

畳み込みでさらに深くなる

クリスチャン・セゲディ1、魏劉2、Yangqing Jia1、ピエール・セルマネット1、スコットリード3、Dragomir Anguelov1、ドゥミトル・エルハン1、ヴィンセントヴァンホーク1、アンドリュー・ラビノビッチ4 1Google Inc. 2ノースカロライナ大学チャペルヒル校 3ミシガン大学、アナーバー 4Magic Leap Inc.

1fszegedy、jiayq、sermanet、dragomir、dumitru、vanhouckeg @ google.com

2wliu@cs.unc.edu 、3reedscott@umich.edu 、4arabinovich@magicleap.com

概要

Im-ageNet大規模視覚認識チャレンジ2014（ILSVRC14）で、分類と検出の新しい最先端を実現する、コードネームInceptionという深い畳み込みニューラルネットワークアーキテクチャを提案します。このアーキテクチャの主な特徴は、ネットワーク内のコンピューティングリソースの使用率が向上していることです。慎重に作成された設計により、計算予算を一定に保ちながら、ネットワークの深さと幅を拡大しました。品質を最適化するために、アーキテクチャの決定は、ヘッブの法則とマルチスケール処理の直感に基づいていました。ILSVRC14の提出で使用された特定のカーネーションの1つは、22層の深さのネットワークであるGoogLeNetと呼ばれ、その品質は分類と検出のコンテキストで評価されます。

1.はじめに

過去3年間で、ディープラーニングと畳み込みネットワークの進歩により、オブジェクトの分類と検出の機能が劇的に向上しました[10]。心強いニュースの1つは、この進歩のほとんどが、より強力なハードウェア、より大きなデータセット、より大きなモデルの結果であるだけでなく、主に新しいアイデア、アルゴリズム、および改善されたネットワークアーキテクチャの結果であるということです。たとえば、ILSVRC 2014コンテストの上位エントリでは、検出目的で同じコンテストの分類データセット以外に新しいデータソースは使用されませんでした。ILSVRC 2014へのGoogLeNetの提出では、実際には2年前のKrizhevsky et al [9]の勝利アーキテクチャの12分の1のパラメータを使用しますが、大幅に正確です。物体検出の面では、最大の利益は、大きなアプリケーションの素朴なアプリケーションからもたらされたものではありません。より大きく、より大きなディープネットワークですが、Girshick et al [6]によるR-CNNアルゴリズムのように、ディープアーキテクチャと古典的なコンピュータビジョンの相乗効果によるものです。

もう1つの注目すべき要因は、モバイルコンピューティングと組み込みコンピューティングの継続的な牽引により、アルゴリズムの効率、特にその電力とメモリの使用が重要性を増していることです。このホワイトペーパーで紹介するディープアーキテクチャの設計につながる考慮事項には、精度の数値に完全に固執するのではなく、この要素が含まれていることは注目に値します。ほとんどの実験では、モデルは推論時に1：5億の乗算加算の計算予算を維持するように設計されているため、純粋に学術的な好奇心になることはありませんが、実際に使用することができます。大規模なデータセットでも、妥当なコストで。

このホワイトペーパーでは、コンピュータビジョン用の効率的なディープニューラルネットワークアーキテクチャに焦点を当てます。コードネームはIn-ceptionです。このアーキテクチャの名前は、Lin et al [12]によるネットワークペーパーのネットワークに由来します。より深く進むために」インターネットミーム[1]。私たちの場合、「深い」という言葉は2つの異なる意味で使用されます。まず、「インセプションモジュール」の形で新しいレベルの組織を導入するという意味と、ネットワークの増加というより直接的な意味です。深さ。一般に、インセプションモデルは、Arora et al [2]による理論的研究からインスピレーションとガイダンスを得ながら、[12]の論理的な集大成と見なすことができます。アーキテクチャの利点は、ILSVRC 2014の分類と検出の課題で実験的に検証されており、現在の最先端技術を大幅に上回っています。

2.関連作業

LeNet-5 [10]以降、畳み込みニューラルネットワーク（CNN）は通常、標準構造になっています。積み重ねられた畳み込み層（オプションでコントラストの正規化と最大プーリングが続く）の後に、1つ以上の完全に接続された層が続きます。この基本設計のバリエーションは、画像分類の文献で広く使用されており、MNIST、CIFAR、特にImageNet分類の課題でこれまでで最高の結果をもたらしています[9、21]。Imagenetなどのより大きなデータセットの場合、最近の傾向は、過剰適合の問題に対処するためにドロップアウト[7]を使用しながら、レイヤー数[12]とレイヤーサイズ[21、14]を増やすことです。

