4M1-3

幾何学的不変性獲得のための多段CNNの提案

Proposal of multi-stage CNN for geometric invariance acquisition

高橋良 *1 松原崇 *1 上原邦昭 *1
Takahashi Ryo Matsubara Takashi Uehara Kuniaki

*1神戸大学 大学院システム情報学研究科計算科学専攻
Graduate School of System Infomatics, Kobe University

Convolutional neural networks (CNNs) have demonstrated remarkable results in image classification tasks. The deeper CNNs have achieved higher performances thanks to their numerous parameters and resulting high expression ability as well as robustness to parallel shift of objects in images. However, the CNNs have a limitation to their robustness to other geometric transformations such as scaling and rotation. Of course, this problem limits performance improvement of the deep CNNs, but there is no established solution. This study focuses on scale transformation and proposes a novel network architecture called *weight-shared multi-stage network* (WSMS-Net), consisting of multiple stages of CNNs. The WSMS-Net is easily combined with existing deep CNNs, such as DenseNet, and enables them to acquire a robustness to scaling of objects. The experimental results demonstrate that existing deep CNNs combined with the WSMS-Net achieve higher accuracy.

1. はじめに

画像識別の分野において，畳み込みニューラルネットワーク（Convolutional Neural NetWork; CNN）を用いた手法 [LeCun 1989] が大きな成果を挙げている．さらに，CNN による画像識別の精度は，年を経るごとに新たなネットワーク構造の開発とともに向上している．例えば，8 層構造の CNN モデル，AlexNet [Krizhevsky 2012] をはじめ，19 層構造の VGG [Simonyan 2015]，22 層構造の GoogLeNet [Szegedy 2014] などの多層構造のネットワークが精度の更新を可能にしている．これらのネットワークは，層を深くして内部パラメータの数を増やし，ネットワークの表現能力を向上させ，複雑な画像処理にも高い精度を達成している．100 以上の総数を持ち，さらに高い識別精度を達成したのが，2015 年の ILSVRC で最高精度を記録した ResNet [He 2015] である．ResNet は，従来の畳み込み層に加えて，畳み込みを行わないショートカット経路を層と層の間に追加している．この新たな構造によって，勾配情報をショートカット経路を通して伝播することができ，多層構造において識別精度向上の障害となる勾配消失問題を大きく克服している．また，ネットワークの形を工夫して勾配消失を防ぐ研究の一例として，DenseNet[Huang 2016]，PyramidalNet[Han 2016]，WideResnet[Zagoruyko 2016] などがある．

しかし，勾配消失問題の克服に大きな焦点が当てられる中で，CNN の幾何学的不変性の問題については注目を浴びていない．幾何学的不変性とは，画像中のある物体が，他の画像中で拡大，縮小されていても同じものだと認識できる拡大縮小の不変性，回転されていても同じものだと認識できる回転の不変性，位置が変わっていても同じものだと認識できる平行移動の不変性などに分類される．幾何学的不変性の中でも，CNN ではプーリング層により，微小な平行移動に対して頑健である [Le 2010] が，拡大縮小，回転などの変化には強くない．

そこで本研究では，幾何学的不変性の中でも，拡大縮小不変性に焦点を当て，拡大，縮小された物体の特徴を同一であると識別できる新たな CNN モデル，重み共有多段ネットワーク (Weight shared multi-stage network; WSMS-Net) を提案する．WSMS-Net は，畳み込み層を積み重ねて構成される従来の CNN に対して，この CNN を平行で多段（multi-stage）に構築したネットワークである．さらに WSMS-Net は，異なるステージの同一の深さの層では，畳み込みに用いる重みを完全に共有するようにしている．異なるステージには，互いに異なるサイズの同じ入力画像が入力され，各ステージの各サイズの画像から得られた特徴がネットワークの最後でひとつに統合される．異なるステージにおける，特徴抽出と統合処理，さらに畳み込みの重み共有の構造によって，拡大縮小された物体の識別を可能としている．そして，この新たな提案モデルを既存の多層 CNN モデルと組み合わせ，さらなる精度の向上を果たすことを目的としている．

