Analyzing the Performance of Multilayer NeuralNetworks for Object Recognition

備考

3章まとめ

- 訓練データを増やすと、微調整されたネットワークとゼロから学習させたネットワークとでほぼ同等の性能を得ることができる。
- クラス選択性は、1層から7層に向けて増加している。
- 微調整によるエントロピーの変化は全結合層で優位である。
- 少量データ(16k 程度の画像量)で行う微調整は、全結合層を「再配線」するに等しい結果となった。
- より多くの微調整データ(13k+25k 程度の画像量)が利用可能な場合、全てのパラメータを微調整することによる利益が期待できる。

著者

Pulkit Agrawal, Ross Girshick, and Jitendra Malik

掲載

European Conference on Computer Vision(ECCV), pp 329--344, 2014.

Abstract

ここ2年間で、畳み込みニューラルネットワーク（CNN）は、標準的な認識データセットやタスクで目覚しい成果を上げてきた。CNNベースの特徴量は、SIFTやHOGのような人工的な表現にすぐに取って代わる準備ができているように思われる。しかし，SIFTやHOGに比べて，大規模なCNNが学習する特徴の性質についてはあまり理解されていない．この論文では，CNNの特徴学習のいくつかの側面を実験的に探り，実務家がCNNをコンピュータビジョン問題に適用する方法について，証拠に裏付けられた有用な直観を得ることができるようにすることを目的としている．

Introduction

過去2年間のベンチマーク視覚認識タスクでの一連の結果から、畳み込みニューラルネットワーク（CNN）[6,14,18]が、さまざまな問題においてSIFT[15]やHOG[2]のような人工的な特徴に取って代わる可能性が高いことが実証された。この一連の流れは，Krizhevskyら[11]が報告した画期的なImageNet[3]の分類結果から始まった．その後すぐに，Donahueら[4]は，ImageNet分類のために訓練された同じネットワークが効果的なブラックボックス特徴抽出器であることを示した．彼らはCNN特徴を用いて、いくつかの標準的な画像分類データセットで最先端の結果を報告した。同時に，Girshickら[7]は，このネットワークがどのように物体検出に適用できるかを示した．R-CNNと呼ばれる彼らのシステムは，ボトムアップのグルーピング機構（例えば，選択的探索[23]）によって生成された物体提案を分類する．検出の訓練データが限られているため，彼らは，CNNを最初に教師付きでImageNet分類のための事前訓練を行い，その後，小さなPASCAL検出データセ

ット[5]上で微調整を行う転移学習戦略を提案した．この最初の結果が得られて以来，他のいくつかの論文がより広範囲のタスクで同様の結果を報告している（例えば，[17]のRazavianらが報告した結果などを参照のこと）．

SIFTやHOGのような特徴変換は、空間ブロックに配置された有向エッジフィルタの応答のヒストグラムとして直感的に解釈できる。しかし、CNNの異なる層がどのような視覚的特徴をコード化しているのかについては、ほとんど理解できていない。今後数年の間に、CNNが提供する豊かな特徴階層が、コンピュータビジョンモデルの主要な特徴抽出器として浮上してくるであろうことを考えると、このような理解を深めることは、興味深い科学的探求であり、CNNを用いたコンピュータビジョン手法の設計の指針となる必須の課題であると考えている。そこで本論文では、経験的なレンズを通してCNNのいくつかの側面を研究する。

1.1. Summary of Findings(所見のまとめ)

Effects of Fine-Tuning and Pre-training (微調整とプレトレーニングの効果).

Girshickら[7]は、教師付き事前訓練や微調整が訓練データが少ない場合に有効であることを示した。しかし，学習データが豊富になるとどうなるかについては調べていない．我々は，ランダムな初期化（ImageNetによる教師付き事前学習なし）からR-CNNを学習する際に，適度な量の検出訓練データ（37k個の基底真実境界ボックス）であれば，良好な性能を得ることができることを示した．しかし，このデータ領域では，教師付き事前訓練が有益であり，検出性能の大幅な向上につながることも示しています．画像分類についても同様の結果が得られた。

ImageNet Pre-training Does not Overfit (ImageNetの事前トレーニングはオーバーフィットしない).

