DNN）は，多層のニューラルネットワークを用いた機械学習モデルであり，コンピュータビジョン [2]，音声認識 [5]，自然言語処理 [5] などの様々な分野でそれぞれ高い性能が報告されている．

DNN の一種である畳み込みニューラルネットワーク (Convolutional Neural Network: CNN) は，畳み込み層とプーリング層など，特に画像処理を指向した演算を行う層を含む DNN の一種であり，一般物体認識の分野において目覚ましい成果を上げ注目されている．将来的に物体の認識が高い精度で可能になれば，例えば周りの物体を認識しながら行動し人間の身の回りの世話をするようなロボットなど様々な場面で応用可能になると考えられる．

現在の物体認識技術の研究では特に汎化性能の向上に対して注力され，非常に多くのクラスの画像を高い精度で認識できるようになっている．一方で，汎化性能の高いモデルを作成する上では以下のような課題がある．

まず第一に学習にかかる計算コストの問題である．CNN が画像認識において成功を収めた理由の一つは，ニューラルネットの層を深くすることでモデルが高い表現力を学習できたことであると言われている．例えば画像認識コンペティション ILSVRC2015 の勝者である ResNet は 152 もの層を持つモデルである．しかし，層の数が多くなるにつれてモデルのパラメータ数は増大し，1 回の学習にかかる時間も非常に大きくなる [4]．近年発表されているモデルでは，学習に数日から数週間かかることも珍しくはない．学習を効率良く行うことが実際の応用では不可欠となる．

第二に学習に使える訓練データ数の問題である．CNN によって高い画像認識精度を持つモデルの構築には大量のラベル付きデータが必要となる．データへのラベル付けは基本的に人間の手で行われるため，訓練データの作成は非常にコストがかかる．さらに，物体検出やセグメンテーションなどのより高次な画像認識を行う際にはラベル付のコストはさらに大きくなる．認識する物体のクラスが多い場合にはより多くの訓練データが必要となる．

先に述べた応用を考えた場合，搭載する CNN は現実世界にあるすべての物体を認識できる必要はなく，例えば対象ロボットが活動する環境に存在する物体のみを認識できれば十分である．また，そのロボット自身がセンサを用いてその環境のデータを取得することで，その場で学習し環境に適応することも可能である．このようにモデルの汎化性能を向上させるのではなく，局所的な環境に適応させる手法が実応用において有効であると考えられる．しかし，実際に画像認識を行いたい環境において十分な数の訓練データを収集することは難しい場合も多い．また，学習にかかる計算コストを削減することも実用上は重要である．

これらの理由から，少ない訓練データ，低い計算コストでモデルを環境に適応させることが求められている．このような場合にはゼロから学習を行うのではなく，ImageNet に

[1] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
[a)] ueno@hal.ipc.i.u-tokyo.ac.jp

© 2017 Information Processing Society of Japan

代表される大規模なデータセットを用いて事前に学習したモデルを，対象となるタスクに適応するように微調整する，fine-tuning が有効であることが知られている．fine-tuning を行うことで，認識したい物体について少数の訓練データしか用意できない場合でも，高い精度で認識することが可能になる．このように，ある領域において学習させたモデルを別の領域に転用し適応させることを一般に転移学習という．

転移学習では，事前の学習に用いられるタスクをソースタスク，適応先のタスクをターゲットタスクと呼ぶ．ソースタスクとターゲットタスクの関連性が高いほど，転移学習が成功しやすいと考えられている [12][13]．

fine-tuning により物体認識の精度が向上する理由の一つとして以下が考えられている．CNN において各畳込み層は特徴マップを出力するが，入力に近い層ほどデータに依らない汎用的な特徴を，出力に近い層ほどデータセットに依存した具体的な特徴を学習していると言われている [10]．そのため，予め大規模なデータセットを使って学習したモデルは，あらゆる画像認識において有効である普遍的な特徴を学習していると考えられる．認識したい物体についての訓練データを用いて学習を行い，そのモデルの出力に近い層のパラメータを更新することでデータセットに依存した具体的な特徴を学習し，すでに学習していた汎用的な特徴と合わせることで目的の物体を認識できるようになると考えられる．

本稿では画像認識問題において fine-tuning を効率的に行うために，転用するモデルの選択指標について検討する．ソースタスクのモデルとターゲットタスクのモデル同士を比較し，モデル同士の類似度を定義することで，上記の研究目的の達成を試みる．

