How transferable are features in deep neural networks?

備考

著者

Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson

掲載

``How transferable are features in deep neural networks?.'' Procs. Advances in neural information processing systems 27 (NIPS 2014), pp. 3320–3328, 2014.

Abstract

自然画像上で訓練された多くのディープニューラルネットワークには、共通して不思議な現象が見られます：第1層では、ガボールフィルタやカラーブロブに似た特徴を学習します。 このような第1層の特徴は、特定のデータセットやタスクに特化したものではなく、多くのデータセットやタスクに適用できるという点で一般的なもののように見えます。このような特徴は、最終的にはネットワークの最終層で一般的なものから特定のものへと移行しなければならないが、この移行についてはこれまで広く研究されていなかった。この論文では、深層畳み込みニューラルネットワークの各層のニューロンの一般性と特異性を実験的に定量化し、いくつかの驚くべき結果を報告する。伝達可能性は、2つの異なる問題によってネガティブな影響を受ける。 (1)目標タスクの性能を犠牲にして、高次の層のニューロンが元のタスクに特化してしまうこと、(2)共同適応したニューロン間のネットワーク分割に関連した最適化の難しさ、である。本研究では、ImageNet上で学習したネットワークの例を用いて、ネットワークの下部、中部、上部のいずれから特徴を伝達するかによって、これら2つの問題のいずれかが支配的になることを実証した。 また、ベースタスクとターゲットタスクの距離が長くなるほど特徴の伝達性は低下するが、遠くのタスクからでもランダムな特徴を使うよりは、特徴を伝達する方が優れていることを示した。 最後の驚くべき結果は、ほとんど何層からでも転送された特徴を用いてネットワークを初期化することで、ターゲットデータセットを微調整した後でも、一般化を促進することができるということである。

Introduction

最近のディープニューラルネットワークには不思議な現象が見られます：画像上で学習すると、Gaborフィルターやカラーブロブに似た第一層の特徴を学習する傾向があります。これらのフィルタの出現は非常に一般的で、自然な画像データセットでそれ以外のものを取得すると、ハイパーパラメタの選択の誤りやソフトウェアのバグが疑われます。 この現象は，異なるデータセットだけでなく，教師付き画像分類（Krizhevskyet al. 2012），教師なし密度学習（Leeet al. 2009），疎な表現の教師なし学習（Leet al. 2011）など，非常に異なる学習目的でも発生する.

これらの標準的な特徴は、正確なコスト関数や自然画像のデータセットに関係なく、第1層の特徴を見つけることができるように思われるので、これらの **第1層の特徴を一般的なものと呼ぶことにする。** 一方、学習

されたネットワークの最後の層で計算される特徴は、選択されたデータセットとタスクに大きく依存することがわかっています。例えば、教師付き分類の目的に向かってうまく訓練されたN次元ソフトマックス出力層を持つネットワークでは、各出力ユニットは特定のクラスに固有のものになります。このように、**最後の層の特徴を特定のもの** と呼んでいます。これらの一般的な概念と特異的な概念は直感的なものであり、以下ではより厳密な定義を提供します。 第一層の特徴が一般的であり、第二層の特徴が特異的であるならば、ネットワークのどこかで一般的なものから特異的なものへの移行があるに違いありません。この観察は、いくつかの疑問を投げかけています。

- 特定の層が一般的なのか、特定の層が特定的なのか、その程度を定量化できるか。
- 遷移は単一のレイヤーで突然起こるのか、それとも複数のレイヤーに分散して起こるのか？
- この遷移はどこで起こるのか：ネットワークの最初の層、中間層、最後の層の近くか？

なぜなら、ネットワーク内の特徴が一般的なものであれば、転送学習に利用できるからである(Caruana, 1995; Bengio et al., 2011; Bengio, 2011)。伝達学習では、まずベースとなるデータセットとタスクでネットワークを学習し、学習した特徴を第2のターゲット・ネットワークに再利用し、ターゲット・データセットとタスクで学習する。 このプロセスは、特徴が一般的なもの、つまりベースタスクに特化したものではなく、ベースタスクとターゲットタスクの両方に適したものであれば、うまくいく傾向があります。