教師付き事前学習を使用する際に懸念されることの 1 つは、例えば ImageNet へのより良いモデルフィットを達成すると、学習した特徴を別のデータセットやタスクに適用する際に、より高い汎化エラーを引き起こす可能性があるということです。このような場合には、事前学習中に早期停止などの何らかの形で正則化を行うことが有益である。事前学習の時間を長くすると、より良い結果が得られ、リターンは減少するが、一般化誤差は増加しないという驚くべき結果が得られる。このことは，CNNをImageNetに適合させることで，一般的で移植性の高い特徴表現が得られることを示唆している．さらに，学習過程はよく振舞われ，早期停止のようなアドホックな正則化を必要としない.

Grandmother Cells and Distributed Codes (祖母細胞と分散記述).

重層ネットワークにおける中間レベルの特徴表現については、まだ十分に理解されていない。最近の特徴の可視化に関する研究(例えば、[13,26]など)では、このようなネットワークは主に "祖母 "細胞から構成されている可能性が示唆されている[1,16]。我々の分析では、中間層での表現はより微妙であることを示している。祖母細胞のような特徴は少数であるが、特徴コードの大部分は分散しており、クラス間で効果的に識別するためには、複数の特徴が協調して発火しなければならない。

Importance of Feature Location and Magnitude (特徴の位置と大きさの重要性).

我々の最後の実験では、特徴の空間的な位置と大きさが画像分類と物体検出においてどのような役割を果たしているかを調査した。直感的には、空間的な位置は物体検出には重要ですが、画像分類にはほとんど重要ではないことがわかりました。さらに驚くべきことに、特徴の大きさはほとんど重要ではないことがわかり

ました。例えば、特徴を2値化しても（0のしきい値で）性能はほとんど低下しません。このことは，大規模な画像検索[8,24]に有用な疎な2値特徴がCNNの表現から "タダで "得られることを示している.

2. Experimental Setup(実験設定)

2.1. Datasets and Tasks(データセットとタスク).

本論文では、いくつかの標準的なデータセットとタスクを用いた実験結果を報告する。

画像分類

画像分類のタスクでは、2つのデータセットを考えています。このデータセットとタスクを「PASCAL-CLS」と呼ぶ。PASCAL-CLSの結果は、標準平均精度(AP)と平均精度(mAP)を用いて報告する。

PASCAL-CLS は，学習用画像が 5k，テスト用画像が 5k，物体クラスが 20 個と，かなり小規模なデータセットである．そこで，ここでは，約108k枚の画像と397個のクラスを持つ中規模のSUNデータセット[25]についても検討します．SUNに関する実験を "SUN-CLS "と呼ぶ。これらの実験では、[25]で提案された10個の標準部分集合ですべての実験を実行することは計算上不可能であったため、非標準的な訓練-テスト分割を使用しています。その代わりに、データセットを無作為に3つの部分 (train、val、test) に分割しました (それぞれ50%、10%、40%)。クラスの分布は3つのセットすべてで一様であった．これらの分割結果は、CNNの特性を調べるためにのみ用いられ、他のシーン分類手法との比較には用いられないことを強調します。SUN-CLSについては、このデータセットの標準的な指標である全クラスの平均の分類精度が397分の1であることを報告しています。選択された実験については、5回の実行における精度の平均±標準偏差として、パフォーマンスのエラーバーを報告します (すべての実験についてエラーバーを計算することは計算上不可能でした)。各実行では、訓練セット、valセット、およびテストセットの異なるランダムな分割が使用された。

物体検出

物体検出のタスクには、PASCAL VOC 2007を使用しています。このデータセットを用いて学習を行い，テストセットを用いてテストを行う．このデータセットとタスクを「PASCAL-DET」と呼ぶ。PASCAL-DETはPASCAL-CLSと同じ画像を使用している。PASCAL-DETの性能は、標準的なAPとmAPを用いて報告する。実験の一部では、PASCAL-DETのバウンディングボックスのみを使用していますが、その場合は "PASCAL-DET-GT "を用いています。

特定の実験のためのより大きな検出訓練セットを提供するために、我々はまた、"PASCAL-DET+DATA "データセットを利用しています。このデータセットは、VOC 2007 trainvalとVOC 2012 trainvalを統合したものとして定義しています。このデータセットには約37k個のラベル付きバウンディングボックスが含まれており、これはPASCAL-DETの約3倍の数である。

2.3. Supervised Pre-training and Fine-Tuning (教師付き事前トレーニングと微調整)

小さなデータセットで大規模なCNNを訓練すると，しばしば壊滅的なオーバーフィッティングを引き起こす．教師付き事前学習のアイデアは，ImageNet分類のようなデータが豊富な補助データセットとタスクを使って，CNNのパラメータを初期化することである．そして，CNNを直接，特徴抽出器として，小さなデータセット上で利用することができる（[4]のように）．あるいは，小さなデータセットで学習を続けることでネットワークを更新することもでき，これは微調整と呼ばれるプロセスである.