2. 関連研究

本章では転移学習を効率的に行うことを目的とした関連研究について述べる．

2.1 特徴マップによる方法

ニューラルネットワークが従来の機械学習に比べて高い画像認識能力を得ることができた理由のうちの一つに，ネットワークが特徴抽出とパラメータ学習を同時に行うため，人間が特徴量を設計する必要がないということ点がある．一方でニューラルネットワークによって学習された特徴量を人間が解釈できないという問題点もある．そこで CNN の中間層を可視化することで特徴量を解釈し，CNN の挙動を理解するアプローチが提案されてきた [10]．

Bau, Zhou らによる Network Dissection[11] は CNN の特徴マップを見て，そのモデルがどの程度の「識別能力」を持っているかを定量的に評価することで CNN の挙動の理解しようとした研究である．この研究では CNN に画像を入力した時に畳み込み層の出力する特徴マップについて，セマンティックセグメンテーションの手法を用いて，その特徴マップがある概念（クラス）を識別できているかどうかを判断する．モデル全体で識別できた概念の数を，そのモデルの識別能力として定量的に評価している．また，この研究により，入力に近い層が汎用的な特徴を，出力に近い層が具体的な特徴を学習しているという主張が正しいことが確かめられたと報告されている．

本稿での目的においても，モデルの類似度を評価するために，各モデルの学習した特徴量を解釈することは重要であると考えられる．

2.2 クラス分類確率による方法

同一の入力画像に対する各モデルの Softmax 出力，すなわち各クラスに分類される確率同士の類似度を定義する手法も提案されている．Frogner らの研究 [14] は Softmax 出力同士の距離指標として Earth mover's distance(EMD) を用いている．EMD は輸送最適化問題の考え方に基づいて定義された分布間の距離尺度である．分布 P， Q の間の EMD は以下の輸送最適化問題を解くことで得られる f^*_{ij} を用いて，(7) 式のように書ける．

$$\text{minimize } W = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \tag{1}$$

$$\text{subject to } f_{ij} \geq 0 (1 \leq i \leq m, 1 \leq j \leq n) \tag{2}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i} (1 \leq i \leq m) \tag{3}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{q_j} (1 \leq j \leq n) \tag{4}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}) \tag{5}$$

$$\tag{6}$$

$$\text{EMD}(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f^*_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f^*_{ij}} \tag{7}$$

ここで m, n はそれぞれ P, Q の次元，d_{ij} は P_i と Q_j の間の距離であり事前に与えられる．f_{ij} は P_i から Q_j への流量を表し，総仕事量 W を最小化するために最適化される変数である．計算された EMD が小さいほど 2 つの分布 P, Q は類似度が高いことを意味する．

Frogner らの研究 [14] ではマルチラベル予測問題にこの EMD を損失関数として用いている．この問題設定において，分布 P, Q をそれぞれクラスに属する確率の予測値と実際の値とし，d_{ij} は word2vec により計算された P_i, Q_j のクラス名の類似語ベクトルのユークリッド距離である．クラス名同士の類似度が低いほどベクトル間のユークリッド距離は大きくなるので，輸送最適化における重みが大きく

© 2017 Information Processing Society of Japan

なることに相当する．具体的に例示すると，真のクラスが boat である画像の分類に失敗した場合，lake と分類するよりも bird と分類することのほうがより強い罰則を課されることとなる．損失関数を EMD とすることで，マルチラベル予測問題で一般的な KL 損失を損失関数として用いた場合よりも高い学習性能となっている．

また，2 つのモデルの Softmax 出力同士の類似度を EMD を使って定義した研究に Lu ら [9] によるものがある．例えば AlexNet による画像クラス分類問題における fine-tuning を考えると，ソースタスクとターゲットタスクのモデルは最後の全結合層以外は共通で，モデルの出力は Softmax 関数の出力である．この 2 つのモデルの Softmax 出力同士の EMD 距離を計算し，各モデルの予測値とラベルのクロスエントロピー誤差にこの EMD を加えたものをロス関数として学習を行う．各モデルの予測誤差を抑えつつ，両モデルの出力を近づけようとする方向に学習が進む．

ターゲットドメインに含まれる画像を，2 つのモデルに入力した際のそれぞれの Softmax 出力の類似度の指標となるロス関数を定義し，それを小さくする方向に学習を進める．この手法を用いることで従来手法よりも効率的に転移学習を進めることができたと報告されている．