最大プーリング層が正確な空間情報の損失をもたらすという懸念にもかかわらず、[9]と同じ畳み込みネットワークアーキテクチャもローカリゼーション[9、14]、オブジェクト検出[6、14、18]にうまく採用されています。 5]および人間の姿勢推定[19]。

霊長類の視覚野の神経科学モデルに触発されて、Serre等。[15]は、複数のスケールを処理するために、さまざまなサイズの一連の固定ガボールフィルターを使用しました。ここでも同様の戦略を使用します。ただし、[15]の固定2層ディープモデルとは異なり、Inceptionアーキテクチャのすべてのフィルターが学習されます。さらに、インセプションレイヤーは何度も繰り返され、GoogLeNetモデルの場合は22レイヤーのディープモデルになります。

Network-in-Networkは、Lin etalによって提案されたアプローチです。[12]ニューラルネットワークの表現力を高めるため。彼らのモデルでは、追加の1x1畳み込み層がネットワークに追加され、その深さが増しています。このアプローチは、アーキテクチャで多用しています。ただし、私たちの設定では、1x1畳み込みには2つの目的があります。最も重要なのは、主に次元削減モジュールとして使用され、計算上のボトルネックを取り除くためのモジュールです。そうしないと、ネットワークのサイズが制限されます。これにより、パフォーマンスを大幅に低下させることなく、ネットワークの深さだけでなく幅も増やすことができます。

最後に、オブジェクト検出の現在の最先端技術は、Girshick et al。による畳み込みニューラルネットワーク（R-CNN）法を使用した領域です。[6]。R-CNNは、全体的な検出問題を2つのサブ問題に分解します。カテゴリに依存しない方法でオブジェクトの場所の提案を生成するために色やテクスチャなどの低レベルの手がかりを利用することと、CNN分類器を使用してそれらの場所のオブジェクトカテゴリを識別することです。 。このような2段階のアプローチは、低レベルの手がかりによるバウンディングボックスのセグメンテーションの精度と、最先端のCNNの非常に強力な分類力を活用します。検出の提出でも同様のパイプラインを採用しましたが、オブジェクトの境界ボックスのリコールを高めるためのマルチボックス[5]予測や、境界ボックスの提案をより適切に分類するためのアンサンブルアプローチなど、両方の段階での機能強化を検討しました。

3.動機と高レベルの考慮事項

パフォーマンスを向上させる最も簡単な方法は、ディープニューラルネットワークのサイズを大きくすることです。これには、深さの増加とネットの数の両方が含まれます。

これは、特に大量のラベル付きトレーニングデータが利用可能であることを考えると、より高品質のモデルをトレーニングする簡単で安全な方法です。ただし、この単純なソリューションには2つの大きな欠点があります。

サイズが大きいということは、通常、パラメーターの数が多いことを意味します。これにより、特にトレーニングセット内のラベル付きの例の数が限られている場合、拡大したネットワークが過剰適合しやすくなります。強くラベル付けされたデータセットは取得するのに手間と費用がかかるため、これは大きなボトルネックです。多くの場合、専門家の人間の評価者は、以下に示すように、ImageNet（1000クラスのILSVRCサブセットでも）などのさまざまなきめの細かい視覚カテゴリを区別する必要があります。図1。

ネットワークサイズが均一に増加するもう1つの欠点は、計算リソースの使用が劇的に増加することです。たとえば、ディープビジョンネットワークでは、2つの畳み込み層が連鎖している場合、フィルターの数が均一に増加すると、計算が2次関数的に増加します。追加された容量が非効率的に使用される場合（たとえば、ほとんどの重みがゼロに近くなる場合）、計算の多くが無駄になります。計算バジェットは常に有限であるため、パフォーマンスの品質を向上させることが主な目的である場合でも、サイズを無差別に増やすよりも、計算リソースを効率的に分散することをお勧めします。

これらの問題の両方を解決する基本的な方法は、スパース性を導入し、完全に接続されたレイヤーを、コンボリューション内であってもスパースレイヤーに置き換えることです。生物学的システムを模倣することに加えて、これには、アロラらの画期的な研究により、より強固な理論的基盤という利点もあります。[2]。彼らの主な結果は、データセットの確率分布が大規模で非常にまばらなディープニューラルネットワークで表現できる場合、先行するレイヤーのアクティブ化の相関統計を分析することにより、レイヤーごとに最適なネットワークトポロジを構築できることを示しています。相関性の高い出力を持つニューロンのクラスタリング。厳密な数学的証明には非常に強い条件が必要ですが、このステートメントが よく知られているヘッブの法則と共鳴します。一緒に発火し、一緒に配線するニューロンは、実際には、それほど厳しくない条件下でも、根底にある考えが適用可能であることを示唆しています。