ターゲット・データセットがベース・データセットよりも著しく小さい場合、転移学習は、オーバーフィッティングなしで大規模なターゲット・ネットワークを訓練することを可能にする強力なツールとなり得る；最近の研究では、この事実を利用して、より高い層から転移する際に最先端の結果を得ている(Donahueet al., 2013a; Zeiler and Fergus, 2013; Sermanetet al., 2014)が、これらの層のニューラル・ネットワークが実際にかなり一般的な特徴を計算していることをまとめて示唆している。 これらの結果は、この一般性の正確な性質と範囲を研究することの重要性をさらに強調している。

通常の転移学習アプローチは、ベースネットワークを訓練し、その最初の n 層をターゲットネットワークの最初の n 層にコピーします。次に、ターゲット・ネットワークの残りの層はランダムに初期化され、ターゲット・タスクに向かって訓練される。 新しいタスクからの誤差をベース（コピーした）特徴量にバックプロパゲーションして**新しいタスクに合わせて微調整する**か、または**転送された特徴量の層を凍結したままにしておく（つまり、新しいタスクでの訓練中は変化しない）** こともできる。ターゲットネットワークの**最初の層を微調整するかどうかの選択は、ターゲットデータセットのサイズと最初の層に含まれるパラメータの数に依存する。** ターゲットデータセットのサイズが小さく，パラメータの数が多い場合，微調整を行うとオーバーフィッティングになる可能性があり，特徴量は凍結されたままになることが多い．一方、ターゲットデータセットが大きく、パラメータ数が少ないため、オーバーフィットが問題にならない場合は、ベースとなる特徴量を新しいタスクに合わせて微調整することで、パフォーマンスを向上させることができる。 もちろん、ターゲットデータセットが非常に大きい場合は、下位レベルのフィルタをターゲットデータセット上でスクラッチから学習するだけなので、移行の必要性はほとんどありません。 以下のセクションでは、微調整された特徴量と凍結された特徴量の2つの技術のそれぞれの結果を比較する。

この論文では、いくつかの貢献をしている。

1. 我々は、特定の層が一般的か特定のものかを定量化する方法、すなわち、その層の特徴があるタスクから別のタスクにどれだけうまく転移するかを定義する（第2節）。次に、ImageNetデータセット上で畳み込みニューラルネットワークのペアを訓練し、一般的な層から特定の層への層ごとの移行を特徴づける（第4節）。

2. 本研究では、微調整を行わずに転送された特徴量を用いた場合の性能低下の原因として、以下の2つの問題があることを実験的に示した。 (i)特徴量自体の特異性。 (ii) 隣り合う層の共適応ニューロン間で

ベースネットワークを分割することによる最適化の難しさ。 我々は、これら2つの効果のそれぞれが、ネットワークの異なる層でどのように支配的になるかを示す。(第4.1節)

3. ベースタスクとターゲットタスクが異なればなるほど、特徴を転移することによる性能上のメリットがどのように減少するかを定量化する (第4.2節)

4. 比較的大規模なImageNetデータセットでは，訓練された重みと比較してランダムな下層重みから計算された特徴量を使用した場合，以前に報告された小規模なデータセット（Jarrettet al., 2009）よりも性能が低下することがわかった．我々は，ランダム重みと転移された重み（凍結と微調整の両方）を比較し，転移された重みの方が性能が良いことを発見した（第4.3節）．

5. 最後に、我々は、ネットワークを初期化する際に、ほとんど何層からでも特徴量を転移して初期化すると、新しいデータセットに微調整した後に、一般化性能が向上することを発見した。これは特に驚くべきことで、最初のデータセットを見たことによる効果は、大規模な微調整を行った後でも持続するからです。(第4.1節)

2. Gnerality vs. Specificity Measured as Transfer Performance

自然画像上で学習されたニューラルネットワークの第1層には、ガボールフィルタやカラーブロブが現れるという不思議な傾向があることを指摘してきた。この研究では、タスクAで学習した特徴のセットの一般性の程度を、その特徴が別のタスクBで使用できる程度と定義しています。これらのサブセットは、互いに類似しているか、または異なるように選択することができる。

タスクAとタスクBを作成するために、1000個のImageNetクラスをランダムに2つのグループに分割し、それぞれ500個のクラスとデータの約半分、約645,000個の例を含むグループを作成しました。これらのネットワークをベースAとベースBと呼び、図1の上2行に示します。次に、{1, 2, ..., 7}の中から層nを選択し、いくつかの新しいネットワークを訓練する。以下の説明および図１では、選択された層の例として、層ｎ = 3 を用いる。まず、以下の２つのネットワークを定義し、訓練する。