本稿ではこの EMD を用いることで，各モデルとターゲットタスクの類似度を定義し，その類似度が効率的な点学習のためのモデル選択の指標として妥当であるかどうかを確認する．

3. 提案手法

問題設定として，規模の大きなデータセットのクラス分類用に設計されたモデルをより規模の小さいデータセットのクラス分類問題に転用することを考える．本章ではターゲットタスクと転用元のモデルの類似度を定義する手法について述べる．

以下，ターゲットタスクのデータセットの画像データおよび画像データ集合を x_i および X とする．また，x_i の属するクラス名およびクラスの番号を l_i および t_i とし，画像 x_i を転用元のモデルに入力した際の Softmax 出力を $f(x_i)$ とする．また，転用元のモデルとターゲットタスクの出力の次元をそれぞれ n_s, n_t とする．この n_s, n_t はそれぞれソースタスクのデータセットのクラス数，ターゲットタスクのデータセットのクラス数に相当する．

3.1 Softmax 出力の最大値に基づく方法

本節では Softmax 出力の最大値に着目した類似度指標を提案する．転用元のモデルに対してターゲットタスクのデータセット中の画像 x_i を入力することにより，たとえ対象画像のクラスを用いた学習が行われていないとしても，CNN の最終的な出力として Softmax 関数が用いられている場合は，ソースタスク中の各クラスに分類される確率が出力される．通常は出力中で最大の確率を持つクラスが分類クラスとして選択される．

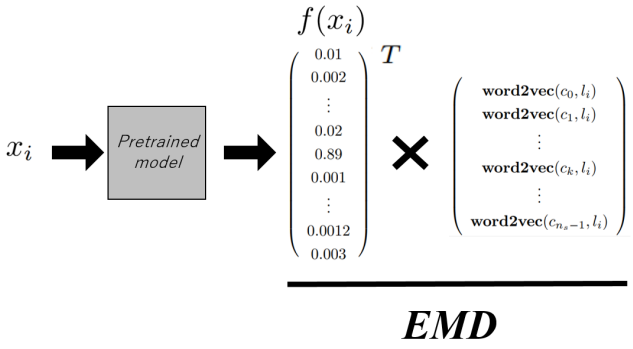

図 1　EMD 計算方法の概要

ここで、その分類されたクラスの確率の値が大きい場合は，そのクラスに属することを判定しやすいネットワークが構築されていることになる．そのため，Softmax 出力である $f(x_i)$ の各要素中で最大の値が大きいほど，そのモデルはターゲットタスクにとって効率良い学習が行われてると見なすことができ，転用元の画像とターゲットタスクの画像の分類のしやすさが近い，すなわち類似度が高いと考えられる．そこで，以下の式を類似度指標として用いる．

$$\sum_i \mathbf{max}\ v_k$$

ここで，$\boldsymbol{v} = f(x_i)$ で v_k は $\boldsymbol{v}$ の k 番目の要素を表す．この数値が高いほど，転用元のモデルとターゲットタスクの類似度が高いとする．

一方で，ImageNet のようにもともとのクラス数が多く，クラス名が例えば「bird（鳥）」や「airplane（飛行機）」のような抽象的な分類ではなく，「limpkin（ツルモドキ）」や「spoonbil（ヘラサギ）」のような具体的なものである場合には，Softmax 出力の値は相対的に低い値をとることが予想される．また，実際のクラス分類が間違う，あるいは似通っていないものの確率が高いと認識された場合も，本手法で考慮されるのは各クラスに分類される確率の最大値のみであるため，正しい，あるいは近そうなクラスに分類されたかどうかは全く考慮されない．この点で，本指標の有効性が制限される可能性がある．

3.2 EMD に基づく方法

本節では 2.2 節で述べた EMD に基づいた類似度指標を説明する．この場合，以下の手順で転用元モデルとターゲットタスクの類似度を算出する．

(1) ターゲットタスクとソースタスクの各クラス名の間の距離を word2vec を用いて算出する
(2) ターゲットタスクの訓練データ x_i を転用元のモデルに入力し，Softmax 出力 $f(x_i)$ を得る
(3) (1) で求めたクラス名の間の距離を重みとして，Softmax 出力 $f(x_i)$ と画像 x_i に対応するクラス番号 t_i の one-hot ベクトル表現の EMD を算出する

