Deep Residual Learning for Image Recognition

備考

Kaiming He, Xiangyu He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

掲載

Procs. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770--778, 2016.

概要

より深いニューラルネットワークの学習はより困難です。本研究では、従来よりも大幅に深いネットワークの学習を容易にする残差学習フレームワークを提案する。我々は、参照されない関数を学習するのではなく、層の入力を参照して残差関数を学習するように、層を明示的に再構成する。このような残差ネットワークは最適化が容易であり、深さを大幅に増やしても精度を向上させることができることを示す包括的な実証的証拠を提供する。ImageNetデータセットにおいて，VGGネット[40]よりも8倍深い152層までの残差ネットを評価したが，複雑さは依然として低い．これらの残差ネットのアンサンブルは，ImageNetテストセットで3.57%のエラーを達成した．この結果は，ILSVRC 2015の分類タスクで1位を獲得した．また、100層と1000層のCIFAR-10での解析結果も紹介します。

表現の深さは、多くの視覚認識タスクにおいて中心的な重要性を持っています。我々の非常に深い表現により、COCOオブジェクト検出データセットで28%の相対的な改善を得ました。深層残差ネットは、ILSVRCとCOCO2015のコンペに提出したもので、ImageNet検出、ImageNetローカライズ、COCO検出、COCOセグメンテーションのタスクで1位を獲得しました。

1. 緒言

深層畳み込みニューラルネットワーク[22, 21]は，画像分類において一連のブレークスルーをもたらした[21, 49, 39]．深層ネットワークは，低・中・高レベルの特徴量[49]と分類器をエンド・ツー・エン

ドの多層構造で自然に統合し，特徴量の「レベル」は積層数（深さ）によって豊かにすることができます．ネットワークの深さが非常に重要であることを示す最近の証拠[40, 43]があり、挑戦的なImageNetデータセット[35]における主要な結果[40, 43, 12, 16]はすべて、深さが16[40]から30[16]の「非常に深い」[40]モデルを利用しています。他の多くの非自明な視覚認識タスク[7, 11, 6, 32, 27]も、非常に深いモデルから大きな恩恵を受けている。

深さの重要性に迫られて、ある疑問が生まれました。より良いネットワークを学習することは，より多くの層を積み重ねることと同じくらい簡単なのだろうか？この問題を解決するための障害は，グラデーションの消失や爆発の問題 [14, 1, 8] であり，これが最初から収束の妨げになっていた．しかし，この問題は，正規化された初期化[23, 8, 36, 12]や中間正規化層[16]によってほぼ解決され，数十層のネットワークでもバックプロパゲーションを用いた確率的勾配降下法(SGD)の収束を開始できるようになりました[22].

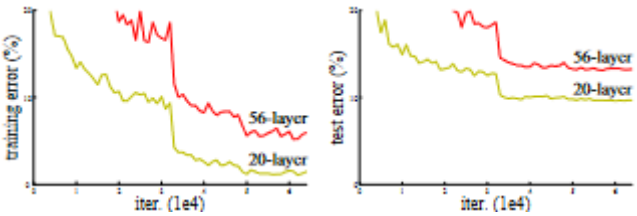

Fig1. 20層および56層の「プレーン」ネットワークを用いたCIFAR-10の学習誤差（左）とテスト誤差（右）。深いネットワークほど学習誤差が大きく、テスト誤差も大きくなっています。ImageNetでの同様の現象を図4に示します。

より深いネットワークが収束し始めると、劣化の問題が露呈します。ネットワークの深さが増すにつれて、精度が飽和し（当然かもしれませんが）、その後急速に劣化していきます。意外なことに，このような劣化はオーバーフィッティングによるものではなく，適切な深さのモデルに層を追加すると学習誤差が大きくなることが[10, 41]で報告されており，我々の実験でも十分に検証されている．図1はその典型的な例です。

学習精度の）低下は、すべてのシステムが同様に最適化しやすいわけではないことを示しています。より浅いアーキテクチャと、その上にさらにレイヤーを追加したより深いモデルを考えてみましょう。より深いモデルには、構築による解が存在します。追加された層はIDマッピングであり、その他の層は学習された浅いモデルからコピーされます。このような解法が存在するということは、深いモデルは浅いモデルよりも学習誤差が大きくならないはずです。しかし、実験によると、現在手元にあるソルバーでは、構築された解と同等以上の解を見つけることができません（または、実現可能な時間内に見つけることができません）。