連絡先: 高橋良，神戸大学大学院システム情報学研究科計算科学専攻，takahashi@ai.cs.kobe-u.ac.jp

2. WSMS-Net

2.1 WSMS-Net の構造

WSMS-Net の概略図を図 1 に示す．WSMS-Net に入力となる画像が入ると，まず画像のリサイズが行われ，複数の異なるサイズの画像が生成される．そして，もとのサイズの入力とリサイズされた複数の入力が，それぞれ別のステージへの入力となり，別々のステージによって，それぞれのサイズの入力画像ごとに特徴抽出が行われる．複数ステージの CNN から得られた特徴は，各ステージを抜けた箇所でチャネル方向についての結合が行われる．さらに，結合され，ひとつになった特徴マップを全結合層へ入力する前に，緩衝材の役割を果たす畳み込み層をひとつ追加する．この層を Integration Layer と呼称する．Integration Layer の調整は，複数ステージから得られた特徴マップをうまく統合する際に重要となる．詳細な調整については 2.3 節において詳しく説明する．最後に，Integration Layer から得られた出力が，全結合層への入力となって，画像識別が行われる．

2.2 拡大縮小不変性獲得のコンセプト

WSMS-Net によって拡大縮小不変性が獲得できると考えられる理由を詳しく説明する．WSMS-Net による拡大縮小不変性獲得の様子を図 2 に示す．図 2 のように，2 つのステージで構成されるネットワークにおいて，2 つの画像 A，B を順に入

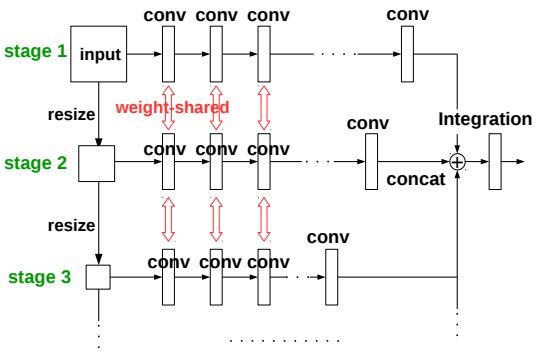

図 1: Weight-shared multi-stage network (WSMS-Net).

力する場合を考える．ステージ 1 には入力画像がそのまま，ステージ 2 には 1/2 に縮小されて入力されるものとする．まず初めに画像 A をネットワークに入力する．このとき，例えば，ステージ 1 で，図に赤く示した領域の特徴が，車であるという識別を行う際に有効な特徴であるとネットワークが学習し，重みが更新されたとする．次に画像 B が同様にネットワークに入力される．このとき，ステージ 1 において，さきほどと同じ領域で特徴を抽出したところ，ネットワークは，画像 A を入力した際に図で赤く示した領域の特徴と全く異なる特徴が抽出されていることに気づく．そして，もしステージが 1 だけであった場合，ネットワークは画像 B のステージ 1 で抽出した特徴も必要であるものであると学習してしまい，画像 A のステージ 1 における特徴との間で，互いに大きく異なる特徴を学習してしまう．これが拡大縮小不変性を獲得できないメカニズムである．そこで，図 2 のように，ネットワークにステージ 2 を追加する．すると，画像 B を入力し，ステージ 2 において画像全体から特徴を抽出したとき，重みをステージ 1 と共有していることによって，画像 A の時に有効とした，赤く示した領域の特徴と一致する．そして，特徴の一致により，ネットワークは初めから画像 B を車であると識別できるため，画像 B からステージ 1 で得た特徴に引っ張られた学習が行われることはなく，特徴抽出の整合性は失われない．従って画像 A, B のような 2 つの画像を学習させたとしても，両者を正しく同じクラスに識別することが可能となる．ここでは例として，一部拡大された画像について説明したが，同様のことは縮小の場合でも言え，よって拡大縮小不変性の獲得がなされる．

