**Abstract**

ImageNet LSVRC-2010コンテストの120万の高解像度画像を1000の異なるクラスに分類するために、大規模で深い畳み込みニューラルネットワークをトレーニングしました。テストデータでは、トップ1とトップ5のエラー率が37.5％と17.0％を達成しました。これは、従来の最新技術よりもかなり優れています。6,000万のパラメーターと650,000のニューロンを持つニューラルネットワークは、5つの畳み込み層で構成され、そのうちのいくつかは最大プール層、最後に1000カテゴリのソフトマックスを持つ3つの完全に接続された層が続きます。トレーニングを高速化するために、非飽和ニューロンと畳み込み演算の非常に効率的なGPU実装を使用しました。完全に接続された層の過剰適合を減らすために、非常に効果的であることが証明された「ドロップアウト」と呼ばれる最近開発された正則化方法を採用しました。また、ILSVRC-2012コンペティションでこのモデルのバリアントを入力し、2位のエントリで達成された26.2％と比較して、15.3％高いテストエラー率でトップ5に勝利しました。

1. **Intro**

　オブジェクト認識への現在のアプローチは、機械学習法を本質的に使用しています。パフォーマンスを向上させるために、より大きなデータセットを収集し、より強力なモデルを学習し、オーバーフィットを防ぐためのより良いテクニックを使用できます。最近まで、ラベル付き画像のデータセットは比較的小さく、数万枚の画像でした（たとえば、NORB [16]、Caltech-101 / 256 [8、9]、およびCIFAR-10 / 100 [12]）。単純な認識タスクは、このサイズのデータ​​セットを使用して、特にラベル保存変換で強化されている場合、非常によく解決できます．例えば、MNIST数字認識タスクの現在の最高エラー率（<0.3％）は、人間のパフォーマンスに近づいています[4]。しかし、現実的な設定のオブジェクトにはかなりのばらつきがあるため、それらを認識することを学ぶには、はるかに大きなトレーニングセットを使用する必要があります。実際、小さな画像データセットの欠点は広く認識されていますが（たとえば、Pinto et al。[21]）、数百万の画像を含むラベル付きデータセットを収集できるようになったのはごく最近です。新しい大規模なデータセットには、数十万の完全にセグメント化された画像で構成されるLabelMe [23]、および22,000以上のカテゴリの1500万を超えるラベル付き高解像度画像で構成されるImageNet [6]が含まれます。

　数百万の画像から数千のオブジェクトを学習するには、大きな学習能力を持つモデルが必要です。ただし、オブジェクト認識タスクは非常に複雑であるため、ImageNetほど大きなデータセットでもこの問題を特定できないため、モデルには、保有していないすべてのデータを補うための多くの事前知識が必要です。 たたみ込みニューラルネットワーク（CNN）は、そのようなモデルのクラスの1つを構成します[16、11、13、18、15、22、26]。それらの容量は、その深さと幅を変えることによって制御でき、また、画像の性質（つまり、統計の定常性とピクセル依存の局所性）について強力でほとんど正しい仮定を行います。したがって、同様のサイズのレイヤーを持つ標準のフィードフォワードニューラルネットワークと比較して、CNNは接続とパラメーターがはるかに少ないため、トレーニングが容易であり、理論的に最高のパフォーマンスはわずかに劣るだけです。

1. **The Dataset**

　ImageNetは、約22,000のカテゴリに属する​​1500万を超えるラベル付き高解像度画像のデータセットです。画像はウェブから収集され、AmazonのMechanical Turkクラウドソーシングツールを使用して人間のラベラーによってラベル付けされました。2010年から、Pascal Visual Object Challengeの一環として、ImageNet Large-Scale Visual Recognition Challenge（ILSVRC）と呼ばれる年次大会が開催されました。ILSVRCは、1000個のカテゴリごとに約1000個の画像を持つImageNetのサブセットを使用します。全体で、およそ120万のトレーニング画像、50,000の検証画像、および150,000のテスト画像があります。

　ILSVRC-2010は、テストセットラベルが利用できるILSVRCの唯一のバージョンです。したがって、これは、ほとんどの実験を実行したバージョンです。ILSVRC-2012コンペティションでもモデルに参加したため、セクション6では、テストセットラベルが使用できないこのバージョンのデータセットに関する結果も報告します。ImageNetでは、トップ1とトップ5の2つのエラー率を報告するのが一般的です。トップ5エラー率は、正しいラベルがモデルによって最も可能性が高いと考えられる5つのラベルの中にないテスト画像の割合です。

　ImageNetは可変解像度の画像で構成されていますが、システムには一定の入力次元が必要です。したがって、画像を256 x 256の固定解像度にダウンサンプリングしました。長方形の画像が与えられた場合、まず短辺の長さが256になるように画像を再スケーリングし、次に結果の画像から中央の256 x 256パッチを切り取りました。各ピクセルからトレーニングセットの平均アクティビティを減算することを除いて、他の方法で画像を前処理しませんでした。そのため、ピクセルの（中心にある）生のRGB値でネットワークをトレーニングしました。

The Architecture

　私たちのネットワークのアーキテクチャを図2に示します。これには、5つの畳み込み層と3つの完全に接続された8つの学習層が含まれています。以下では、ネットワークアーキテクチャの斬新な機能や珍しい特徴について説明します。セクション3.1〜3.4は、重要性の評価に従ってソートされています。

ReLU Nonlinearity

The standard way to model a neuron’s output f as a function of its input x is with f(x) = tanh(x) or f(x) = (1 + e􀀀x)􀀀1. In terms of training time with gradient descent, these saturating nonlinearities are much slower than the non-saturating nonlinearity f(x) = max(0; x). Following Nair and Hinton [20], we refer to neurons with this nonlinearity as Rectified Linear Units (ReLUs). Deep convolutional neural networks with ReLUs train several times faster than their equivalents with tanh units. This is demonstrated in Figure 1, which shows the number of iterations required to reach 25% training error on the CIFAR-10 dataset for a particular four-layer convolutional network. This plot shows that we would not have been able to experiment with such large neural networks for this work if we had used traditional saturating neuron models.