本論文では、深層残差学習フレームワークを導入することで、この劣化問題に対処します。このフレームワークでは、いくつかの非線形層を重ねることで、目的のマッピングに直接適合させるのではなく、これらの非線形層に残差マッピングを適合させます。形式的には、望ましい基本的なマッピングを $H(x)$ とすると、積層された非線形層に $F(x) := H(x) - x$ の別のマッピングを適合させ、元のマッピングを $F(x) + x$ に再構成する。我々は、参照されていないオリジナルのマッピングを最適化するよりも、残余のマッピングを最適化する方が簡単であるという仮説を立てた。極端に言えば、もしアイデンティティーマッピングが最適であれば、非線形層の積み重ねによってアイデンティティーマッピングをフィットさせるよりも、残差をゼロにする方が簡単であろう。

$F(x)+x$ の定式化は，「ショートカット接続」を持つフィードフォワード・ニューラル・ネットワークで実現できる（図2）．ショートカット接続 [2, 33, 48] は，1つ以上の層をスキップする接続です．今回の例では，ショートカット接続は単に同一性マッピングを行い，その出力は積層された層の出力に追加されます（図2）．アイデンティティ・ショートカット接続は，パラメータの追加や計算量の増加を伴わない．このネットワーク全体は、バックプロパゲーションを用いたSGDによってエンドツーエンドで学習することができ、ソルバーを変更することなく、一般的なライブラリ（Caffe [19] など）を用いて簡単に実装することができます。

ImageNet [35]を用いた包括的な実験を行い、劣化問題を示し、我々の手法を評価します。その結果、以下のことがわかりました。

1. 我々が開発した非常に深い残差ネットは，最適化が容易であるが，対応する「プレーン」ネット（単に層を重ねるだけのネット）は，深さが増すと学習誤差が大きくなる．
2. 我々の深い残差ネットは、深さを大幅に増やすことで容易に精度を向上させることができ、以前のネットワークよりも大幅に良い結果を得ることができる。

同様の現象は，CIFAR-10セット[20]でも示されており，最適化の難しさと我々の手法の効果が，特定のデータセットだけではないことを示唆している．我々は、このデータセットで100以上の層を持つモデルの学習に成功し、1000以上の層を持つモデルを探求している。

ImageNet分類データセット[35]では，非常に深い残差ネットによって優れた結果が得られた．我々の152層の残差ネットは，VGGネット[40]よりも複雑さが低いにもかかわらず，ImageNetでこれまでに発表された中で最も深いネットワークです．我々のアンサンブルは、ImageNetテストセットで3.57%のトップ5エラーを示し、ILSVRC 2015の分類コンペティションで1位を獲得しました。極めて深い表現は、他の認識タスクにおいても優れた汎化性能を発揮し、さらに1位を獲得するに至った。ILSVRC & COCO 2015のコンペティションでは、ImageNet detection、ImageNet localization、COCO detection、COCO segmentationで1位を獲得しました。このように、残差学習原理には汎用性があり、他の視覚問題や非視覚問題にも適用できることが期待されます。

2. 関連研究

残差表現: 画像認識において、VLADは辞書に対する残差ベクトルによって符号化する表現であり、フィッシャーベクトル[30]はVLADの確率的バージョン[18]として定式化できる。どちらも画像検索や分類のための強力な浅い表現である[4, 47]。ベクトル量子化においては，残差ベクトルの符号化[17]が元のベクトルの符号化よりも有効であることが示されている．

ローレベルビジョンやコンピュータグラフィックスでは，偏微分方程式（PDE）を解くために，広く使われているマルチグリッド法[3]が，システムを複数のスケールの部分問題として再構成し，各部分問題が粗いスケールと細かいスケールの間の残差解を担当する．マルチグリッド法に代わる手法として，2つのスケール間の残差ベクトルを表す変数に依存する階層的基底の事前条件付け[44, 45]があります．これらのソルバーは，解の残差を意識しない標準的なソルバーよりもはるかに速く収束することが示されています[3, 44, 45]．これらの手法は，優れた再定式化や前提条件の設定により，最適化を簡略化できることを示唆している．

ショートカット接続: ショートカット接続[2, 33, 48]につながる実践と理論は、長い間研究されてきました。多層パーセプトロン（MLP）を学習する初期のプラクティスは，ネットワークの入力から出力に接続された線形層を追加することである[33, 48]．[43, 24]では，いくつかの中間層を補助的な分類器に直接接続して，バニシング／エクスポージング・グラジェントに対処している．[38, 37, 31, 46]の論文では，層の応答，勾配，伝播した誤差をセンタリングする方法を提案しており，ショートカット接続によって実装されている．[43]では、「インセプション」層は、ショートカットブランチといくつかの深いブランチで構成される。