2.3 Integration Layer

複数ステージから得られた特徴をうまく統合する際に重要となるのが，Integration Layer である．WSMS-Net では，複数ステージから得られた特徴をチャネル方向に結合するため，ひとつに結合された特徴マップでは通常の CNN よりもチャネル数が増大する．さらに CNN では，畳み込みから得られた特徴を全結合層に入力する際に，最大プーリングを行って 1 次元の特徴ベクトルを生成して入力する．そのため，通常の CNN に対して，チャネル数が増大した特徴マップを用いることで，全結合層におけるクラス分類に影響が出ることが考えられる．よって，結合した特徴マップに対して，緩衝材の働きを持たせる畳み込み層を加え，畳み込みによってチャネル数を減らし，クラス分類への影響を調整する．

ここで，Integration Layer の種類として，3 通りのパターンを考え，識別に最適なパターンを模索するものとする．ひとつめに考えられるのが，チャネル数の増大が識別に影響しない

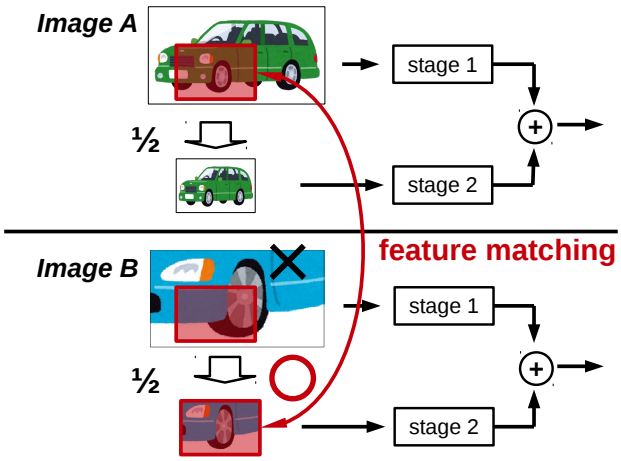

図 2: Conceptual explanation of scale invariance in the WSMS-Net.

として，畳み込み層を追加せず，直接全結合層に入力するというパターンである．本研究では，このパターンの Integration Layer を *no conv* と呼称する．ふたつめに考えられるのが，通常の CNN においてよく用いられる，カーネルサイズ 3 の畳み込みを用いるパターンである．このパターンを 3×3 *conv* と呼称する．最後に考えられるのが，カーネルサイズ 1 の畳み込みを用いるパターンである．このパターンでは，3×3 *conv* の場合と比較して，特徴マップの縦，横方向への畳み込みを考慮せず，純粋にチャネル方向についてのみの畳み込みとなる．最後のカーネルサイズ 1 の Integration Layer を 1×1 *conv* と呼称する．また，畳み込みを行う 3×3 *conv* と 1×1 *conv* のパターンにおいて，出力となる特徴マップのチャネルサイズは一貫して 128 チャネルに設定する．

3. WSMS-DenseNet

実験では，既存の多層 CNN に WSMS-Net を組み合わせ，拡大縮小不変性の獲得と識別精度向上の検証を行う．ここで，既存の多層 CNN として DenseNet[Huang 2016] を用いる．DenseNet は，画像識別精度において，state-of-the-art な結果を出しているモデルのひとつであり，図 3 に示すように，密構造のネットワークを特徴とするモデルである．畳み込み層と畳み込み層の間のショートカット経路を，全ての深さの層に対して繋げて，勾配消失問題を克服し，ResNet[He 2015] の画像識別精度を大きく上回る．また，図 3 は DenseNet の全体のうちの 1 ブロックである，DenseBlock と呼ばれる．DenseNet は，この DenseBlock を 3 つ並べることで，全体のネットワークが構成されている．DenseNet の詳細については文献を参考されたい．DenseNet と WSMS-Net を組み合わせた新たなモデルを，WSMS-DenseNet と呼称し，概略図を図 4 に示す．通常の DenseNet は，前述の通り，3 つのブロックに大きく分割されるネットワーク構造をしており，1 ブロックごとに特徴マップのダウンサイジングが行われる．よって DenseNet を複数ステージに拡張する際は，特徴マップのサイズが統合 (concat) 部分で合致するように，ステージ 2 を 2 ブロック，ステージ 3 を 1 ブロックとしている．各ブロックの下部に，そのブロックにおける特徴マップのサイズを示している．以上の WSMS-DenseNet を用いて実際に精度実験を行い，拡大縮小不変性獲得の有無と識別精度向上の有無を検証する．

