A review on deep convolutional neural networks

Abst

コンピュータビジョンの問題を解決するための従来の手法が成功するかどうかは、特徴抽出プロセスに大きく依存する。しかし、畳み込みニューラルネットワーク（CNN）は、ドメイン固有の特徴を自動的に学習するための代替手段を提供しています。現在では、コンピュータ・ビジョンの幅広い分野のあらゆる問題が、この新しい手法の観点から再検討されています。そのため、問題に特化したネットワークのタイプを把握することが不可欠です。本研究では、深層学習のフレームワークとして広く利用されている畳み込みニューラルネットワークについて、徹底的な文献調査を行いました。AlexNetを基本的なCNNモデルとし、様々なアプリケーションに対応するために時間をかけて生み出されたすべてのバリエーションと、同じものを実装するために利用可能なフレームワークについて少し検討しました。この記事が、この分野の初心者にとってのガイドとなることを願っています。

1. Introduction

畳み込みニューラルネットワークは，広く用いられている深層学習フレームワークであり[1]，動物の視覚野にヒントを得て開発されたものである[2]．当初は物体認識タスクに広く用いられていましたが，現在では，物体追跡[3]，姿勢推定[4]，テキスト検出・認識[5]，視覚的強調度検出[6]，行動認識[7]，シーンラベリング[8]など，他の領域でも検討されています[9].

1980年のネオコグニトロン[10]がConvNetsの前身とされています。LeNetは、1990年にLeCunらによってConvolutional Neural Networksの先駆的な研究が行われ[11]、その後改良が加えられました[12]。LeNetは、特に手書きの数字を分類するために設計され、前処理なしで入力画像から直接、視覚パターンを認識することに成功しました。しかし，十分な学習データと計算能力がなかったため，このアーキテクチャは複雑な問題ではうまく機能しなかった．その後、2012年にKrizhevskyら[13]がCNNモデルを開発し、ILSVRC競技会でのエラー率を下げることに成功しました[14]。その後、彼らの研究はコンピュータビジョンの分野で最も影響力のあるものの一つとなり、多くの人がCNNアーキテクチャのバリエーションを試すために使用しています。AlexNetは、従来のConvNetsのモデル[15, 16, 17, 18, 19, 20]と比較して、純粋に教師付き学習を使用し、ネットをシンプルに保つために教師なしの事前学習を行わずに、顕著な結果を得ることができました。このアーキテクチャは、5つの畳み込み層と3つの完全連結層を持つLeNetの主要な改良型と考えることができます。AlexNetはILSVRC-2012で