我々の研究と同時に、「ハイウェイ・ネットワーク」[41, 42]は、ゲート機能[15]を持つショートカット接続を提示している。これらのゲートはデータに依存し，パラメータを持っているが，我々のアイデンティティ・ショートカットはパラメータを持たない．ゲート付きのショートカットが「閉じている」（ゼロに近づいている）場合、ハイウェイ・ネットワークのレイヤーは非永続的な機能を表している。これに対して、我々の定式化は常に残余機能を学習する。我々のアイデンティティ・ショートカットは決して閉じられることはなく、すべての情報は常に通過し、学習すべき残余機能が追加される。また、high-wayネットワークは、極端に深さを増しても（例えば100層以上）、精度の向上は見られない。

3. 深層残差学習

3.1. 残差学習

ここでは、$H(x)$ を、いくつかの層（ネット全体でなくてもよい）でフィットさせるための基礎的なマッピングと考え、xをこれらの層の最初の層への入力とします。複数の非線形層が複雑な関数を漸近的に近似できると仮定すると、残差関数、すなわち $H(x) - x$ を漸近的に近似できると仮定するのと同じことになります（入力と出力が同じ次元であると仮定します）。そのため、積み重ねられた

層が $H(x)$ を近似することを期待するのではなく、これらの層が残差関数 $F(x) := H(x) - x$ を近似することを明示します。どちらの形式でも（仮説通り）目的の関数を漸近的に近似できるはずですが、学習のしやすさは異なるかもしれません。

この再定式化は、劣化問題（図1左）に関する直観に反する現象が動機となっています。冒頭で述べたように、もし追加された層が同一性のマッピングとして構成されるならば、深いモデルは浅いモデルよりも学習誤差が大きくならないはずです。劣化の問題は、ソルバーが複数の非線形層で同一のマッピングを近似することが困難であることを示唆しています。残差学習の再定式化により、同一性マッピングが最適であれば、ソルバーは複数の非線形層の重みをゼロに近づけるだけで、同一性マッピングに近づけることができる。

実際のケースでは、同一性マッピングが最適である可能性は低いが、我々の再定式化は問題の前提条件を整えるのに役立つだろう。最適関数がゼロマッピングよりもアイデンティティマッピングに近い場合、ソルバーにとっては、関数を新たに学習するよりも、アイデンティティマッピングを参照して摂動を見つける方が簡単なはずである。我々は実験により（図7）、学習された残差関数は一般に小さな応答を持つことを示し、同一性マッピングが合理的な前提条件を提供することを示唆している。

3.2. ショートカットによるアイデンティティマッピング

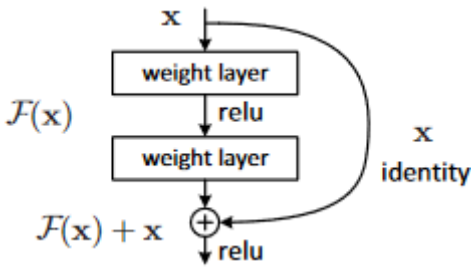

Fig2. 残留学習：ビルディングブロック

また、数枚のスタック層ごとに残差学習を採用しています。図2にビルディング・ブロックを示す。形式的には、本稿では次のように定義されるビルディング・ブロックを考える。

$$y = F(x, \{W_i\}) + x$$

ここで，x と y は，考慮する層の入力と出力のベクトルである。関数 $F(x, W_i)$ は，学習すべき残差マッピングを表しています．2つの層を持つ図2の例では，$F = W_2\sigma(W_1x)$ となり，σ はReLU [29] を表し，ノーテーションを簡略化するためにバイアスは省略されている．演算 $F + x$ は、ショートカット接続と要素ごとの加算によって行われる。加算後の第2の非線形性 ($\sigma(y)$，図2参照) を採用する。