微調整については、[7]で述べた手順に従う。まず，CNNの分類層を削除する．これは事前学習タスクに固有のものであり，再利用できない．次に，ランダムに初期化された新しい分類層を追加し，目標タスクの出力ユニット数を設定する．最後に、学習率を0.001（ImageNet分類のためのネットワークのトレーニングに使用される初期学習率の1/10）に設定して、ターゲット損失関数上で確率的勾配降下（SGD）を実行します。この選択は，オーバーフィッティングを制御するために，CNNの初期化を妨げないようにするために行われた．20,000回の微調整を繰り返すごとに，学習率を10倍に減らしている．

3. The Effects of Fine-Tuning and Pre-training on CNNPerformance and Parameters (微調整と事前学習がCNNの性能とパラメータに与える影響)

[7]の結果(R-CNN)では，ImageNet分類のための教師付き事前学習とPASCAL物体検出のための微調整が，事前学習したネットワークの特徴を直接使用する場合(微調整なし)に比べて大きな利益をもたらすことが示されています．しかし、[7]では、微調整の3つの重要な点については調査されていない。

1. 検出データ上でネットワークを「ゼロから」（つまりランダムな初期化から）訓練するとどうなるか？
2. 微調整データの量によって画像はどのように変化するのか？
3. 微調整によってネットワークのパラメータはどのように変化するのか？

本節では、これらの疑問を物体検出と画像分類のデータセットを用いて検討する。

3.1. Effect of Fine-Tuning on CNN Performance (CNNのパフォーマンスに対する微調整の効果)

表1. ゼロから学習したCNN、ImageNet上で事前に学習したCNN、微調整したCNNの性能を比較。
PASCAL-DET+DATAにはVOC 2012 trainvalの追加データが含まれています。
(検出結果にはバウンディングボックス回帰は使用していない)

SUN-CLS			PASCAL-DET			PASCAL-DET+DATA		
scratch	pre-train	fine-tune	scratch	pre-train	fine-tune	scratch	pre-train	fine-tune
40.4 ± 0.2	53.1 ± 0.2	56.8 ± 0.2	40.7	45.5	54.1	52.3	45.5	59.2

本節の主な結果を表1に示す．まず、オープンソースのR-CNNコードを用いて実装した検出実験に注目します。全ての結果はレイヤfc-7の特徴量を使用しています。

少し驚くべきことに、VOC 2007 trainval（13kのバウンディング・ボックス・アノテーション）からの訓練データのみを用いてゼロからCNNを訓練すると、それなりの結果（40.7%のmAP）を得ることができます。しかし、これは事前に学習したネットワークを微調整せずに直接使った場合（45.5%）よりもまだ悪い。さらに驚くべきことに、VOC 2007 trainvalデータにVOC 2012データ（25k個のバウンディングボックスアノテーションを追加）を追加したところ、52.3%のmAPをゼロから達成することができました。この結果は、ImageNetでプレトレーニングを行い、その後VOC 2007 train-valで微調整を行った場合のパフォーマンス（54.1%のmAP）とほぼ同等である。これらの結果は、同じネットワークアーキテクチャに基づく最近の検出システムであるDetectorNet [21]で得られた30.5%のmAPと比較することができる。

次に、PASCAL-DET+DATAの設定でImageNetの事前トレーニングが有用かどうかを尋ねます。ここでは、ゼロからトレーニングを行うことで良好な性能を得ることができるにもかかわらず、事前トレーニングがかなり有効であることがわかります。追加の検出データを用いて微調整した場合の最終的なmAPは59.2%であり，[7]で報告された最良の結果（いずれもバウンディングボックス回帰なし）よりも5%ポイント高い．この結果は，R-CNNの性能はデータが飽和しておらず，他の変更をせずに検出訓練データを追加するだけで，結果が大幅に改善される可能性があることを示唆している．

また、SUN画像の分類についても同様の結果が得られています。ここでも同様の傾向が見られます：ゼロから学習した場合はそれなりの性能が得られますが、ImageNetから初期化した後に微調整を行うと、かなり良い性能が得られます。