残念ながら、今日のコンピューティングインフラストラクチャは、不均一なスパースデータ構造の数値計算に関しては非常に非効率的です。算術演算の数が100削減されたとしても、ルックアップとキャッシュミスのオーバーヘッドが支配的になります。スパース行列への切り替えは効果がない可能性があります。基盤となるCPUまたはGPUハードウェアの細部を活用して、非常に高速で高密度のマトリックス乗算を可能にする、着実に改善され高度に調整された数値ライブラリを使用することで、ギャップはさらに拡大します[16、9]。また、不均一なスパースモデルには、より高度なエンジニアリングおよびコンピューティングインフラストラクチャが必要です。現在のほとんどの視覚指向の機械学習システムは、畳み込みを採用するという理由だけで、空間領域のスパース性を利用しています。しかしながら、畳み込みは、前のレイヤーのパッチへの密な接続のコレクションとして実装されます。ConvNetsは、[11]以来、対称性を破り、学習を改善するために、特徴次元でランダムでスパースな接続テーブルを伝統的に使用してきましたが、パラメーターをさらに最適化するために、傾向は[9]との完全な接続に戻りました。コンピュータビジョンの現在の最先端のアーキテクチャは、均一な構造を持っています。多数のフィルターとより大きなバッチサイズにより、高密度の計算を効率的に使用できます。

これは、次の中間ステップ、つまり理論で示唆されているようにフィルターレベルのスパース性を利用するが、密な行列での計算を利用することによって現在のハードウェアを拡張するアーキテクチャに希望があるかどうかという疑問を提起します。スパース行列の計算に関する膨大な文献（[3]など）は、スパース行列を比較的密度の高い部分行列にクラスタリングすると、スパース行列の乗算に対して競争力のあるパフォーマンスが得られる傾向があることを示唆しています。近い将来、不均一な深層学習アーキテクチャの自動構築に同様の方法が利用されると考えるのは、遠慮がちではないようです。

インセプションアーキテクチャは、ビジョンネットワークの[2]によって暗示されるスパース構造を近似しようとする高度なネットワークトポロジ構築アルゴリズムの仮想出力を評価し、高密度ですぐに利用できるコンポーネントによって仮想結果をカバーするためのケーススタディとして始まりました。非常に投機的な取り組みであるにもかかわらず、[12]に基づく参照ネットワークと比較した場合、早い段階で適度な増加が観察されました。少し調整することで、ギャップが広がり、Inceptionは、[6]と[5]のベースネットワークとしてのローカリゼーションとオブジェクト検出のコンテキストで特に有用であることが証明されました。興味深いことに、元のアーキテクチャの選択のほとんどは、個別に徹底的に質問およびテストされていますが、ローカルで最適に近いことが判明しました。ただし、注意が必要です。インセプションアーキテクチャはコンピュータビジョンで成功を収めていますが、これがその構築につながった指針に起因するのかどうかは依然として疑問です。これを確認するには、はるかに徹底的な分析と検証が必要になります。

4.アーキテクチャの詳細

インセプションアーキテクチャの主なアイデアは、畳み込みビジョンネットワークの最適なローカルスパース構造をどのように近似し、すぐに利用できる高密度コンポーネントでカバーできるかを検討することです。並進不変性を仮定することは、私たちのネットワークが複雑なビルディングブロックから構築されることを意味することに注意してください。必要なのは、最適なローカル構造を見つけて、それを空間的に繰り返すことです。アロラら。[2]は、最後の層の相関統計を分析し、それらを相関の高いユニットのグループにクラスター化する必要がある層ごとの構築を提案しています。これらのクラスターは次のレイヤーのユニットを形成し、前のレイヤーのユニットに接続されます。前のレイヤーの各ユニットが入力画像のある領域に対応し、これらのユニットがフィルターバンクにグループ化されていると仮定します。下位層（入力に近い層）では、相関ユニットはローカル領域に集中します。したがって、[12]で提案されているように、多くのクラスターが1つの領域に集中し、次のレイヤーで1x1畳み込みのレイヤーでカバーできるようになります。ただし、より大きなパッチでの畳み込みによってカバーできる、より空間的に分散したクラスターの数が少なくなり、より大きな領域でのパッチの数が減少することも予想できます。パッチアラインメントの問題を回避するために、Incep-tionアーキテクチャの現在の化身はフィルターサイズ1x1、3x3、および5x5に制限されています。この決定は、必要性よりも利便性に基づいていました。また、提案されたアーキテクチャは、これらすべてのレイヤーと、次のステージの入力を形成する単一の出力ベクトルに連結された出力フィルターバンクの組み合わせであることも意味します。さらに、プーリング操作は現在のコンボリューションネットワークの成功に不可欠であるため、そのような各段階で代替のパラレルプーリングパスを追加すると、さらに有益な効果が得られるはずです（図2（a）を参照）。

