Vol.2018-ARC-231 No.11
2018/6/15

効率的な転移学習のための学習モデル構築およびモデル選択手法

上野 洋典[1,a)]　近藤 正章[1]

概要：畳み込みニューラルネットワークによる画像認識において，訓練データが十分に得られない場合に大規模なデータセットで学習済みのモデルを対象とするデータセットの学習へと転用する転移学習が有効であることが知られている．この際に，転用元の学習モデルと転用先のデータセットと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．本稿では，効率的な転移学習を行うために，あるモデルが獲得した特徴量に着目し，各モデルと対象データセットとの親和性を定量化するための指標をいくつか検討する．また，種々の転用元となる学習モデルをあらかじめ構築しておくためのデータセット構築手法についても検討する．評価の結果，提案した指標のいくつかにもとづき転用元モデルを選択することで，効率的に転移学習が行えることが確認できた．

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，自然言語処理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あげている．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重きが置かれ，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・使用するためには以下のような課題がある．

まず第 1 に，学習にかかる計算コストが高いことが問題となる．CNN が画像認識において成功を収めた理由の一つは，多層化により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．また，IoT デバイス数が飛躍的に増加しセンサによるデータの収集が容易となった今日において，IoT デバイスや組み込み機器上で学習を行うこと要求も増加すると予想される．そのため，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 2 に学習に使える訓練データが制約されるという問題がある．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．学習データへのラベル付けは基本的に人手により行われるため，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，ラベル付けのコストはさらに大きくなる．また，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．このように，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困難である場合も多い．

これらの理由から，少ない訓練データかつ低い計算コストで CNN の学習をすることが重要と考えられているが，このような場合には転移学習が有効であることが知られている [7]．転移学習とは，ある領域において事前に学習させたモデルを別の領域に転用し適応させる技法である [11]．CNN による画像認識では，転移学習の一種である fine-tuning[8] が良く用いらる．fine-tuning は，ゼロからモデルの学習を行うのではなく，ImageNet[5] に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方法である．fine-tuning により，認識したい物体について少数の

[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

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．

この転移学習を応用することにより，例えばIoTデバイスや組み込み機器などのエッジと中央サーバを協調させ，中央サーバ上では多数の学習データを利用して得られた様々な学習済みモデル群を提供しつつ，エッジ上で低コストかつ高精度を得られるCNNの学習を行うことが可能になると考えられる．この際には，モデル群の中からエッジで収集したデータと親和性の高いモデルを選択し，そのモデルを元に転移学習を行うことが効率的な学習には重要である．

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が，一般的に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 [11][12]．しかしながら，ソースタスクとターゲットタスクの関連性，あるいは学習済みモデルとターゲットタスクの親和性を定量的に評価する指標は確立されていない．上述のようなシステムで効率的な転移学習を行うには，このような指標の構築が必要不可欠である．

本稿では画像分類問題において，fine-tuningを効率的に行うために，複数の学習済みモデルの中から転用元モデルを選択するための指標を提案し，その初期評価を行う．転用元候補となる各モデルのターゲットタスクに対する親和性を定量的に評価することで，指標の有効性を評価する．さらに，上述のエッジデバイス側での転移学習時に，応用環境に応じて変わり得る様々なターゲットタスクに対応するためには，異なる特徴を持つ種々の学習モデルを予め準備しておくことが望ましい．本稿では種々の学習モデルをあらかじめ構築しておくためのデータセット構築手法についても検討する．

2. 関連研究

本章では，本稿の関連研究として，CNNの獲得した特徴量を解釈することを目的とした研究，および転移学習を効率的に行うことを目的とした研究について述べる．

2.1 CNNの特徴量の解釈を目的とした研究

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1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ことがあげられる．一方，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 [9].

