CNN Features off-the-shelf: an Astounding Baseline for Recognition

Abstract

最近の結果から、畳み込みニューラルネットワークから抽出された一般的な記述子は非常に強力であることが示されている。本論文では，これが実際にそうであることを示す証拠を追加する．我々は，ILSVRC13上で物体分類を行うために訓練されたOverFeatネットワークのコードとモデルを用いて，さまざまな認識タスクに対して行った一連の実験について報告する．我々は，一般的な画像表現としてOverFeatネットワークから抽出された特徴量を用いて，物体画像分類，シーン認識，細粒度認識，属性検出，画像検索の多様な認識タスクに取り組み，多様なデータセットに適用した．これらの課題やデータセットは，OverFeatネットワークが解くために訓練した本来の課題やデータから徐々に遠ざかっていくように選択した．驚くべきことに，さまざまなデータセット上のすべての視覚分類タスクにおいて，高度にチューニングされた最先端のシステムと比較して一貫して優れた結果が得られたことを報告する．例えば，検索では，彫刻のデータセットを除いて，メモリフットプリントの低い手法を一貫して凌駕しています．結果は、ネット上のレイヤーから抽出されたサイズ4096の特徴表現に適用された線形SVM分類器（検索の場合はL2距離）を用いて達成された。この特徴表現は、ジッタリングなどの単純な増強技術を用いてさらに修正されている。この結果は、畳み込みネットを用いた深層学習から得られた特徴が、ほとんどの視覚認識タスクにおいて第一の候補となるべきであることを強く示唆している。

備考

元論文

著者

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Cralsson

掲載

The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 806-813, 2014.

Abstract

最近の結果から、畳み込みニューラルネットワークから抽出された一般的な記述子は非常に強力であることが示されている。本論文では，これが実際にそうであることを示す証拠を追加する．我々は，ILSVRC13上で物体分類を行うために訓練されたOverFeatネットワークのコードとモデルを用いて，さまざまな認識タスクに対して行った一連の実験について報告する．我々は，一般的な画像表現としてOverFeatネットワークから抽出された特徴量を用いて，物体画像分類，シーン認識，細粒度認識，属性検出，画像検索の多様な認識タスクに取り組み，多様なデータセットに適用した．これらの課題やデータセットは，OverFeatネットワークが解くために訓練した本来の課題やデータから徐々に遠ざかっていくように選択した．驚くべきことに，さまざまなデータセット上のすべての視覚分類タスクにおいて，高度にチューニングされた最先端のシステムと比較して一貫して優れた結果が得られたことを報告する．例えば，検索では，彫刻のデータセットを除いて，メモリフットプリントの低い手法を一貫して凌駕しています．結果は、ネット上のレイヤーから抽出されたサイズ4096の特徴表現に適用された線形SVM分類器（検索の場合はL2距離）を用いて達成された。この特徴表現は、ジッタリングなどの単純な増強技術を用いてさらに修正されている。この結果は、畳み込みネットを用いた深層学習から得られた特徴が、ほとんどの視覚認識タスクにおいて第一の候補となるべきであることを強く示唆している。

1. Introduction

"ディープラーニング。コンピュータビジョンの問題にどのくらい効果があると思いますか？" ほとんどの場合、この質問はあなたのグループの喫茶室で提起されています。それに対して誰かが最近の成功例[29, 15, 10]を引用し、他の誰かが懐疑的であることを公言した。あなたは、「残念ながら、自分のネットワークを訓練して素早く答えを見つけるための時間も、GPUプログラミングスキルも、ラベル付きの大量のデータも持っていない」と考えて、少し落ち込んで喫茶室を後にしたかもしれません。しかし，最近，畳み込みニューラルネットワークである OverFeat [38] が公開されたことで，いくつかの実験が可能になりました．特に今、私たちが疑問に思っているのは、与えられたタスクに特化したディープネットワークを訓練できるかどうかではなく、ディープネットワーク（画像分類という特定のタスクを実行するために多様なImageNetデータベース上で注意深く訓練されたもの）によって抽出された特徴が、多種多様な視覚タスクに利用できるかどうかということでした。コンピュータビジョンの研究者であれば、同じ疑問を持ったことがあると思いますので、我々の議論と一般的な知見を関連付けることにしました。

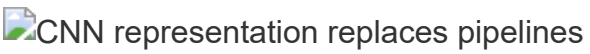

図1.

上) CNN 表現は s.o.a 法のパイプラインに取って代わり，より良い結果を得ることができる．

下）線形SVMを用いた拡張CNN表現は，複数のタスクにおいて一貫してs.o.a.よりも優れている．特化されたCNNとは，そのタスクのためにCNNを特別に設計した他の研究を指す．

教授：
最初に、この問題を調査した人はいますか？

学生：
そうですね、Donahueら[10]やZeilerとFergus[48]、Oquabら[29]が一般的な特徴を示唆しています。Donahueら[10]、Zeiler and Fergusら[48]、Oquabら[29]が大規模なCNNから一般的な特徴を抽出できることを示唆し、この主張を裏付ける初期の証拠を提供している。しかし、彼らは少数の視覚認識タスクしか考慮していない。これらのCNN特徴がどれほど強力なのかをもっと徹底的に調べるのは楽しいことだろう。どのようにして始めればよいのだろうか？