これらの「インセプションモジュール」は互いに積み重ねられているため、それらの出力相関統計は変化するはずです。より高い抽象化の特徴がより高い層によってキャプチャされるため、それらの空間集中は減少すると予想されます。これは、上位層に移動するにつれて、3x3と5x5の畳み込みの比率が増加するはずであることを示唆しています。

上記のモジュールの大きな問題の1つは、少なくともこのナイーブな形式では、わずかな数の5x5コンボリューションでさえ、多数のフィルターを備えたコンボリューションレイヤーの上に非常に高価になる可能性があることです。この問題は、プーリングユニットがミックスに追加されるとさらに顕著になります。出力フィルターの数はnumに等しくなります。前の段階のフィルターのber。プーリング層の出力と畳み込み層の出力をマージすると、ステージからステージへの出力数が必然的に増加します。このアーキテクチャは最適なスパース構造をカバーしているかもしれませんが、それは非常に非効率的であり、数段階以内に計算の爆発につながります。

これは、インセプションアーキテクチャの2番目のアイデアにつながります。それは、計算要件が大幅に増加する場合は常に、次元を慎重に削減することです。これは、埋め込みの成功に基づいています。低次元の埋め込みでも、比較的大きな画像パッチに関する多くの情報が含まれている可能性があります。ただし、埋め込み-dingは情報を高密度の圧縮形式で表し、圧縮された情報は処理が困難です。表現は、ほとんどの場所でスパースに保たれ（[2]の条件で要求されるように）、信号をまとめて集約する必要がある場合にのみ信号を圧縮する必要があります。つまり、1x1畳み込みは、高価な3x3および5x5畳み込みの前に削減を計算するために使用されます。縮小として使用されることに加えて、それらはまた、それらを二重の目的にする整流された耳の活性化の使用を含みます。

一般に、インセプションネットワークは、上記のタイプのモジュールが互いに積み重ねられたネットワークであり、グリッドの解像度を半分にするためにストライド2の最大プーリングレイヤーがときどきあります。技術的な理由（トレーニング中のメモリ効率）から、Inceptionモジュールは、従来の畳み込み方式で下位層を維持しながら、上位層でのみ使用を開始することが有益であるように思われました。これは厳密には必要ではなく、現在の実装におけるインフラストラクチャの非効率性を反映しているだけです。

このアーキテクチャの有用な側面は、後の段階での計算の複雑さの制御されていない爆発なしに、各段階でユニットの数を大幅に増やすことができることです。これは、パッチサイズが大きい高価な畳み込みの前に次元削減を広く使用することによって実現されます。さらに、設計は、視覚情報をさまざまなスケールで処理してから集約し、次の段階でさまざまなスケールの特徴を同時に抽象化できるようにするという実際的な直感に従います。

計算リソースの使用法が改善されたことで、計算が困難になることなく、各ステージの幅とステージ数の両方を増やすことができます。Inceptionアーキテクチャを利用して、わずかに劣るが計算コストの低いバージョンを作成できます。使用可能なすべてのノブとレバーにより、計算リソースのバランスを制御できるため、同様に実行されるネットワークよりも3〜10高速なネットワークが実現します。ただし、この時点では、慎重な手動設計が必要です。

5. GoogLeNet

「GoogLeNet」という名前は、ILSVRC2014コンペティションの提出で使用されたInceptionアーキテクチャの特定の化身を指します。また、品質がわずかに優れた、より深く幅の広いInceptionネットワークを使用しましたが、それをアンサンブルに追加しても、結果はわずかにしか改善されなかったようです。経験的証拠が正確なアーキテクチャパラメータの影響は比較的小さいことを示唆しているため、そのネットワークの詳細は省略します。表1は、コンテストで使用されるInceptionの最も一般的なインスタンスを示しています。このネットワーク（さまざまな画像パッチサンプリング方法でトレーニングされた）は、アンサンブルの7つのモデルのうち6つに使用されました。