Bau, Zhouらによる Network Dissection[10] 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の挙動を理解しようとする研究である．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，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．具体的には，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．

2.2 転移学習を効率的に行うことを目的とした研究

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を元に転移学習を効率的に行う手法が提案されている．

Luら [3] の研究では，転用元のモデルとターゲットタスクに対して適応させたいモデルのSoftmax出力同士の距離を定義し，その距離を小さくする方向に学習を進めることで効率的に転移学習を行う手法を提案している．この研究では，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．

$$\begin{aligned}
\text{minimize} \quad & W = \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} && (1)\\
\text{subject to} \quad & f_{ij} \geq 0 (1 \leq i \leq m, 1 \leq j \leq n) && (2)\\
& \textstyle\sum_{j=1}^{n} f_{ij} \leq w_{p_i} (1 \leq i \leq m) && (3)\\
& \textstyle\sum_{j=1}^{n} f_{ij} \leq w_{q_j} (1 \leq j \leq n) && (4)\\
& \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}) && (5)\\
& && (6)
\end{aligned}$$

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

ここで d_{ij} は分布 P,Q の各要素 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．

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．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が行えたと報告されている．

また，筆者らは，ターゲットタスクを入力した際の学習済みモデルのSoftmax出力に基づいて，fine-tuningにおいて転用元のモデルを選択するための指標を以前に提案して

いる [1]．そのうちの 1 つは，上述の EMD を使い Softmax 出力とラベルを表す one-hot ベクトルの距離を算出し，学習済みモデルとターゲットタスクの親和性を測る方法である．EMD の算出に必要な 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を用いている．すなわち行列 d はソースタスクに含まれるクラス名のベクトル表現とターゲットタスクに含まれるクラス名のベクトル表現の各組み合わせのユークリッド距離を表している．当該指標に基づいて転用元モデルを選択することで，ある程度効率的に転移学習をおこなえることが実験により示されている．しかしながら，指標の算出にはターゲットタスクとソースタスクの組み合わせに固有の情報である d を予め計算しておく必要があるという欠点があった．

3. 特徴マップに基づくモデル選択指標

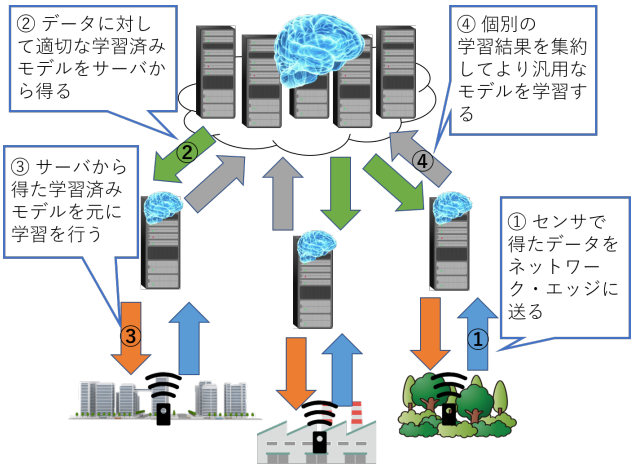

図 1　提案するシステム形態の概要

本稿で想定する応用システム形態の一つを図 1 に示す．IoT デバイスや組み込み機器などのエッジと中央サーバを協調させ，中央サーバ上では多数の学習データを利用して得られた様々な学習済みモデル群を提供しつつ，エッジ上で fine-tuning を行い，低コストに高精度な環境に合わせた CNN 学習モデルを得ることが目的である．サーバ上にあるモデル群の中から，エッジで収集したデータと親和性の高いモデルを選択し，そのモデルを元に転移学習を行うことで効率的な学習を行う．

fine-tuning により物体認識の精度が向上する理由の 1 つとして，以下が考えられている．CNN では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よ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抽出していると言われている [9], [16]．そのため，大規模なデータセットを使って学習したモデルは，入力に近い層ではあらゆる画像認識に有効な普遍的な特徴を学習していると考えられる．認識したい物体の訓練データを用いて再学習を行い，出力に近い層のパラメータを更新することで，そのデータセットに特化した具体的な特徴抽出器を学習し，すでにある汎用的な特徴抽出器と合わせて，対象の物体を高精度に認識できるようになると考えられる．

この際に，ソースタスクとターゲットタスクの関連性が高ければ，出力に近い層での fine-tuning による再学習が効率良く行えると考えられる．そこで本章では，主に出力に近い層の特徴マップ出力に着目し，画像分類問題において，fine-tuning を効率的に行うための，転用元モデル選択指標を提案する．

問題設定として，様々なデータセットのクラス分類用に訓練されたモデルを異なる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親和性を定義する手法について述べる．本稿では転用元の候補となるモデルにターゲットタスクのデータセットの画像を入力した際の特徴マップに着目する．各モデルの構造は AlexNet とし，着目する特徴マップ出力は最も出力に近い畳み込み層である conv5 の出力とする (図 2 内赤丸部分参照)．