大成功を収めて以来、様々なバリエーションが登場しています。この記事は、この分野の初心者のためのガイドとなるでしょう。

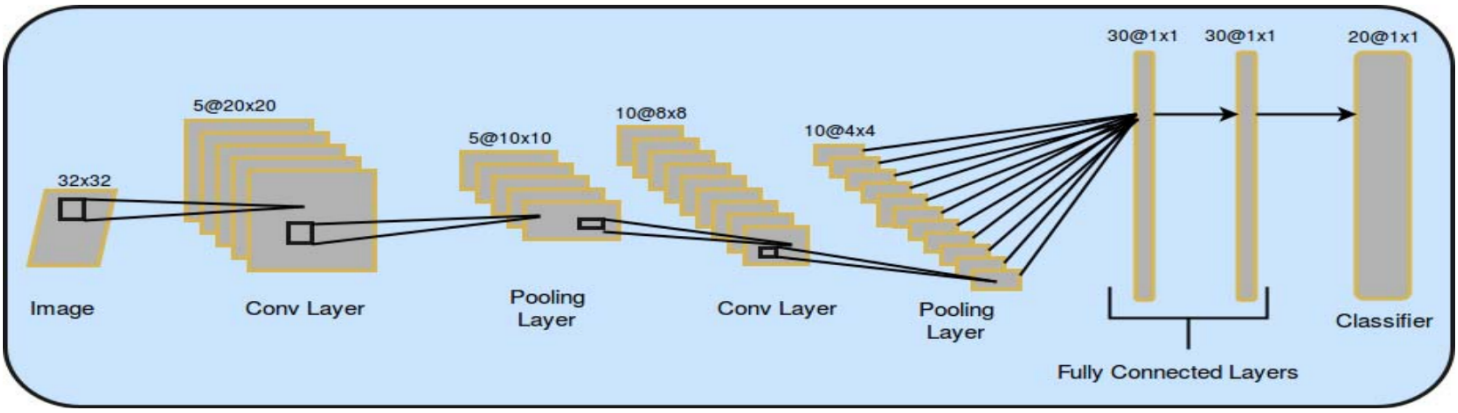

図1. 畳み込み層とプーリング層を交互に配置した基本的なConvNetアーキテクチャ。強調された小箱は受容領域である。接続は、特徴の暗黙的な階層的学習を示している。

論文の構成は以下の通りです。セクションIIでは、Convnet層について説明しています。セクションIIIでは、これらの作業の活性化関数について説明します。セクションIVでは、CNNのトレーニングとテストについて説明します。セクションVでは、一般的なCNNのアーキテクチャについて説明します。セクションVIでは、実装について説明します。セクションVIIでは、未解決の問題について述べ、セクションVIIIでは論文を締めくくります。

2. コンブネットのレイヤー。構造を説明し、各段階でバリエーションを導入

コンベットは通常のニューラルネットワークと非常によく似ており、ニューロンの集合体が非周期的なグラフとして配置されたものとして視覚化されます。ニューラルネットワークとの主な違いは、隠れ層のニューロンが前の層のニューロンのサブセットにしか接続されていないことです。この疎な接続性により、暗黙的に特徴を学習することができます。例えば、第1層で学習されたフィルターはエッジやカラーブロブのセットとして、第2層ではいくつかの形状として視覚化され、次の層のフィルターはオブジェクトのパーツを学習し、最終層のフィルターはオブジェクトを識別することができます。

A. Convolutional Layer

この層は、ほとんどの計算が行われるConvNetの基本単位を形成しています。この層は、ニューロンが配置された特徴マップのセットです。この層のパラメータは、学習可能なフィルタまたはカーネルのセットです。これらのフィルターは、特徴マップと畳み込まれて、別々の2次元活性化マップを生成し、これを奥行き方向に重ねると、出力ボリュームが生成されます。同じ特徴マップにあるニューロンは、重みを共有することで（パラメータ共有）、パラメータの数を少なくしてネットワークの複雑さを軽減しています[21]。2つの層のニューロン間の疎結合の空間的広がりは、受容野と呼ばれるハイパーパラメータです。出力ボリュームの大きさを制御するハイパーパラメータは、デプス（各層のフィルタの数）、ストライド（フィルタの移動）、ゼロパディング（出力の空間的な大きさを制御）です。ConvNetsはバックプロパゲーションを用いて学習され、バックワードパスも同様に畳み込み演算を行いますが、フィルターは空間的に反転します。図2は、コンブネットの基本的な畳み込み演算を示しています。