Incep-tionモジュール内のものを含むすべての畳み込みは、修正された線形アクティベーションを使用します。私たちのネットワークの受容野のサイズは、RGB色空間で224x 224であり、平均はゼロです。「＃3x3reduce」および「＃5x5 reduce」は、3x3および5x5畳み込みの前に使用されたリダクションレイヤーの1x1フィルターの数を表します。pool proj列に組み込まれたmax-poolingの後、プロジェクションレイヤーの1x1フィルターの数を確認できます。これらのすべての縮小/投影レイヤーは、修正された線形アクティベーションも使用します パラメータ付きのレイヤーのみをカウントする場合、ネットワークの深さは22レイヤーです（プーリングもカウントする場合は27レイヤー）。ネットワークの構築に使用されるレイヤー（独立したビルディングブロック）の総数は約100です。正確な数は、機械学習インフラストラクチャによってレイヤーがどのようにカウントされるかによって異なります。分類器の前の平均プーリングの使用は[12]に基づいていますが、実装には追加の線形レイヤーがあります。線形レイヤーを使用すると、ネットワークを他のラベルセットに簡単に適合させることができますが、主に便宜上使用されており、大きな効果は期待できません。完全に接続されたレイヤーから平均プーリングに移行すると、トップ1の精度が約0.6％向上することがわかりましたが、完全に接続されたレイヤーを削除した後でも、ドロップアウトの使用は依然として不可欠でした。

ネットワークの深さが比較的大きいことを考えると、効果的な方法ですべてのレイヤーを介して勾配を伝播する機能が懸念事項でした。このタスクでの浅いネットワークの強力なパフォーマンスは、ネットワークの中央のレイヤーによって生成される機能が非常に識別力があることを示しています。これらの中間層に接続された補助分類器を追加することにより、分類器の下位段階での識別が期待されました。これは、正則化を提供しながら、勾配消失問題と戦うと考えられていました。これらの分類器は、Inception（4a）および（4d）モジュールの出力の上に配置されたより小さな畳み込みネットワークの形式を取ります。トレーニング中、それらの損失は、割引の重みでネットワークの総損失に追加されます（補助分類器の損失は0.3で重み付けされました）。推論時に、これらの補助ネットワークは破棄されます。その後の対照実験では、補助ネットワークの効果は比較的小さく（約0.5％）、同じ効果を達成するために必要なのはそのうちの1つだけであることが示されています。

補助分類器を含む、側面の追加ネットワークの正確な構造は次のとおりです。

* 5x5フィルターサイズとストライド3の平均プーリング層で、（4a）の場合は4x4x512の出力、（4d）ステージの場合は4x4x528の出力になります。
* 次元の縮小と修正された線形アクティブ化のための128個のフィルターを備えた1x1畳み込み。
* 1024ユニットと修正された線形アクティベーションを備えた完全に接続されたレイヤー。
* ドロップされた出力の比率が70％のドロップアウトレイヤー。
* 分類器としてソフトマックス損失を持つ線形層（メイン分類器と同じ1000クラスを予測しますが、推論時に削除されます）。

結果として得られるネットワークの概略図を図3に示します。

6.トレーニング方法

GoogLeNetネットワークは、DistBe-lief [4]分散型機械学習システムを使用して、最新の量のモデルとデータ並列処理を使用してトレーニングされました。CPUベースの実装のみを使用しましたが、概算では、GoogLeNetネットワークは1週間以内に少数のハイエンドGPUを使用して収束するようにトレーニングでき、主な制限はメモリ使用量です。私たちのトレーニングでは、0.9モーメントの非同期確率的勾配降下法[17]、固定学習率スケジュール（8エポックごとに学習率を4％ずつ減少）を使用しました。Polyak平均化[13]を使用して、推論時に使用される最終モデルを作成しました。

画像のサンプリング方法は、競争につながる数か月の間に大幅に変更され、すでに収束したモデルは、ドロップアウトや学習率などの変更されたハイパーパラメータと組み合わせて、他のオプションでトレーニングされました。したがって、これらのネットワークをトレーニングするための最も効果的な単一の方法に明確なガイダンスを与えることは困難です。さらに複雑なことに、いくつかのモデルは主に小さな相対的な作物で訓練され、他のモデルは[8]に触発されて大きなもので訓練されました。それでも、競合後に非常にうまく機能することが確認された1つの処方には、アスペクト比がに制限された画像領域の8％から100％の間でサイズが均等に分散されている画像のさまざまなサイズのパッチのサンプリングが含まれます。間隔[34; 43]。また、