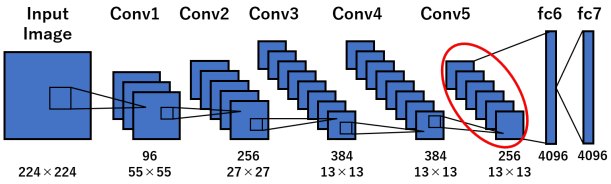

図 2　AlexNet のモデル構造概要図

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_n および X とする．データセットには x_1 から x_N の N 枚の画像が含まれている．また，x_n の属するクラスを c_n とする．画像データ x_n をモデルに入力した際の conv5 層における特徴マップ出力を $T_n \in \mathbb{R}^{256\times13\times13}$ とする．

ターゲットタスクの画像を入力した際のモデルの特徴マップ出力が全結合層で分類しやすいほど，あるいは特徴マップの持つ情報量が多いほど，そのモデルはターゲットタスクに対して親和性が高いと考えられる．これらの方針に従い，以下 5 つのモデル選択指標を提案する．

3.1 特徴マップの成分ごとの分散に基づく方法

特徴マップ出力のうち値が大きくなる箇所は特徴量ごとに局所性があると考えられている．例えば特徴マップ出力のある部分は犬の画像を入力した際にのみ値が大きくなる，ということがこれまでの研究により明らかにされている [16]．よってターゲットタスクのそれぞれの画像を入力した際の特徴マップ出力が大きく異なっていると，全結合層による分類が容易になり，そのモデルはより良いモデルであると考えることができる．そこで，以下の選択指標を提案する．

(1) $T_1, T_2, ..., T_N$ の成分ごとの分散を求める
(2) 各成分ごとに求めた分散の合計をモデル選択指標 S_1 とし，指標 S_1 が大きいほど，モデルとターゲットタスクの親和性が高いと判断する

なお，与えられた複数の特徴マップ出力 $T_1, T_2, ..., T_N$ における (i, j, k) 成分ごとの分散 $\sigma_{i,j,k}$ は以下により求める．

$$\sigma_{i,j,k} = \frac{1}{N}\sum_{n=1}^{N}(T_n(i,j,k) - \mu_{i,j,k})^2 \quad (8)$$

ただし，$\mu_{i,j,k} = \frac{1}{N}\sum_{n=1}^{N} T_n(i,j,k)$ で，$T_n(i,j,k)$ は T_n の (i, j, k) 成分を表すとする．3.2 節，3.5 節でも同様の計算を用いる．

3.2 クラスの平均特徴マップ出力の成分ごとの分散に基づく方法

異なるクラスの画像を入力した際の特徴マップ出力が大きく異なっている場合に，全結合層による分類が容易になり，そのモデルはより良いモデルであると考えることができる．選択指標は以下となる．

(1) 同一クラスの画像に対する T_i を平均し，各クラスの平均特徴マップ出力 $T'_{class-j}$ を求める
(2) $T'_{class-1}, T'_{class-2}, ...$ の成分ごとの分散を求める
(3) 各成分ごとに求めた分散の合計をモデル選択指標 S_2 とし，指標 S_2 が大きいほど，モデルとターゲットタスクの親和性が高いと判断する．

3.3 特徴マップ出力のチャネル同士の相関に基づく方法

CNN の畳み込み層はパラメータの異なる複数のフィルタを持ち，入力から様々な情報を抽出する．$256 \times 13 \times 13$ 次元の特徴マップ出力は，256 個の異なるフィルタにより抽出された 13×13 次元の情報が集約された特徴量となる．この 13×13 次元の情報 1 枚を 1 チャネルと呼ぶ．

それぞれのチャネルが異なる情報を保持しているほど特徴マップ出力全体の持っている情報量は多く，それだけターゲットタスクに対して良いモデルと考えられる．先行研究 [14] では各チャネル同士の相関の絶対値が小さいほど，各チャネルの保持している情報は異なるため，良い特徴マップ出力が得られていると判断している．よって，各チャネル同士の相関の絶対値に基づいたモデル選択指標を提案する．

(1) T_i について，任意の 2 つのチャネル同士の相関の絶対値を合計する
(2) 各 T_i について求めた値の合計を選択指標 S_3 とし，この S_3 の値が小さいほど，モデルとターゲットタスクの親和性が高いと判断する．