3.2. Effect of Fine-Tuning on CNN Parameters (CNNパラメータに対する微調整の効果)

識別的に事前に訓練されたネットワークを微調整することが、タスクのパフォーマンスの面で非常に効果的であることを示す追加の証拠を提供しました。このネットワークの内部を見て、微調整がそのパラメータをどのように変化させるかを見てみましょう。

そのために、フィルタのセットのクラス選択性を測定する方法を定義します。直感的には、ある画像のセットに対してフィルタがある閾値を超えて活性化したときのクラスラベルのエントロピーを使います。この測定値はエントロピーに基づくものなので、 値が低ければフィルタのクラス選択性が高いことを示し、 値が大きければフィルタがクラスに関係なく発光していることを示します。この指標の正確な定義は付録にあります。

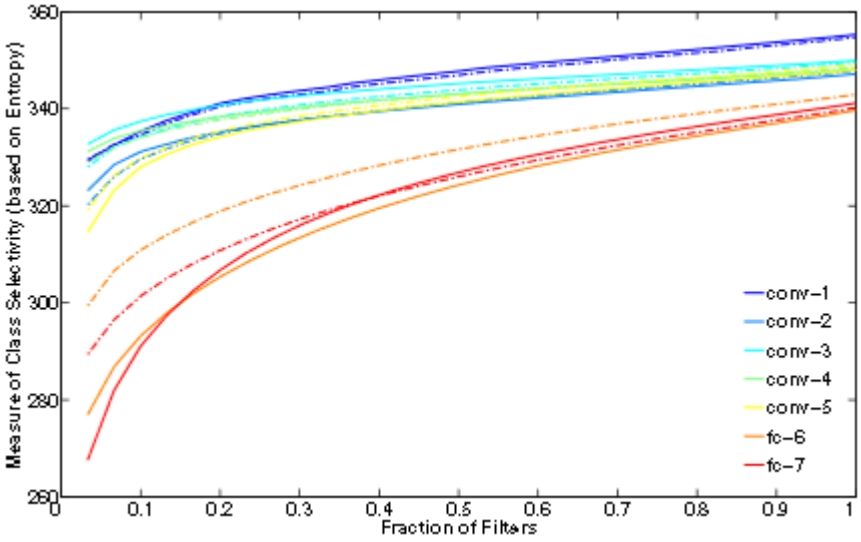

図1: PASCAL オブジェクトのクラス選択性は、各レイヤーについて、微調整前（ダッシュ点線）と微調整後（実線）のフィルタの割合に対してプロットされています。値が低いほどクラス選択性が高いことを示しています。ネットワークの上位に行くほどレイヤーの識別性は高くなりますが、限られたデータでの微調整（PASCAL-DET）は、最後の2つのレイヤー（fc-6とfc-7）にのみ有意に影響を与えます。

フィルタセットのクラス選択性をまとめるために、最も選択性の高いフィルタから最も選択性の低いフィルタまでソートし、ソートされたリストを掃引しながら、最初のフィルタの平均選択性をプロットする。図 1 は、微調整前後のレイヤ 1〜7 のフィルタ群のクラス選択率を示したものである（VOC 2007 trainval 上）。選択率は、フィルタの応答が画像背景との相関ではなく、関心のあるオブジェクトカテゴリの存在による直接的な結果であることを確認するために、画像全体の分類タスクの代わりに、PASCAL-DET-GT のグランドトゥルースボックスを使用して測定されています。

表2: ネットワーク全体を微調整した場合（ft）と、全結合層のみを微調整した場合（fc-ft）の性能比較

SUN-CLS		PASCAL-DET		PASCAL-DET+DATA	
ft	fc-ft	ft	fc-ft	ft	fc-ft
56.8 ± 0.2	56.2 ± 0.1	54.1	53.3	59.2	56.0

図1は、微調整の有無にかかわらず、クラス選択性が層1から層7に向かって増加していることを示しています。興味深いことに、微調整によるエントロピーの変化は、層6と7でのみ有意である。この観察は、微調整データが限られている場合には、6層と7層のみを微調整することで十分な性能が得られる可能性があることを示しています。この仮説をSUN-CLSとPASCAL-DETで検証し、微調整したネットワーク(ft)と、fc-6とfc-7の重みのみを更新して微調整したネットワーク(fc-ft)の性能を比較しました。これらの結果は、表2に示すように、少量のデータでは、微調整は完全に接続された層を「再配線」することに等しいことを示している。しかし、より多くの微調整データが利用可能な場合（PASCAL-DET+DATA）には、すべてのネットワークパラメータを微調整することによる大きな利益がまだ存在する。

3.3.