1. ILSVRC2014分類チャレンジの設定と結果

ILSVRC 2014分類の課題には、画像をImagenet階層内の1000個のリーフノードカテゴリの1つに分類するタスクが含まれます。トレーニング用に約120万枚の画像、検証用に50,000枚、テスト用に100,000枚の画像があります。各画像は1つのグラウンドトゥルースカテゴリに関連付けられており、パフォーマンスは最高スコアの分類子予測に基づいて測定されます。通常、2つの数値が報告されます。グラウンドトゥルースを最初の予測クラスと比較する上位1の正解率と、グラウンドトゥルースを最初の5つの予測クラスと比較する上位5のエラー率です。画像は次のように見なされます。グラウンドトゥルースがトップ5に含まれる場合、ランクに関係なく正しく分類されます。チャレンジでは、ランク付けの目的で上位5つのエラー率を使用します。

トレーニングに外部データを使用せずにチャレンジに参加しました。このホワイトペーパーで前述したトレーニング手法に加えて、テスト中に一連の手法を採用して、より高いパフォーマンスを実現しました。これについて次に説明します。

1. 同じGoogLeNetモデルの7つのバージョン（1つのより広いバージョンを含む）を個別にトレーニングし、それらを使用してアンサンブル予測を実行しました。これらのモデルは、同じ初期化（見落としのために同じ初期重みを使用した場合でも）と学習率ポリシーを使用してトレーニングされました。それらは、サンプリング方法とランダム化された入力画像の順序のみが異なりました。
2. テスト中、Krizhevskyらよりも積極的なトリミングアプローチを採用しました。[9]。具体的には、画像のサイズを4つのスケールに変更し、短い方の寸法（高さまたは幅）をそれぞれ256、288、320、および352にし、これらのサイズ変更された画像の左、中央、および右の正方形を取ります（ポートレート画像の場合、上部、中央、下部の正方形を取ります）。次に、各正方形について、4つのコーナーと中央の224 224の切り抜き、および224224にサイズ変更された正方形とそれらのミラーバージョンを取得します。これにより、画像あたり4 3 6 2 = 144クロップになります。同様のアプローチが、前年のエントリでAndrew Howard [8]によって使用されました。これは、提案されたスキームよりもわずかにパフォーマンスが悪いことを経験的に検証しました。このような積極的なトリミングは、実際のアプリケーションでは必要ない場合があることに注意してください。
3. ソフトマックス確率は、最終的な予測を得るために、複数の作物とすべての個々の分類器で平均化されます。私たちの実験では、作物の最大プーリングや分類器の平均化など、検証データに対する代替アプローチを分析しましたが、単純な平均化よりもパフォーマンスが低下します。

このホワイトペーパーの残りの部分では、最終的な提出物の全体的なパフォーマンスに寄与する複数の要因を分析します。

チャレンジへの最終的な提出では、検証データとテストデータの両方で上位5つのエラーが6.67％得られ、他の参加者の中で1位にランクされています。これは、2012年のSuperVisionアプローチと比較して56.5％の相対的削減であり、分類器のトレーニングに外部データを使用した前年の最良のアプローチ（Clarifai）と比較して約40％の相対的削減です。表2は、過去3年間で最もパフォーマンスの高いアプローチのいくつかの統計を示しています。

また、モデルの数とモデルの数を変えることにより、複数のテストの選択肢のパフォーマンスを分析して報告します。 表3の画像を予測するときに使用されたクロップの数。1つのモデルを使用する場合、検証データでトップ1エラー率が最も低いモデルを選択しました。テストデータの統計に過剰適合しないように、すべての数値が検証データセットに再移植されます。