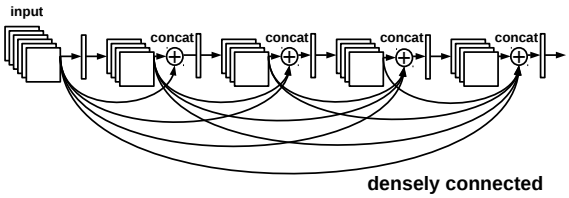

図 3: A dense block of DenseNet.

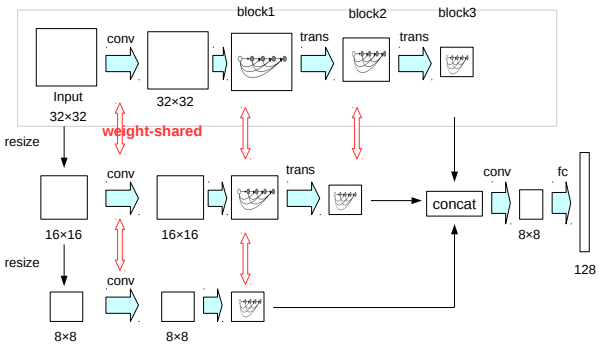

図 4: Weight-shared multi-stage DenseNet (WSMS-DenseNet)

4. 実験と結果

学習データセットとして CIFAR-10，CIFAR-100 画像データセット [Krizhevsky2009] を用いる．CIFAR10，CIFAR-100 は DenseNet が精度評価を行うために用たデータセットである．これらのデータセットは，本研究の提案モデルと通常の DenseNet との比較を行うために用いる．CIFAR-10 は，あらかじめ人の手によって 10 クラスのラベルが付けられた画像データの集合であり，訓練用データ 50,000 枚とテスト用データ 10,000 枚で構成される．訓練用データのうち，5,000 枚はネットワークのハイパパラメータ決定のための検証用データとして用いる．各画像のサイズは全て 32×32 ピクセル，3 チャネルのカラー画像である．CIFAR-100 は 100 クラスのラベルが付与されている点のみ CIFAR-10 と異なっている．

実際の識別精度結果が表 1 である．Network は使用したネットワークモデルを表し，DenseNet は通常の DenseNet，WSMS-DenseNet は多段ネットワークの重みを共有した場合の提案手法モデル，MS-DenseNet は重みを共有しなかった場合のモデルを示す．#params はネットワークの重み数，Error(%) は誤識別率を表す．growth rate(成長率) は，DenseNet の畳み込み層で，特徴をどの程度詳しく取得させるかを調整するハイパパラメータである．この値が大きいほど，ネットワークの重み層数も大きくなる．C10 は CIFAR-10，C100 は CIFAR-100 の結果であることを示す．また，ネットワーク名の右には，使用した Integration Layer の種類を示している．