3.4 ベクトル化した特徴マップ出力同士のユークリッド距離に基づく方法

AlexNet において，conv5 層の特徴マップ出力は，$256 \times 13 \times 13 = 43264$ 次元のベクトルに変形されて全結合層に入力される．それらのベクトルが大きく異なっていると，全結合層による分類が容易になると考えられる．よって，これらのベクトル同士のユークリッド距離に基づいたモデル選択指標を提案する．

(1) T_i を 43246 次元ベクトル V_i に変形
(2) 任意の i, j について，V_i と V_j のユークリッド距離を計算し，その合計を選択指標 S_4 とし，この S_4 の値が大きいほど，モデルとターゲットタスクの親和性が高いと判断する．

3.5 同一クラス画像入力に対する特徴マップテンソルの成分ごとの分散に基づく方法

同一クラスに属する異なる画像に対する特徴マップ出力が似ているほど，全結合層による分類が上手くいくと考えられる．よって，同一クラスに属する画像に対する特徴マップ出力同士の成分ごとの分散に基づいたモデル選択指標を提案する．

(1) 同一クラスの画像に対する特徴マップ出力同士の成分ごとの分散を求め，その合計を計算する
(2) 各クラスについて（1）で計算した値を合計し，選択指標 S_5 とし，この S_5 の値が小さいほど，モデルとターゲットタスクの親和性が高いと判断する．

4. 様々な学習モデルを構築するためのデータセット構築手法

本章では，様々な特徴を持つ学習済みモデルを用意するための学習データセット構築手法について述べる．3 節で述べたモデル選択指標の有効性を確認するためにも，本データセット構築手法を利用して得られた複数の学習済みモデルを利用する．具体例として ImageNet2012[5] のデータセット利用しつつ，それらを分割したデータセットを複数構築する．それぞれのデータセットで予め学習を行うことで学習済みモデルを構築する．

4.1 単純な分割によるデータセットの構築

ImageNet のデータセットを例えば単純に分割することで，複数のデータセットを構築する．すなわち，比較的ランダムなデータセットに近いと考えられる．ここでは ImageNet の 1000 クラス分のデータセットを 10 分割し，100 クラスのデータセットを 10 個作成する．それぞれ個別に学習を行うことで，学習データ異なるモデルを 10 個得ることができる．

データセットの分割は，クラス番号の 0 番から 99 番のデータを Subset1, 100 番から 199 番のデータを Subset2, …，というように行う．以下，このデータセット群のこと

を “単純サブセット” と呼ぶ.

4.2 スーパークラスを考慮した分割によるデータセットの構築

ImageNet のデータのクラスには，より一般的な概念であるスーパークラスが設定されている．例えば，“tench” クラスのデータと “tiger fish” クラスのデータはどちらもスーパークラス “fish” が設定されている．このようにスーパークラスを考慮してデータを分割する手法である．ここでは，ImageNet2012 のデータの一部を 12 のデータセットに分割する．スーパークラスの名前と各スーパークラスに含まれるクラスの数，およびスーパークラスを構成するクラスの例を表 1 に示す．以下，このデータセット群のことを “スーパークラス指向サブセット” と呼ぶ.

表 1　スーパークラス指向サブセットの概要

	class num	含まれるクラスの例
ball	7	baseball, croquet_ball
bear	4	black_bear, brown_bear
bike	3	mountain_bike, scooter
bird	17	bald_eagle, chickadee
bottle	7	beaker, beer_glass
cat	13	Egyptian, Siamese
dog	123	Afghan_hound, Shih_Tzu
fish	5	goldfish, great_white
fruit	11	Granny_Smith, bell_pepper
sign	2	sign_street_sign sign_traffic_light
turtle	5	box, leatherback
vehicle	14	ambulance, limousine

5. 評価

本章では提案する各モデル選択指標が，効率的な転移学習を行うためのモデル選択において有効であるかどうかについて評価を行う.

5.1 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，特定のターゲットタスクに対して各モデルの指標を 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，算出した指標との関連性を考察する.

5.1.1 転用元モデルの準備

本評価ではニューラルネットワークの構成として AlexNet を用いる．転用元となるモデルは 3.2.1 項で述べた ImageNet2012[5] の単純サブセットで学習したモデル 10 個，および 3.2.2 項で述べた ImageNet2012 のスーパークラス指向サブセットで学習したモデル 12 個の，合計 22 個である．転用元のモデルの学習条件は表 2 にまとめた.

