Abst

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes.

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。

On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art.

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。

The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax.

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カテゴリー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。

To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation.

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。

To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective.

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。

We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

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。

Intro

Current approaches to object recognition make essential use of machine learning methods.

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。

To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting.

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。

Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]).

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]）。

Simple recognition tasks can be solved quite well with datasets of this size, especially if they are augmented with label-preserving transformations.

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

For example, the currentbest error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4].

たとえば、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]。

But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images.

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。

The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

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。

To learn about thousands of objects from millions of images, we need a model with a large learning capacity.

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。

However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots of prior knowledge to compensate for all the data we don’t have.

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。

Convolutional neural networks (CNNs) constitute one such class of models [16, 11, 13, 18, 15, 22, 26].

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]。

Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies).

Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train, while their theoretically-best performance is likely to be only slightly worse.