まず CIFAR-10 の結果に注目する．1×1 *conv* を用いた WSMS-DenseNet は，誤識別率が 3.51%であり，同じ成長率を設定された DenseNet($k = 24$) の場合の誤識別率は 3.74%である．この結果より，通常の DenseNet に対して，提案手法 WSMS-DenseNet の画像識別精度が向上していることが分かる．また，3×3 *conv* では，*no conv* より識別精度が低いものの，通常の DenseNet よりは高い識別精度を出していることが分かる．しかし，1×1 *conv* よりも重みの数が 4M 以上増えており，識別を行う上で，重みの無駄が多い手法になっている．さらに *no conv* の場合，通常の DenseNet の結果よりも悪い精度が出ている．この結果からは，Integration Layer において，畳み込みによる特徴マップ統合の調整が大きな意味を持つことが確認できる．また，提案手法 1×1 *conv* の WSMS-DenseNet で，通常の DenseNet に比べて重みの数が増加している．DenseNet ($k = 26$) の結果は，単純に重みの増加によって識別精度が上がっているのではないことを裏付けるためのものである．この結果では，通常の DenseNet($k = 24$) に対して重みが増えているにも関わらず，識別精度は逆に悪くなっている．そして提案手法 1×1 *conv* WSMS-DenseNet では，同様に重みの数は増えているが，精度は良くなっており，この点から単純な重みの増加によって識別精度が上がってるわけではないことが分かる．最後に，異なるステージでの重み共有の必要性を確認するために MS-DenseNet の結果を見る．MS-DenseNet では，重みの共有をしていないため，重みの総数が大きく増大していることに加えて，識別精度は通常の DenseNet に比べて悪くなっている．よって，ネットワークをマルチステージに構築した効果が全く確認できない．逆に，重みの共有を行い，Integration Layer を MS-DenseNet と揃えた 1×1 *conv* WSMS-DenseNet は，識別精度を更新する結果を出している．この結果の違いから，異なるステージにおける重み共有の必要性が確認できる．次に，CIFAR-100 の結果に注目する．CIFAR-10 の場合と同様に，提案手法 1×1 *conv* WSMS-DenseNet の誤識別率 18.45%は，DenseNet($k = 24$) の誤識別率 19.25%よりも低く，識別精度が向上していることが分かる．さらに，Integration Layer のパターンごとの比較でも，1×1 *conv* が最も識別精度が高くなっている．また，DenseNet ($k = 26$) と MS-DenseNet による，重み総数と重み共有のふたつの観点からの比較でも，CIFAR-10 の場合と同様の結果が得られている．

次に，提案手法による，拡大縮小不変性獲得の有無を検証する結果を示す．図 5 は，CIFAR-10 のテストデータ画像から，ラ ンダムに 20 枚取り出した結果の画像集合である．また，図 6 は，DenseNet($k = 24$) と，DenseNet ($k = 26$) のテスト時に識別不可能であり，かつ提案手法 1×1 *conv* WSMS-DenseNet のテスト時に，新たに識別可能になった画像のうちから，ランダムに 20 枚取り出した結果の画像集合である．まず，図 5 を見ると，テスト画像の中には，物体全体が写った画像が比較的多く，物体が拡大縮小されたタイプの画像は少ない．このことから，CIFAR-10 では，拡大縮小されたタイプの画像がテストデータ全体と比べて，比較的少ないことが分かる．次に図 6 を見ると，ほぼ全ての画像が，物体全体のうちの一部分のみを写しているか，画像サイズに対して小さく写しているような種類の画像であることが確認できる．ここで，提案手法による画像識別精度の向上が，拡大縮小不変性の獲得によるものであれば，図 6 には，拡大縮小されたタイプの画像ばかりが集まっていることが予想される．図 6 の結果は，予想の通りの結果になっていると言え，拡大縮小不変性が獲得されたことを強く示すものとなっている．

以上をまとめると，本研究の目的である，拡大縮小不変性の新規獲得による，画像識別精度の更なる向上が，提案手法 WSMS-Net と，WSMS-DenseNet によって実現できている．

本研究は JSPS 科研費 (16K12487 及び 26280040)，柏森情

表 1: Test Error Rates of WSMS-DenseNet and Original DenseNets on CIFAR-10 and CIFAR-100 Datasets.