式(1)のショートカット接続は、余分なパラメータや計算の複雑さをもたらさない。このことは、実際には魅力的であるだけでなく、プレーンネットワークと残差ネットワークを比較する上でも重要です。私たちは、パラメータ数、深さ、幅、計算コスト（無視できる要素ごとの追加を除く）が同時に同じであるプレーンネットワークと残差ネットワークを公平に比較することができます。

$$y = F(x, w_i) + W_s x$$

式(1)において、x と F の寸法は等しくなければなりません。そうでない場合（入出力チャンネルを変更した場合など）は、寸法を合わせるためにショートカット接続による線形投影 W_s を行えばよい。

$$y = F(x, W_i) + W_s x$$

式(1)で正方行列 W_s を使うこともできます。しかし、劣化問題を解決するためにはIDマッピングで十分であり、経済的であることを実験で示すので、W_s は次元を合わせるときにのみ使用する。

残差関数 F の形は自由です。本稿の実験では、関数 F が2つまたは3つの層を持つ場合を想定しているが（図5）、それ以上の層も可能である。しかし、単一の層を持つ場合、式（1）は線形の層に似ています: $y = W_1 x + x$，これに対する利点は観察されていません。

また、上記の表記は、簡単のために完全連結層に関するものですが、畳み込み層にも適用できることに注意してください。関数 $F(x, Wi)$ は、複数の畳み込み層を表すことができます。要素ごとの加算は，2つの特徴マップに対して，チャネルごとに行われます．

3.3. Network の構造

我々は、様々なプレーン／リジッドネットをテストし、一貫した現象を観察しました。議論のための事例を提供するために、ImageNetの2つのモデルを以下のように説明します。

プレーンなネットワーク: 我々のプレーンベースライン（図3、中）は、主にVGGネット[40]（図3、左）の哲学にインスパイアされています。畳み込み層は主に3×3のフィルタを持ち、2つの単純な設計ルールに従っています。(i) 同じ出力特徴マップサイズであれば，各層は同じ数のフィルタを持つ，(ii) 特徴マップサイズが半分になれば，フィルタの数は2倍になり，層ごとの時間的複雑さを維持する．ダウンサンプリングは，ストライドが2の畳み込み層で直接行います．ネットワークは，グローバル平均プーリング層と，ソフトマックスを用いた1000通りの完全連結層で終わります．重み付けされた層の総数は、図3（中）の34層です。