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際にモデルの更新を行う.

表 2　学習条件

Parameter	Description
エポック数	30
損失関数	クロスエントロピー誤差
Optimizer	確率的勾配降下法 (SGD)
学習率	初期値 0.01 で 7 エポック毎に 0.1 倍
パラメータの更新	各エポックでテストデータにおける認識率が向上した際にのみ更新

ターゲットタスクには caltech-101[6] のクラス分類問題を用いる．なお，学習済みモデルをターゲットタスクの fine-tuning を行う際にも，表 2 の学習手法と同条件で行う.

5.1.2 評価指標

算出した指標の有用性評価には，モデル毎の各指標の値と fine-tuning 後のターゲットタスクにおける精度の相関を見ることで行う．本稿では，相関の指標として Spearman の順位相関係数を用いる．Spearman の順位相関係数は以下の式で表される.

$$\sigma = 1 - \frac{6\sum D^2}{M^3 - M} \tag{9}$$

ただし，M は要素数，D は対応する順位の差を表している．例えば 3 つの転用元モデル A, B, C について，ある指標を算出し，その指標と fine-tuning 後の精度について Spearman の順位相関係数を計算することを考える．ある指標では A, B, C の順番で親和性が高いと判断でき，実際の精度は B, C, A の順番で高かったとする．この場合要素数 M は 3 で，対応する順位の差 D はモデル A, B, C それぞれについて $1, 1, 2$ となる．よってこの場合の指標と fine-tuning 後の精度の間の Spearman の順位相関係数は，-0.5 となる.

定義から $-1 \leq \sigma \leq 1$ であり，一般に $|\sigma| \geq 0.4$ で 2 つの値の間に相関が認められ，$|\sigma| \geq 0.7$ で強い相関が認められるとされている.

5.2 評価結果

評価に用いた転用元モデル 22 個とターゲットタスクである caltech-101 との各モデル選択指標，および各モデルを転用元として fine-tuning を行った際のテスト認識精度を，表 3, 表 4 に示す．それぞれ，単純サブセットで学習したモデル 10 個，スーパークラス指向サブセットで学習したモデル 12 個の結果である．また，Spearman の順位相関係数を算出するのに必要な各指標と fine-tuning 後の精度の順位を表 5 に示す．これは単純サブセットで学習したモデル 10 個およびスーパークラス指向サブセットで学習したモデル 12 個の合計 22 個のモデルでの結果である.

各モデル選択指標と fine-tuning 後の精度の間の Spearman の順位相関係数は表 6 のようになる.

ここで，要素数 M は 22 である．S_3 以外の指標については相関係数の絶対値が 0.7 を超えており，順位に関して

表 3 各モデル選択指標の値と fine-tuning 後の精度（単純サブセットで学習したモデル）

	S_1	S_2	S_3	S_4	S_5	prec
subset1	1.16E+07	1.63E+05	1.83E+07	4.98E+08	1.02E+07	0.445783
subset2	5.05E+06	7.94E+04	1.77E+07	3.31E+08	4.41E+06	0.445337
subset3	1.00E+06	1.97E+04	9.78E+06	1.97E+08	1.61E+06	0.398929
subset4	1.02E+07	1.12E+05	9.24E+06	4.71E+08	8.80E+06	0.46988
subset5	1.86E+07	1.87E+05	1.10E+07	6.40E+08	1.58E+07	0.508255
subset6	1.99E+07	1.98E+05	1.00E+07	6.64E+08	1.71E+07	0.49353
subset7	1.64E+07	1.56E+05	9.66E+06	6.00E+08	1.41E+07	0.507363
subset8	2.18E+07	2.30E+05	1.56E+07	6.93E+08	1.86E+07	0.496653
subset9	1.82E+07	1.83E+05	1.20E+07	6.35E+08	1.55E+07	0.516287
subset10	1.55E+07	1.73E+05	9.44E+06	5.84E+08	1.32E+07	0.480589