		C10		C100	
Integration Type	growth rate k	#params	Error (%)	#params	Error (%)
DenseNet	24	27.2M	3.74	27.2M	19.25
DenseNet	26	31.9M	3.82	31.9M	18.94
MS-DenseNet (1×1 conv)	24	41.3M	4.18	41.3M	18.70
WSMS-DenseNet (no conv)	24	27.4M	4.54	27.8M	20.11
WSMS-DenseNet (3×3 conv)	24	32.7M	3.54	32.7M	19.16
WSMS-DenseNet (1×1 conv)	24	28.0M	**3.51**	28.0M	**18.45**

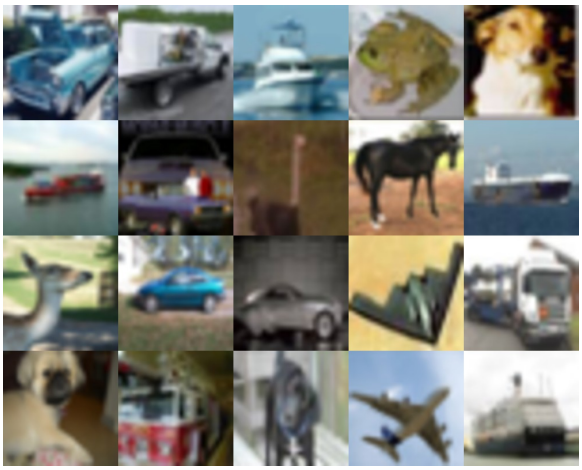

図 5: Examples of CIFAR-10 test images.

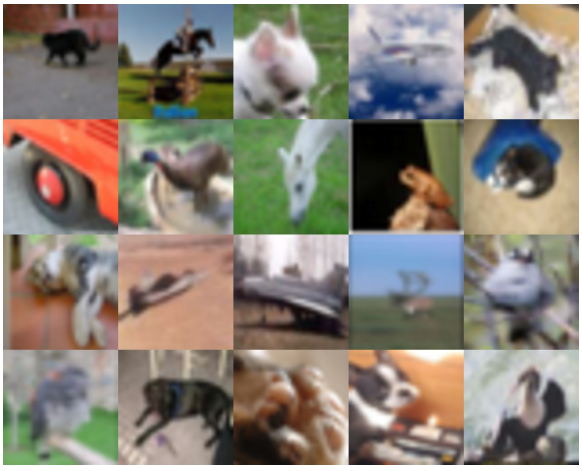

図 6: Examples of CIFAR-10 test images misclassified by the DDenseNet ($k = 24$) and the DenseNet ($k = 26$) but classified by the WSMS-DenseNet ($k = 24$, 1×1 conv) correctly.

報科学振興財団，中島記念国際交流財団，住友電工グループ社会貢献基金の助成を受けて行われた.

参考文献

[LeCun 1989] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “ Backpropagation Applied to Handwritten Zip Code Recognition, ” Neural Computation, vol. 1, no. 4, pp. 541-551, 1989.

[Krizhevsky 2012] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ ImageNet Classification with Deep Convolutional Neural Networks, ”in Advances In Neural Information Processing Systems, F. Pereira, pp. 1097-1105, 2012.

[Simonyan 2015] K. Simonyan and A. Zisserman, “ Very Deep Convolutional Networks for Large-Scale Image Recognition, ” Proc. of International Conference on Learning Representations (ICRL2015), pp. 1-14, 2015.

[Szegedy 2014] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “ Going deeper with convolutions, ”in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[He 2015] S. Wu, S. Zhong, and Y. Liu, “ Deep residual learning for image steganalysis, ”Multimedia Tools and Applications, pp. 1-9, 2017.

[Huang 2016] G. Huang, Z. Liu, and K. Q. Weinberger, “ Densely Connected Convo- lutional Networks, ” arXiv preprint, pp. 1-12, 2016.

[Han 2016] D. Han, J. J. Kim, and J. J. Kim, “ Deep Pyramidal Residual Networks, ”arXiv, pp. 1-9, 2016.

[Zagoruyko 2016] S. Zagoruyko and N. Komodakis, “ Wide Residual Networks, ” arXiv, pp. 1-15, 2016.

[Le 2010] Q. Le, J. Ngiam, Z. Chen, D. H. Chia, and P. Koh, “ Tiled convolutional neural networks. ” Advances in Neural Information Processing Systems 23, pp. 1279-1287, 2010.

[Krizhevsky2009] A. Krizhevsky, “ Learning Multiple Layers of Features from Tiny Im- ages, ” Technical report, University of Toronto, pp. 1-60, 2009.