© 2017 Information Processing Society of Japan

(4) ターゲットタスクの訓練データ全てについて EMD を算出し，その合計をモデルとターゲットタスクの距離とする

図 1 に EMD 算出のフローチャートを示す．上記手順で求めた EMD 値が小さいほど，転用元のモデルとターゲットタスクの類似度が高いと考えられる．

なお，実際の EMD の計算は，d_{ij} = {ソースタスクのクラス i の名前とターゲットタスクのクラス j の名前の類似語ベクトル間のユークリッド距離} となる行列 $d_{ij} \in \mathbb{R}^{n_s \times n_t}$ をあらかじめ求めておき，

$$\sum_i f(x_i)^T \times \boldsymbol{w_i}$$

を計算することで求めることが可能である．ただし，$\boldsymbol{w_i}$ は行列 d_{ij} の t_i 列目の列ベクトルである．

この類似度指標は，ソースタスクのクラスとターゲットタスクのクラスに共通のもの，あるいは似たようなものが多いほど小さい値になりやすいことが予想される．これは 1 章で述べた，ソースタスクとターゲットタスクの関連性が高いほど転移学習が成功しやすい，という点を反映できる手法であると考えられる．

4. 評価

本章では提案する類似度指標が，効率的な転移学習を行うためのモデル選択指標として有効であるかどうかについて評価を行う．

4.1 評価手法

評価においては，転用元となるモデルを複数個用意しておき，各モデルとターゲットタスクの類似度指標を 3 章で述べた方法に基づき求める．そして，各モデルをターゲットタスクへと fine-tuning した際の認識精度を比較し，類似度指標との関連性を考察する．

本評価ではニューラルネットワークの構成として AlexNet と resnet-18 を用いる．転用元となるモデルは ImageNet2012[6] および places365[15] で訓練済みのもの，およびそれらを caltech-256[8] のデータセットを用いて fine-tuning したものを用いる．caltech-256 による fine-tuning の際には，損失関数に交差エントロピー誤差，optimizer には確率的勾配降下法（SGD）を用いて 30 エポックの学習を行った．学習開始時の学習率は 0.001 で，7 エポック毎に 0.1 倍する．各エポック毎にモデルのパラメータを記録し，テスト用データセットにおける認識率が最も高かったパラメータのモデルのを転用元モデルとして採用する．

ターゲットタスクには caltech-101[7] のクラス分類問題を用いる．なお，ターゲットタスクの fine-tuning を行う際にも，上述の学習手法と同条件で行うこととした．

また，3.2 節の EMD に基づく方法により各モデルとターゲットタスクの距離を算出する際に，クラス名の間の距離の算出には facebook research の訓練済みの word2vec モデルを用いた [16]．また，各クラス名が複数の単語で構成されていた場合には，word2vec で計算した各単語のベクトル和をクラス名のベクトルとした．さらに，word2vec の key に存在しない単語がクラス名に含まれていた場合は，類義語あるいは上位の概念にあたる単語に置き換えることで対応した．深層学習フレームワークは PyTorch を用いた。

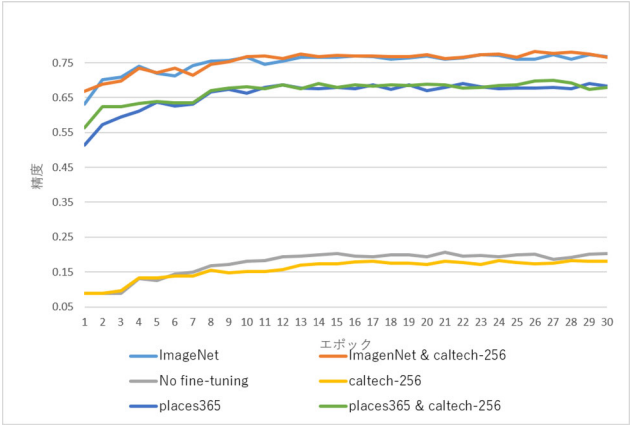

図 2 各モデルの学習の様子: AlexNet

4.2 評価結果

評価に用いた転用元モデルとターゲットタスクである caltech-101 との類似度指標，および fine-tuning を行って 30 エポックの学習した中で最も高かったテスト認識精度と、それを記録したエポックを表 1 に示す．表中，太字で示されているものは，各ネットワーク構成において類似度指標毎に最も良いスコアを達成しているものである．表中 “No training” は，学習を行っていない初期値をそのままパラメータとした場合である．なお，“caltech-256” は caltech-256 のデータセットを使い，4.1 節で述べたのと同じ条件で 30 エポックの学習をしたモデルである．これは，ImageNet や places365 で訓練済みのモデルに比べ，学習回数が非常に少ない場合の転用元モデルの例として評価した．