1. ILSVRC2014検出チャレンジの設定と結果

ILSVRC検出タスクは、200の可能なクラスの中から画像内のオブジェクトの周囲にバウンディングボックスを作成することです。検出されたオブジェクトは、グラウンドトゥルースのクラスと一致し、境界ボックスが少なくとも50％重なっている場合（Jaccardインデックスを使用）、正しいものとしてカウントされます。無関係な検出は誤検知としてカウントされ、ペナルティが課せられます。分類タスクとは異なり、各画像には多くのオブジェクトが含まれる場合と含まれない場合があり、それらのスケールは異なる場合があります。結果は、平均平均精度（mAP）を使用して報告されます。GoogLeNetが検出のために採用したアプローチは、[6]によるR-CNNに似ていますが、領域分類子としてInceptionモデルで拡張されています。さらに、領域提案ステップは、選択的検索[20]アプローチとマルチボックス[5]予測を組み合わせて、より高いオブジェクト境界ボックスのリコールを実現することで改善されています。誤検知の数を減らすために、スーパーピクセルサイズが2増加しました。これにより、選択的検索アルゴリズムからの提案が半分になります。マルチボックス[5]からの200の地域提案を追加し直した結果、[6]で使用された提案の合計で約60％になり、カバレッジが92％から93％に増加しました。全体的な効果-カバー数を増やして提案数を削減する効果-単一モデルの場合の平均精度が1％向上します。最後に、各地域を分類するときに、6つのGoogLeNetのアンサンブルを使用します。これにより、精度が40％から43.9％に向上します。R-CNNとは異なり、時間がないため、バウンディングボックス回帰を使用しなかったことに注意してください。マルチボックス[5]からの200の地域提案を追加し直した結果、[6]で使用された提案の合計で約60％になり、カバレッジが92％から93％に増加しました。全体的な効果-カバー数を増やして提案数を削減する効果-単一モデルの場合の平均精度が1％向上します。最後に、各地域を分類するときに、6つのGoogLeNetのアンサンブルを使用します。これにより、精度が40％から43.9％に向上します。R-CNNとは異なり、時間がないため、バウンディングボックス回帰を使用しなかったことに注意してください。マルチボックス[5]からの200の地域提案を追加し直した結果、[6]で使用された提案の合計で約60％になり、カバレッジが92％から93％に増加しました。全体的な効果-カバー数を増やして提案数を削減する効果-単一モデルの場合の平均精度が1％向上します。最後に、各地域を分類するときに、6つのGoogLeNetのアンサンブルを使用します。これにより、精度が40％から43.9％に向上します。R-CNNとは異なり、時間がないため、バウンディングボックス回帰を使用しなかったことに注意してください。各地域を分類するときは、6つのGoogLeNetのアンサンブルを使用します。これにより、精度が40％から43.9％に向上します。R-CNNとは異なり、時間がないため、バウンディングボックス回帰を使用しなかったことに注意してください。各地域を分類するときは、6つのGoogLeNetのアンサンブルを使用します。これにより、精度が40％から43.9％に向上します。R-CNNとは異なり、時間がないため、バウンディングボックス回帰を使用しなかったことに注意してください。

最初に上位の検出結果を報告し、検出タスクの初版以降の進捗状況を示します。2013年の結果と比較すると、精度はほぼ2倍になっています。最高のパフォーマンスを発揮するチームはすべて、畳み込みネットワークを使用しています。表4の公式スコアと、各チームの一般的な戦略（外部データ、アンサンブルモデル、またはコンテキストモデルの使用）を報告します。外部データは通常、モデルを事前トレーニングするためのILSVRC12分類データであり、後で検出データで改良されます。一部のチームは、ローカリゼーションデータの使用についても言及しています。ローカリゼーションタスクのバウンディングボックスの大部分は検出データセットに含まれていないため、分類が事前トレーニングに使用されるのと同じ方法で、このデータを使用して一般的なバウンディングボックス回帰分析を事前トレーニングできます。

表5では、単一のモデルのみを使用して結果を比較しています。最高のパフォーマンスを発揮するモデルはDeepInsightによるもので、驚くべきことに、3つのモデルのアンサンブルで0.3ポイントしか改善されませんが、GoogLeNetはアンサンブルで非常に強力な結果を取得します。

9.結論

私たちの結果は、すぐに利用できる高密度のビルディングブロックによって予想される最適なスパース構造を近似することが、コンピュータビジョンのニューラルネットワークを改善するための実行可能な方法であるという確かな証拠をもたらします。この方法の主な利点は、浅くて狭いアーキテクチャと比較して、計算要件がわずかに増加しても品質が大幅に向上することです。

コンテキストを利用せず、バウンディングボックス回帰を実行しなかったにもかかわらず、オブジェクト検出作業は競争力があり、Inceptionアーキテクチャの長所のさらなる証拠を示唆しています。 分類と検出の両方で、同様の深さと幅のはるかに高価な非開始タイプのネットワークによって、同様の品質の結果を達成できることが期待されます。それでも、私たちのアプローチは、よりまばらなアーキテクチャへの移行が実行可能であり、一般的に有用なアイデアであるという確かな証拠をもたらします。これは、[2]に基づいて自動化された方法で、よりまばらでより洗練された構造を作成すること、およびInceptionアーキテクチャの洞察を他のドメインに適用することに向けた将来の作業を示唆しています。