表 4 各モデル選択指標の値と fine-tuning 後の精度（スーパークラス指向サブセットで学習したモデル）

	class num	S_1	S_2	S_3	S_4	S_5	prec
ball	7	1.04E+05	2.91E+03	1.94E+07	4.64E+07	8.83E+04	0.144578
bear	4	1.86E+05	3.04E+03	2.38E+07	6.14E+07	1.60E+05	0.166444
bike	3	6.92E+04	3.39E+03	2.59E+07	3.58E+07	5.73E+04	0.112004
bird	17	5.91E+07	8.80E+02	1.22E+07	1.69E+05	1.53E+05	0.257921
bottle	7	1.29E+05	6.58E+03	2.60E+07	5.21E+07	1.12E+05	0.186078
cat	13	1.84E+05	2.78E+03	1.54E+07	6.30E+07	1.65E+05	0.251227
dog	123	2.22E+06	9.98E+03	2.35E+06	2.18E+08	2.04E+06	0.369478
fish	5	4.81E+05	1.90E+03	2.25E+07	8.76E+07	4.08E+05	0.181615
fruit	11	1.70E+05	1.40E+03	7.14E+06	5.99E+07	1.48E+05	0.230701
sign	2	3.84E+04	7.27E+02	3.01E+07	2.60E+07	3.38E+04	0.0888
turtle	5	1.11E+05	8.54E+02	2.81E+07	4.51E+07	9.34E+04	0.191879
vehicle	14	1.56E+05	1.72E+03	8.72E+06	5.76E+07	1.36E+05	0.238733

表 5 各モデル選択指標と fine-tuning 後の精度の順位（全モデル）

	S_1	S_2	S_3	S_4	S_5	prec
subset1	8	6	15	7	16	8
subset2	10	9	14	9	14	9
subset3	12	10	7	11	12	10
subset4	9	8	4	8	15	7
subset5	4	3	9	3	20	2
subset6	3	2	8	2	21	5
subset7	6	7	6	5	18	3
subset8	2	1	13	1	22	4
subset9	5	4	10	4	19	1
subset10	7	5	5	6	17	6
ball	20	15	16	18	3	20
bear	14	14	18	14	9	19
bike	21	13	19	20	2	21
bird	1	20	11	22	8	12
bottle	18	12	20	17	5	17
cat	15	16	12	13	10	13
dog	11	11	1	10	13	11
fish	13	17	17	12	11	18
fruit	16	19	2	15	7	15
sign	22	22	22	21	1	22
turtle	19	21	21	19	4	16
vehicle	17	18	3	16	6	14

表 6 各モデル選択指標と fine-tuning 後の精度の順位順位相関係数

	S_1	S_2	S_3	S_4	S_5
Spearman の順位相関係数	0.878	0.863	0.565	0.819	-0.929

高い相関が認められる．しかし，S_5 に関しては相関の符号が負になっており，選択指標の算出方針とは異なる結果となっている．

図 3，図 4 に，順位相関係数の絶対値が最も低かった S3 と最も高かった S5 についての選択指標と精度の値についての散布図をそれぞれ示す．横軸は選択指標，縦軸は認識精度（対数）である．図 3 に示す S3 指標については，指標と精度の間に関係性はあまり見られない．そのため，表 6 で Spearman の順位相関係数を踏まえても，S3 指標は良い指標ではないことがわかる．一方で，S5 指標については正の相関が見られ，片対数グラフであることに注意すると，本指標と fine-tuning 後の精度には指数的な相関がある．しがたって，S5 指標の値から fine-tuning 後の精度の値を予測可能であることが示唆されており，本指標はモデル選択時の有効な指標になり得ることがわかる．

表 3 より，単純サブセットで学習した各モデルの fine-tuning 後の認識精度を比べると，最高のものと最低のもので 10 %以上の差がある．単純サブセットのそれぞれに含まれる学習データの数は同程度であるので，学習済み