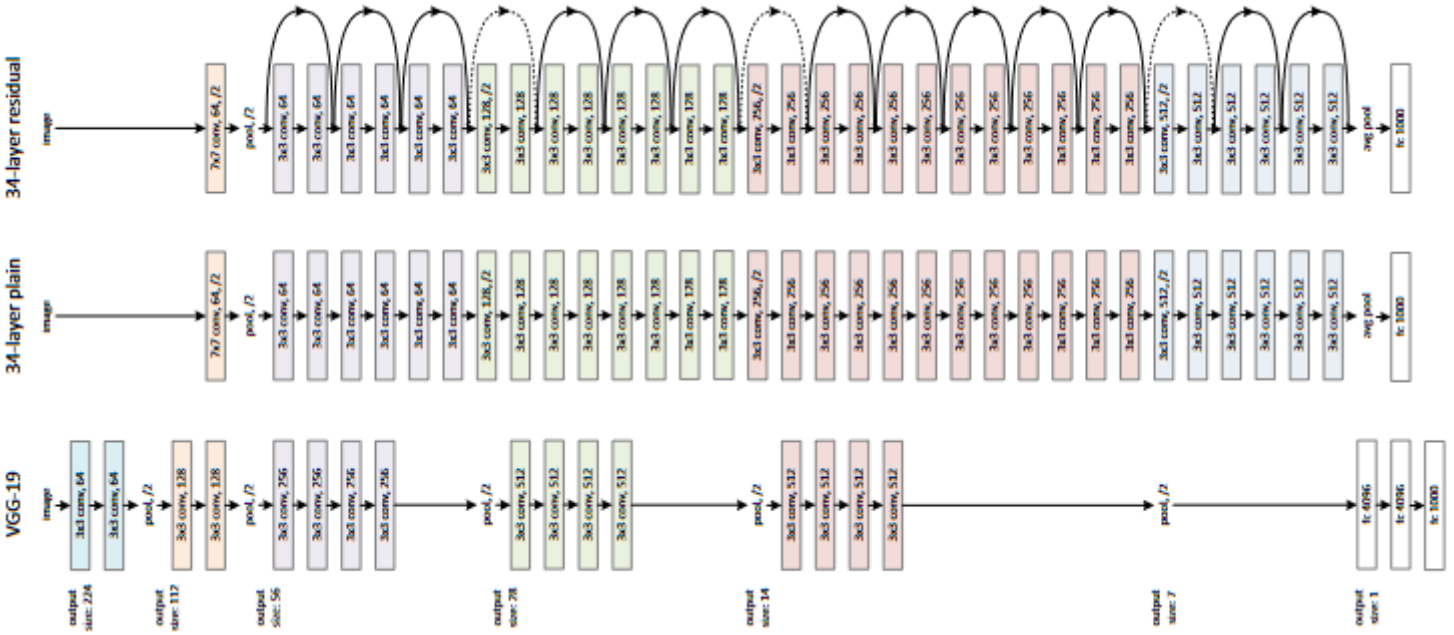

Fig3. 左：VGG-19モデル[40]（196億FLOPs）を参考にしています。中段：34のパラメータ層を持つプレーンネットワーク（36億FLOPs）、右：34のパラメータ層を持つ残差ネットワーク（36億FLOPs）です。点線のショートカットは寸法を大きくしています。表1に詳細と他のバリエーションを示します。

私たちのモデルは，VGGネット[40]よりもフィルタの数が少なく，複雑さも低いことが注目に値します（図3，左）。我々の34レイヤーのベースラインは36億FLOPs（乗算加算）で、VGG-19（196億FLOPs）の18%に過ぎません。

残差ネットワーク。上記のプレーンネットワークをもとに、ショートカット接続（図3右）を挿入して、ネットワークを対応する残差版に変えます。アイデンティティ・ショートカット(式(1))は、入力と出力が同じ寸法の場合には、直接使用することができる(図3の実線のショートカット)。次元が大きくなると（図3の点線のショートカット）、2つの選択肢を考える。(A) ショートカットは，次元が大きくなってもゼロのエントリを追加して，同一性マッピングを実行する．(B) 式(2)の射影ショートカットを使用して次元を合わせる（1×1の畳み込みで行う）。どちらのオプションでも、ショートカットが2つのサイズの特徴マップにまたがる場合は、ストライドを2にして実行されます。

3.4. 実装

ImageNetへの実装は、[21, 40]の方法に従っています。画像は，スケール増強のために，短辺が[256, 480]でランダムにサンプリングされてリサイズされます[40]．224×224のクロップは、画像またはその水平方向の反転からランダムにサンプリングされ、ピクセルごとの平均値が差し引かれます[21]。カラー拡張には，[21]の標準的な方法を用いる．我々は，[16]に従って，各畳み込みの直後と活性化の前にバッチ式正規化（BN）[16]を採用した．重みの初期化は[12]と同様に行い、すべてのプレーン／

リジッドネットをゼロから学習します。ミニバッチサイズを256に設定したSGDを使用します。学習率は0.1から始まり、誤差がピークに達した時点で10で割って、最大60×104回の反復でモデルを学習します。重みの減衰は0.0001，運動量は0.9としました．また，[16]に倣い，ドロップアウト[13]は使用していません．

テストでは，比較研究のために，標準的な 10 クロップテストを採用した[21]．最良の結果を得るためには，[40, 12]のような完全な畳み込み形式を採用し，複数のスケールでスコアを平均化する（画像は，短辺が{224, 256, 384, 480, 640}になるようにリサイズされる）．

4. 実験

ImageNet 分類

我々は，1000個のクラスからなるImageNet 2012分類データセット[35]を用いて，我々の手法を評価した．モデルは，128万枚の訓練画像で学習され，5万枚の検証画像で評価されます．また，テストサーバから報告された100k枚のテスト画像を用いて，最終的な結果を得た．トップ1エラー率とトップ5エラー率の両方を評価します。