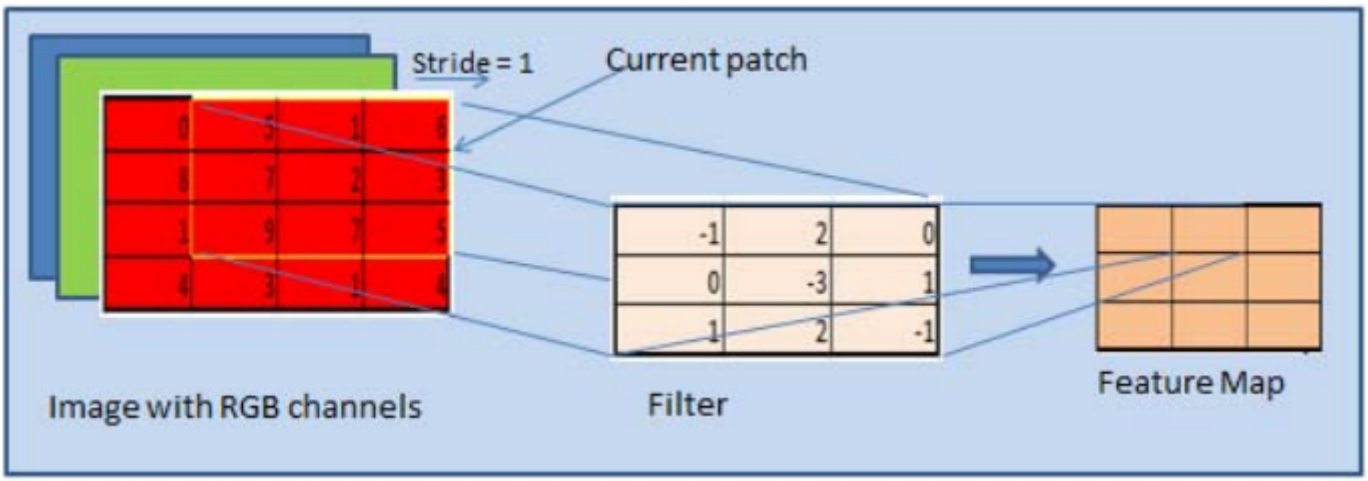

図2. コンボリューション操作。赤(R)、緑(G)、青(B)の色チャンネルを持つ入力画像で、現在の受容領域が黄色のボックスでハイライトされています。コンボリューション操作では、受容領域（R、G、Bチャンネル）とフィルタの対応する要素とのドット積を計算します。 受容野ウィンドウは画像をスライドしながら、空間的に内積を計算し、特徴マップを作成します。

Linら[22]が提案した "Network In Network"(NIN)は、従来のCNNの変種のひとつで、1×1 畳み込みフィルターを従来の線形フィルターから多層パーセプトロン(mlp)に変更し、完全連結層を大域平均プーリング層に変更したものである。その結果、マイクロネットワークがmlpconv層のスタックで構成されていることから、mlpconv層と呼ばれています。CNNとは異なり、NINは潜在的な概念の抽象化能力を高めることができる。彼らは、NINの最後のmlpconv層がカテゴリーの信頼マップであることを証明することに成功し、NINによる物体認識の可能性を導き出した。Fisher YuとVladlen Koltunによる最近の研究[23]では、もともと画像分類のために開発されたCNNを、画像分類とは構造的に異なる

セマンティック・セグメンテーションのような密な予測問題に適用した。解像感を損なうことなく、受容野の指数関数的な拡大をサポートする拡張畳み込みが使用されています。 これは、畳み込み演算子を拡張したもので、拡張されたフィルターの構築を必要としません。密な予測のために特別に設計されたアーキテクチャは、画像分類で使用される伝統的なピラミッド型構造とは対照的に、プーリングやサブサンプリングを行わない、畳み込み層の長方形のプリズムとして配置されています。これにより、緻密な予測のための最先端の結果を得ることができました。

B. Pooling Layer

基本的なConvNetのアーキテクチャには、交互に配置されたConv層とプーリング層があり、後者は活性化マップの空間的次元を（情報を失うことなく）減らし、ネットのパラメータ数を減らし、全体的な計算の複雑さを減らす機能を持っています。これにより、オーバーフィッティングの問題を抑制することができます。一般的なプーリング操作には，最大プーリング，平均プーリング，ストキャスティックプーリング[24]，スペクトルプーリング[25]，空間ピラミッドプーリング[26]，マルチスケールオーダレスプーリング[27]などがある．図2は，マックスプーリングの動作を示している．