参考文献

[1]知っている 君の ミーム： 我々 必要 に 行く もっと深く。

http://knowyourmeme.com/memes/we-need-to-go-deeper。

アクセス：2014-09-15。

1. S. Arora、A。Bhaskara、R。Ge、およびT.Ma。いくつかの深い表現を学習するための確かな限界。CoRR、abs / 1310.6343、2013。
2. UVC¸atalyurek、¨C。Aykanat、およびB.Uc¸ar。2次元スパース行列分割について：モデル、メソッド、およびレシピ。SIAM J.Sci。Comput。、32（2）：656–683、2010年2月。
3. J.ディーン、G。コラード、R。モンガ、K。チェン、M。デヴィン、M。マオ、M。ランザト、A。シニア、P。タッカー、K。ヤン、QVル、AYNg。大規模な分散型ディープネットワーク。P. Bartlett、F。Pereira、C。Burges、L。Bot-tou、およびK. Weinberger、編集者、NIPS、1232〜1240ページ。2012年。
4. D. Erhan、C。Szegedy、A。Toshev、およびD.Anguelov。ディープニューラルネットワークを使用したスケーラブルなオブジェクト検出。CVPR、2014年。
5. RB Girshick、J。Donahue、T。Darrell、およびJ.Malik。正確なオブジェクト検出とセマンティックセグメンテーションのための豊富な機能階層。コンピュータビジョンとパターン認識、2014年。CVPR2014。IEEE会議、2014年。

[7] GEヒントン、N。Srivastava、A。Krizhevsky、I。Sutskever、およびR.Salakhutdinov。機能検出器の共適応を防ぐことによるニューラルネットワークの改善。CoRR、abs / 1207.0580、2012。

1. AGハワード。深い畳み込みニューラルネットワークベースの画像分類に関するいくつかの改善。CoRR、abs / 1312.5402、2013。
2. A. Krizhevsky、I。Sutskever、およびG.Hinton。深い畳み込みニューラルネットワークを使用した画像-agenet分類。神経情報処理の進歩-ingSystems 25、pages 1106–1114、2012。
3. Y. LeCun、B。Boser、JS Denker、D。Henderson、RE Howard、W。Hubbard、およびLDJackel。手書きの郵便番号認識に適用されるバックプロパゲーション。Neural Comput。、1（4）：541–551、1989年12月。
4. Y. LeCun、L。Bottou、Y。Bengio、およびP.Haffner。グラデーションベースの学習は、ドキュメントの認識に適用されます。IEEEの議事録、86（11）：2278–2324、1998。
5. M.リン、Q。チェン、S。ヤン。ネットワーク内のネットワーク。CoRR、abs / 1312.4400、2013。
6. BTポリアクとABジュディツキー。平均化による確率的近似の加速。SIAM J. Control Optim。、30（4）：838–855、1992年7月。
7. P. Sermanet、D。Eigen、X。Zhang、M。Mathieu、R。Fer-gus、およびY. LeCun 克服：畳み込みネットワークを使用した統合された認識、ローカリゼーション、および検出。CoRR、abs / 1312.6229、2013。
8. T. Serre、L。Wolf、SM Bileschi、M。Riesenhuber、T。Poggio 皮質のようなメカニズムによる堅牢なオブジェクト認識。IEEETrans。パターンアナル。マッハ。Intell。、29（3）：411–426、2007。
9. F.ソングとJ.ドンガラ。1000cpuコアの共有メモリメニーコアシステムでの行列計算のスケールアップ。スーパーコンピューティングに関する第28回ACM国際会議の議事録、ICS '14、333〜342ページ、ニューヨーク、ニューヨーク、米国、2014年。ACM。
10. I. Sutskever、J。Martens、GE Dahl、およびGEHinton。深層学習における初期化と勢いの重要性について。ICML、JMLR Proceedingsの第28巻、1139〜1147ページ。JMLR.org、2013年。
11. C. Szegedy、A。Toshev、およびD.Erhan。オブジェクト検出のためのディープニューラルネットワーク。CJC Burges、L。Bot-tou、Z。Ghahramani、およびKQ Weinberger、編集者、NIPS、2553〜2561ページ、2013年。
12. A.トシェフとC.セゲディ。ディープポーズ：ディープニューラルネットワークを介した人間のポーズの推定。CoRR、abs / 1312.4659、2013。
13. KEA van de Sande、JRR Uijlings、T。Gevers、およびAWMSmeulders。オブジェクト認識の選択的検索としてのセグメンテーション。2011年コンピュータビジョン国際会議の議事録、ICCV '11、1879〜1886ページ、米国ワシントンDC、2011年。IEEEComputerSociety。
14. MDツァイラーとR.ファーガス。畳み込みネットワークの視覚化と理解。DJ Fleet、T。Pa-jdla、B。Schiele、およびT. Tuytelaars、編集者、ECCV、コンピュータサイエンスのレクチャーノートの8689巻、818〜833ページ。Springer、2014年。