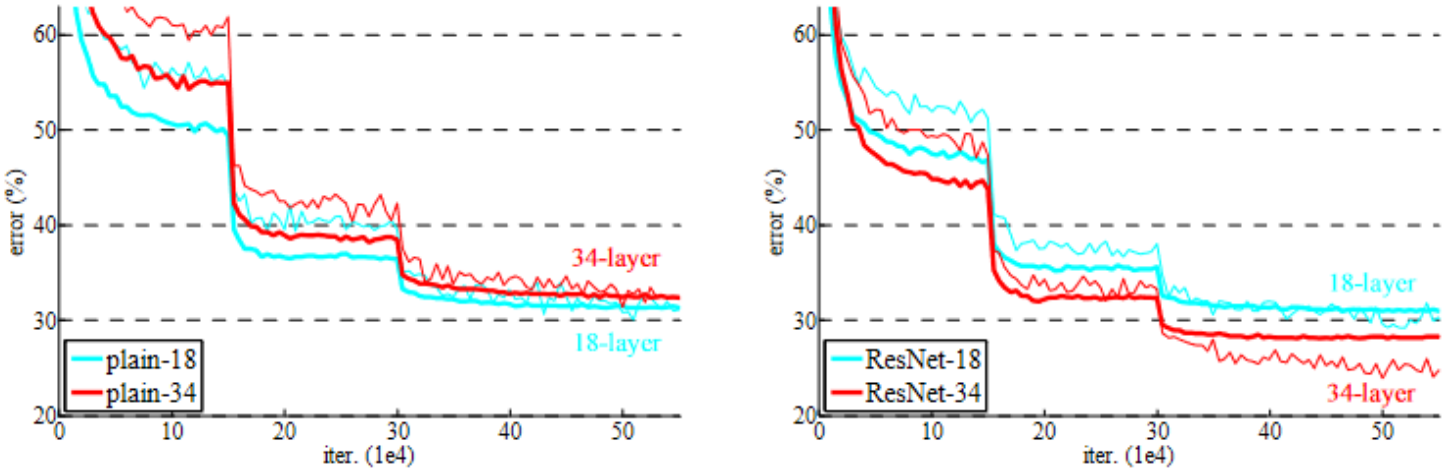

Fig4. ImageNetでの学習。細い曲線は学習誤差、太い曲線は中央の作物の検証誤差を示す。左：18層と34層のプレーンネットワーク。右：18層と34層のプレーンネットワーク。18層と34層のResNets。このプロットでは、残差ネットワークはプレーンネットワークに比べて余分なパラメータを持たない。

プレーンネットワーク。まず，18層と34層のプレーンネットを評価します．34層のプレーンネットは，図3（中）に示すとおりである．18層のプレーンネットも同様の形をしている．詳細なアーキテクチャは表1を参照してください。

表2の結果を見ると，深い34層のプレーンネットは，浅い18層のプレーンネットよりも検証誤差が大きいことがわかる。その理由を明らかにするために，図4（左）では，訓練手順における訓練誤差と検証誤差を比較しています．18層プレーンネットの解空間は34層プレーンネットの解空間の部分空間であるにもかかわらず，訓練手順全体を通して34層プレーンネットの方が学習誤差が大きいという劣化問題が発生していることがわかる．

この最適化の難しさは、勾配の消失が原因であるとは考えにくいことを主張しています。これらのプレーンネットワークはBN[16]を用いて学習されており、前方伝搬する信号が非ゼロの分散を持つことが保証されている。また、BNを用いて、後方に伝搬された勾配が健全なノルムを示すことを検証する。そのため、前方シグナルも後方シグナルも消滅しません。実際、34層のプレーンネットでも競争力のある精度が得られており（表3）、ソルバーがある程度機能していることが示唆される。我々は、深いプレーンネットの収束率が指数関数的に低く、学習誤差の低減に影響を与えているのではないかと推測している。このような最適化の難しさの理由については、今後の研究課題としたい。

残留型ネットワーク 次に，18層および34層の残差ネット（ResNets）を評価します．ベースライン・アーキテクチャは、上記のプレーンネットと同じですが、図3（右）のように、3×3フィルタの各ペアにショートカット接続が追加されています。最初の比較（表2と図4右）では、すべてのショートカットにアイデンティティ・マッピングを使用し、次元が大きくなるとゼロパディングを使用します（オプションA）。そのため、プレーンなカウンターパートと比較して、余分なパラメータはありません。

表2と図4から、3つの主要な見解が得られました。まず、残差学習では状況が逆転し、34層ResNetは18層ResNetよりも優れています（2.8%差）。さらに重要なことは、34層のResNetは学習誤差がかなり少なく、検証データにも一般化できることです。これは、この設定では劣化問題がうまく解決されており、深さを増すことで精度を得ることができたことを示しています。

次に、34層のResNetは、プレーンなResNetと比較して、トップ1エラーを3.5%削減しています（表2）。この比較により，極めて深いシステムにおける残差学習の有効性が検証されました．

最後に，18層のプレーン／残差ネットの精度は同程度ですが（表2），18層のResNetの方が早く収束することにも注目してください（図4右対左）。ネットが「過度に深くない」場合（ここでは18層）でも、現在のSGDソルバーはプレーンネットの良い解を見つけることができます。この場合、ResNetは初期段階での収束を高速化することで、最適化を容易にします。