各モデルをターゲットタスクである caltech-101 のデータセットで fine-tuning した際の学習の様子を，ネットワーク構成毎に図 2, および図 4 に示す．図 3 と図 5 は，図 2 と図 4 のグラフの一部をそれぞれ拡大したものである．

表 1 より，3.1 節で提案した Softmax 出力に基づく指標，3.2 節で提案した EMD に基づく指標ともに，総じて類似度スコアが良いモデルで fine-tuning を行うことで，高い認識精度がより早い学習段階で得られることがわかる．これは，本稿の目的である fine-tuning が成功しやすいモデルを類似度指標により選択することが成功していると言える．

一方で，表 1 より，Softmax 出力に基づく類似度指標のスコアが良いモデルでも，fine-tuning がうまく行えていないもの，あるいはその逆の場合も観測されている．本指標は転用元モデルのクラス数が多いほど類似度との関連性が低くなりやすいことが予想されていたが，実際事前に学習を全く行っていない “No training” モデル（ソースタスク

© 2017 Information Processing Society of Japan

表 1　評価に使用したモデルと類似度指標

	model	softmax 指標	EMD 指標	max accuracy	epoch
AlexNet	No training	7.095	44280	0.2071	21
	ImageNet	2154	40805	0.7724	27
	ImageNet & caltech-256	**4344**	**32285**	0.7831	26
	caltech-256	27.519	52505	0.183	24
	places365	1912	53355	0.6899	22
	places365 & caltech-256	4039	34520	0.7001	27
ResNet18	No training	31.98	44258	0.4065	24
	ImageNet	2012	39016	0.8496	26
	ImageNet & caltech-256	**3953**	**30429**	0.8532	20
	caltech-256	873	49056	0.4449	28
	places365	2396	53217	0.7162	24
	places365 & caltech-256	2759	36465	0.7488	23