教授：
最も単純な方法は，OverFeatネットワークから画像の特徴ベクトルを抽出し，これを単純な線形分類器と組み合わせることです．特徴ベクトルは，画像を入力として，ネットワークの最終層の1つからの応答だけにすることができます．このアプローチはどのようなビジョンタスクに有効だと思いますか？

学生：
間違いなく画像の分類です。いくつかのビジョングループでは、すでにPascal VOC上の最新の手法と比較して、性能が大幅に向上しています。でも、もしかしたら、ネットワークの微調整が必要だったのかもしれませんね。私はPascal VOCで試してみて、MITシーンデータセットを少しトリッキーにします。

回答：
OverFeat は，微調整をしなくても非常に良い仕事をしてくれます（詳細は 3.2 節を参照）．

教授：
なるほど，この結果は以前の知見を裏付けるもので，それほど驚くべきものではないかもしれません．我々は，OverFeat特徴量に，解くために訓練された問題を解くように求めました．そして、ImageNetは多かれ少なかれPascal VOCのスーパーセットです。室内シーンのデータセットの結果には感心しましたが。では、あまり使い勝手の良くない問題はどうでしょうか

学生：
詳細画像認識（fine-grained classification）は知っています。ここでは，花の種類の違いなど，カテゴリの下位カテゴリを区別したいのですが，もっと一般的なOverFeatureの方がいいのでしょうか？よ

り一般的なOverFeat特徴量は，非常に類似したクラス間の微妙な違いを拾うのに十分な表現力を持っていると思いますか？

回答：
これは、標準的な鳥と花のデータベースでは素晴らしい働きをしてくれました。最も単純な形では、最新の最高性能の手法には勝てませんでしたが、改善の余地が十分にある、よりクリーンなソリューションです。実際には、一連のシンプルなデータ増強技術を採用することで（やはり線形SVMを使用して）、最高のパフォーマンスを発揮する手法に勝てます。感動的ですね！（詳細は3.4節を参照）。

教授：
次の課題は属性検出？OverFeatの特徴が人や物の意味的な性質について何かをエンコードしているかどうかを見てみよう。

学生：
人の境界ボックスから抽出されたグローバルCNN機能は、H3Dデータセットに存在する明瞭さ（アーティキュレーション）と物体の重なり具合（オクルージョン）に対処できると思いますか？最良の方法はすべて、分類前とトレーニング中に何らかのパーツの位置合わせを行います。

回答：
驚くべきことに、CNNの特徴は、平均的にposelets（Poseletとは「任意の姿勢の人物に対してある部分を切り出したもの」。見えの異なる画像を用いてPoselet毎の識別器を学習し、腕や脚や胴体といったパーツの検出/セグメンテーションを行うことが目的となります。）とH3Dデータセット（H3DはHuman in 3Dの略です。）でラベル付けされた人の属性に対して変形可能な部分モデルを叩き出しています。うわー、どうやってそんなことができたんだろう？また，物体属性のデータセットでも非常によく動作します．もしかしたら，これらのOverFeat特徴は本当に属性情報を符号化しているのかもしれません．(詳細は3.5節を参照のこと).

学生：
これらの結果から言えることを教えてください。

教授：
すべては特徴量のおかげです！SIFTやHOG記述子は10年前に大きな性能向上を実現しました。SIFTやHOG記述子は10年前に大きな性能向上をもたらしましたが、今では深い畳み込み特徴が認識にも同様のブレークスルーをもたらしています。したがって、確立されたコンピュータビジョンの手順をCNN表現に適用すれば、報告された結果をさらに押し上げる可能性があります。いずれにしても、認識タスクのために新しいアルゴリズムを開発する場合には、一般的な深層特徴＋単純な分類器という強力なベースラインと比較しなければならない。

2. Background Outline

3.3. Object Detection

残念ながら、オブジェクト検出のタスクにCNNの既製の特徴を使用する実験は行っていません。しかし，Girshickら[15]がPASCAL VOC 2007でCaffeコードの既製の特徴量を用いて顕著な数値を報告していることには言及する価値があります．ここでは、関連する結果を繰り返します既製の機能を使用することで、彼らは46.2のmAPを達成しており、これはすでに約10%も最先端の技術を凌駕しています。このことは、既製のCNN特徴が視覚認識タスクに対していかに強力であるかを証明しています。

最後に、PASCAL VOC 2007のデータセット（もう既製品ではなくなりました）の表現をさらに微調整することで、53.1という印象的な結果が得られました。

3.4. Fine grained Recognition

詳細画像認識(Fine grained recognition)は、商用アプリケーションとカタログアプリケーションの両方に大きな可能性があるため、最近人気を博しています。詳細画像認識は、異なる鳥種、犬の品種、花の種類など、同じオブジェクトクラスのサブクラスの認識を含むため、特に興味深いものです。オックスフォードの花などの細かいアノテーションを備えた多くの新しいデータセットの登場[27] 、Caltech鳥種[45]、犬種[1]、料理活動[37]、猫と犬[32]は、フィールドの急速な発展を助けてきました。 （異なるカテゴリではなく）下位のクラス全体の微妙な違いには、詳細な表現が必要です。この特性により、きめの細かい認識は、一般的な表現がこれらの微妙な詳細をキャプチャできるかどうかの優れたテストになります。

3.4.1. Datasets