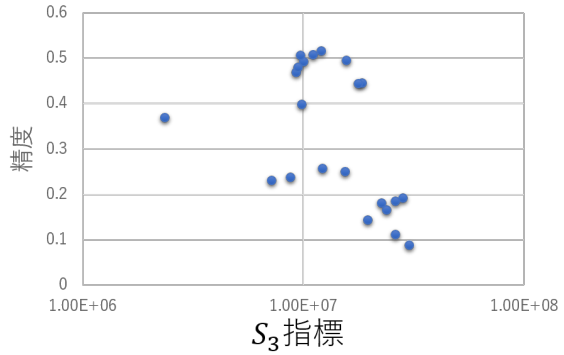

図 3 各モデルの選択指標 S_3 と fine-tuning 後の精度の散布図

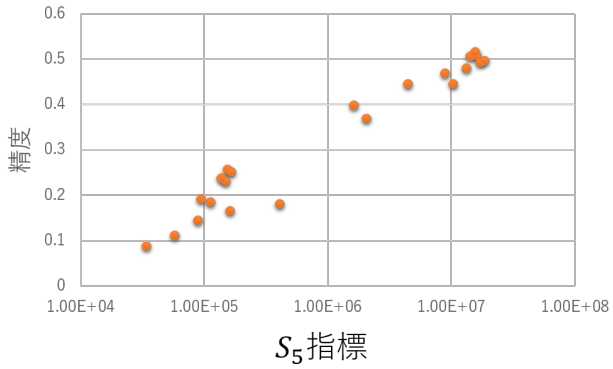

図 4 各モデルの選択指標 S_5 と fine-tuning 後の精度の散布図

モデルの構築方法を工夫することで fine-tuning を効率良く行うことが可能であると考えられる．スーパークラス指向サブセットで学習したモデルのうち，sign や bike 等のスーパークラスに含まれるクラス数の少ないモデルは fine-tuning 後の精度が低くなっている．学習データが少なく，当該モデルが画像認識に必要な汎用的な特徴量を獲得できていないことが示唆される．スーパークラス指向サブセットの “dog” で学習したモデルは 123 クラス分の画像で学習を行っているため，それぞれ 100 クラスから成る単純サブセットで学習したモデルに比べて多くの学習データを使って学習していると考えられる．しかし，“dog” モデルの fine-tuning 後の精度は，単純サブセットで学習したモデルのどれよりも低くなっている．これにより，特定のスーパークラスに属する学習データで学習したモデルよりも，幅広い概念の学習データを使って学習したモデルのほうが fine-tuning の転用元モデルとして適切であることが示唆される．本稿でターゲットタスクとして用いた caltech-101 データセットには幅広い概念の画像が含まれるためこのような結果になっているとも考えられるので，種々のターゲットタスクで検証する必要がある．

6. おわりに

本稿では，CNN による画像認識を行う上で，転用元の候補となるモデルが複数ある際に転移学習を効率的に行うためのモデル選択指標を提案した．転用元モデルとターゲットタスクの親和性を定量的に評価する指標により，適切にモデルを選択可能であることを Spearman の順位相関係数を用いて示した．また，いくつかの指標については，選択指標の値と fine-tuning 後の精度の値が指数的に相関を持つことが示唆された．さらに，学習済みのモデルを複数構築するためのデータセット構築手法についても検討を行った．

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．さらに，今回は CNN のアーキテクチャは AlexNet に限定したが，他の典型的なアーキテクチャに関しても同様の方針で有効なモデル選択指標を算出できるか検証を行う必要もある．

謝辞 本研究の一部は，JST CREST 課題番号 JPMJCR1785（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] 上野 洋典 , 東 耕平 , 近藤 正章 : 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, 研究報告システム・アーキテクチャ（ARC）,2017-ARC-228(3),1-6 (2017-10-31) , 2188-8574
[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] Lu, Ying & Chen, Liming & Saidi, Alexandre. (2017). Optimal Transport for Deep Joint Transfer Learning. .
[4] Canziani, Alfredo & Paszke, Adam & Culurciello, Eugenio. (2016). An Analysis of Deep Neural Network Models for Practical Applications. .
[5] 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.
[6] 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
[7] R. Girshick, J. Donahue, T. Darrell, U. C. Berkeley, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proc. IEEE CVPR, 2014.
[8] P. Agrawal, R. Girshick, and J. Malik. Analyzing the Performance of Multilayer Neural Networks for Object Recognition. In Proc. ECCV, 2014.
[9] 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
[10] 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.
[11] 神嶌敏弘. (2010). 転移学習. 人工知能学会誌, 25(4), 572-

580.
[12] Caruana R. (1998) Multitask Learning. In: Thrun S., Pratt L. (eds) Learning to Learn. Springer, Boston, MA
[13] 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).
[14] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer. Shift: A zero flop, zero parameter alternative to spatial convolutions. arXiv preprint arXiv:1711.08141, 2017
[15] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.
[16] Morcos, A.S., Barrett, D.G., Rabinowitz, N.C., Botvinick, M.: On the importance of single directions for generalization. Int. Conf. on Learning Representations (ICLR), 2018.