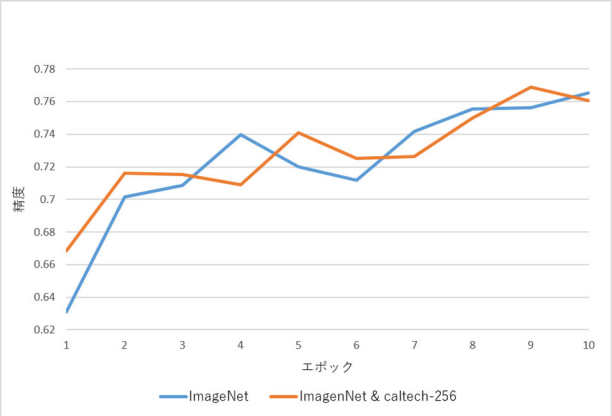

図 3　各モデルの学習の様子: AlexNet（一部拡大）

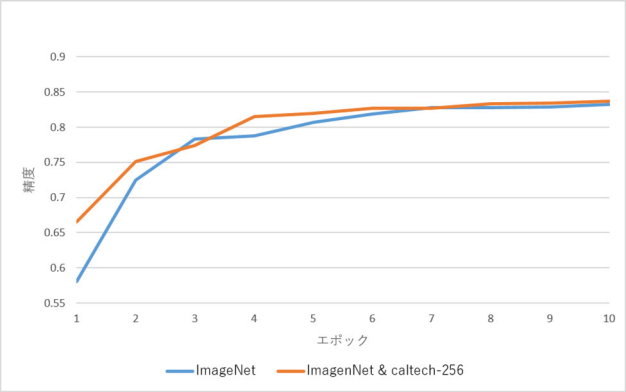

図 5　各モデルの学習の様子: ResNet18（一部拡大）

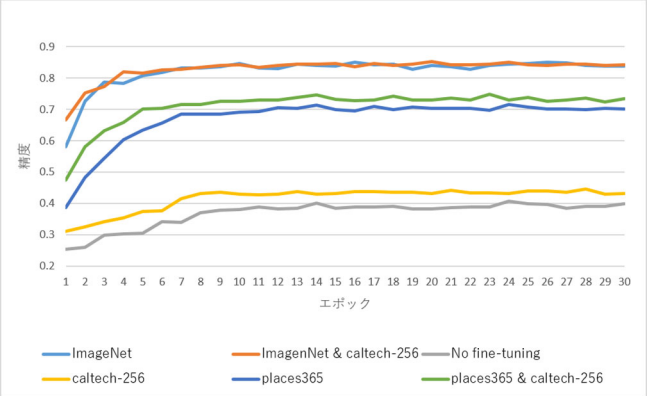

図 4　各モデルの学習の様子: ResNet18

のクラス数 1000）と，学習回数が少ない “caltech-256” のモデル（ソースタスクのクラス数 256）は，fine-tuning 後には同程度の精度となっているが，類似度指標間では大きな差がある．ImageNet におけるクラスは，非常に具体的な動物の種などが多く含まれることも，実際の精度に対して類似度指標が低く見積もられていることの一因であると考えられる．

EMD に基づく類似度指標については，ほぼスコアが良いモデルほど fine-tuning が成功し，高い精度が早い段階で得られる傾向にある．しかし，“ImageNet” と “places365 & caltech-256” の転用元モデルを考えると，類似度指標としては前者が低いスコアである一方，ターゲットタスク向けの fine-tuning 後にはより高い精度が得られている．“ImageNet” は “places365 & caltech-256” よりも汎化性能が高い認識ができる傾向にある．EMD に基づく指標の場合には転用元モデルの汎化性能の高さよりも，転用元モデルを作成した際のソースタスクとターゲットタスクの類似度が重視されてしまうが，fine-tuning を行う上では汎化性能も重要な指標である可能性があると考えられる．

5. おわりに

本稿では，CNN による画像認識を行う上で，環境に特化したモデルを構築するための転移学習を効率的に行うための指標についての検討を行った．転用元モデルとターゲットタスクの類似度を定量的に評価する手法を提案し，類似度指標と転移学習後の認識精度について評価を行った．実際に複数の大規模なデータセットで訓練済みのモデルと，caltech101 のクラス分類問題間の類似度を示し，各モデルを転用元として fine-tuning を行った際の精度を比較したところ，類似度が高いものほど fine-tuning 後に早い学習段階で高い精度が得られることを確認した．これは，提案手法が転移学習のためのモデル選択指標として有効であることを意味している．

© 2017 Information Processing Society of Japan

今後の課題としては，より様々な転用元モデルとターゲットタスクの組み合わせについて評価をすることがあげられる．また，モデルのパラメータや特徴マップなどの比較を行うなど，別の類似度指標について検討することも今後の課題である．

謝辞 本研究の一部は JST CREST（研究課題名「リアルタイム性と全データ性を両立するエッジ学習基盤」）の支援を受けたものである．

参考文献

[1] 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).
[2] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: *Imagenet classification with deep convolutional neural networks*, In Advances in neural information processing systems (pp. 1097-1105) (2012).
[3] LeCun, Y.: *The MNIST database of handwritten digits*, http://yann. lecun. com/exdb/mnist/.
[4] Canziani, Alfredo & Paszke, Adam & Culurciello, Eugenio. (2016). An Analysis of Deep Neural Network Models for Practical Applications. .
[5] Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., and Khudanpur, S.: *Recurrent neural network based language model*, In Interspeech (Vol. 2, p. 3) (2010).
[6] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.
[7] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004, Workshop on Generative-Model Based Vision. 2004
[8] Griffin, G. Holub, AD. Perona, P. The Caltech 256. Caltech Technical Report.
[9] Lu, Ying & Chen, Liming & Saidi, Alexandre. (2017). Optimal Transport for Deep Joint Transfer Learning. .
[10] Zeiler M.D., Fergus R. (2014) Visualizing and Understanding Convolutional Networks. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham
[11] D. Bau*, B. Zhou*, A. Khosla, A. Oliva, and A. Torralba. "Network Dissection: Quantifying Interpretability of Deep Visual Representations." Computer Vision and Pattern Recognition (CVPR), 2017. Oral.
[12] 神嶌敏弘. (2010). 転移学習. 人工知能学会誌, 25(4), 572-580.
[13] Caruana R. (1998) Multitask Learning. In: Thrun S., Pratt L. (eds) Learning to Learn. Springer, Boston, MA
[14] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, Tomaso Poggio. Learning with a Wasserstein Loss. In Advances in Neural Information Processing Systems (NIPS) 28 (2015).
[15] Places: A 10 million Image Database for Scene Recognition B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017
[16] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

© 2017 Information Processing Society of Japan