How Deeply to Fine-Tune a Convolutional Neural Network: A Case Study Using a Histopathology Dataset

備考

著者

Ibrahem Kandel, Mauro Castelli

Abstract

医用画像の正確な分類は、正しい疾患診断のために非常に重要です。経験豊富な医療スタッフが不足している場合には、セカンドオピニオンを提供したり、より良い分類を行うことができるため、医用画像の分類の自動化は非常に必要です。畳み込みニューラルネットワーク(CNN)は、画像の分類に使用する特徴量を手動で選択する必要がなく、画像分類の領域を改善するために導入されました。 ゼロからCNNを学習するには、医療分野では不足している非常に大規模な注釈付きデータセットを必要とする。CNNの重みを別の大規模な非医療用データセットから転送学習することで、医療用画像の希少性の問題を克服することができる。 転写学習は、新しいデータセットに合わせてCNN層を微調整することからなる。 転移学習を利用する際の主な問題は、ネットワークをどの程度深く微調整するか、そしてそれによって一般化にどのような違いが生じるかということである。本論文では、CNNのブロック単位での微調整の効果を系統的に研究するために、3つの最先端アーキテクチャを用いて2つの組織病理学データセットを用いてすべての実験を行った。結果は、ネットワーク全体を微調整することが必ずしも最良の選択肢とは限らないことを示している。特に浅いネットワークでは、トップブロックを微調整することで、時間と計算パワーの両方を節約し、よりロバストな分類器を生成することができる。

Introduction

医用画像は患者さんの治療に非常に重要な役割を果たしていますが、通常、人手不足や判断に要する時間、セカンドオピニオンの必要性などが大きく影響しています。 病理学のような特定の医療画像分野では、画像を正確かつ迅速に分類することが絶対的なニーズとなっています。組織病理学の画像は、癌のような特定の種類の病気を検出するために、あるいは癌の種類やその程度を判断するために非常に重要です。組織病理学とは、生検で採取した組織サンプルを顕微鏡で検査し、特定の疾患を診断することと定義されています[1]。疾患の発見において非常に重要な役割を果たしており、医師は慎重に治療計画を立てることができます。組織病理検査画像の分類を担当する医師は病理医と呼ばれています。 画像の検査は非常に難しく、長年の経験を持つ病理医が必要とされます。米国では、過去10年間で現役の病理医の数は17.5%減少し、作業量は41%増加している[2]が、組織病理画像を高精度に分類できる自律型分類器を提供することで、病理医の作業を支援することが求められている。

最近、人工知能の分野でブレークスルーされているのが機械学習で、画像の特徴を自動的に抽出するアルゴリズムを開発することができます。使用されるアルゴリズムが複数の隠れ層を持つニューラルネットワークである場合、それは深層学習と呼ばれています。ディープラーニングは、フィードフォワード畳み込みニューラルネットワーク（CNN）[3]を用いて画像を自動的に分類する画像分類の領域に実装することができる。 CNNを画像を分類できるようにすることをトレーニングといい，トレーニングでは，研究対象の画像データセットに合わせてCNNの重みを調整する．CNNの最大のポイントは、CNNが明示的にプログラムされていなくても、画像の重要な特徴をマッピングして画像を分類することができることである。CNNは，糖尿病網膜症の検出と分類 [4,5]，アルツハイマー病の検出 [6,7]，皮膚病変の検出 [8,9]など，多くの医療分野の分類において高い精度で機能することが証明されている．

画像の分類は、画像を事前に定義されたラベルに連続してグループ化することで定義され、医療分野のような多くの分野で非常に重要な役割を果たしています。 ディープラーニングアルゴリズムは、人手を介さずに画像の重要な特徴を検出することができますが、これは、画像のクラスを区別する方法を自ら学習する自律的なアルゴリズムを使用していると考えることができます。クラス間の区別方法を学習する自動分類器を構築しようとした最初の試みは、福島ら[10]やHubel and Wiesel[11]の研究に触発されたLeCun[3]によって紹介され、畳み込みニューラルネットワーク（CNN）と名付けられたが、この試みはデータセットの大きさや当時利用可能な計算能力の問題から制限されたものであった。 2012年にはKrizhevskyら[12]がAlexNetと名付けたCNNアーキテクチャを発表し，ILSVRCコンテスト[13]で1位を獲得したが，2位の25%の誤差率に対して16%であった．CNNは，少なくとも1つの畳み込み層を持つフィードフォワード型の人工ニューラルネットワークと考えられている．CNNの図を図1に示す．

CNNを正確に訓練し、フィードフォワード・ニューラルネットワークに関連する過学習の問題を克服するためには、通常、数十万から数百万枚の画像が必要とされ[14]、CNNの利用は自然画像のような特定の領域に限定されてしまう。転移学習は、CNN訓練における前述の問題を克服するために導入された新しい領域であり、ネットワークの重みをゼロから初期化するのではなく、別の大規模なデータセットで訓練された以前のネットワークの重みを使うことができる。医療分野では画像分類にCNNを

利用することで多くの恩恵を受けることができますが、最大の欠点は訓練に利用できる画像の数が少ないことであり、そこで転移学習を利用することができます。医療画像に転移学習を用いることは，文献[4,6,8]で成功している．Chollet [15]が指摘しているように、転移学習には主に2つの手法があります：元の重みを新しいデータセットに合わせて再調整するファインチューニングと、元の重みを固定し、元のレイヤーを特徴抽出にのみ使用する特徴抽出です。どちらの手法も，Yosinskiら[16]が指摘しているように，データセットの大きさに応じて大きな助けになる．データセットが十分に大きければ元のレイヤーを微調整することができ、データセットが小さければ元のレイヤーを特徴抽出に利用することができる。 また，元のデータセットと対象のデータセットの類似度も非常に重要な役割を果たすことがある． 本研究では，以下の理由から，特徴抽出ではなく微調整を行うことにした．

1. 使用するデータセットに含まれる画像数が多いこと
2. ImageNetデータセットとヒストパソロジーデータセットとの間に大きな差（画像の領域に関して）があること
   などの理由から、特徴抽出ではなく微調整を行うことにました。

CNNアーキテクチャ全体の微調整には何時間もかかり、特定のハードウェアを必要とし、各ブロックの効果は非常に重要な役割を果たします。ネットワーク全体を微調整しても，必ずしも最高の性能が得られるとは限らない． 本研究の主な目的は，CNNをブロック単位で微調整することの効果を決定し，各ブロックを学習することで得られる性能を評価することである． 本研究では3つの学習率を持つ3つの最先端のCNNアーキテクチャを用いた。使用した性能指標はROC曲線のAUCである。 CNNアーキテクチャの性能を評価するために，別のラベルなしのテストセットを用いる．本論文の残りの部分は以下のように構成されている。第2節では、組織病理学における転移学習の利用に関する簡単な文献レビューを行う。第3節では、提案された方法論について議論する。第4節では、得られた結果を述べる。第5節では、所見の概要を述べている。第6章では、結論を述べている。

3.2. Transfer Learning

通常，元データセットは，ImageNetデータセット[32]のように，数千のクラスを持つ数百万の画像を含む．文献では，以下の4つの転移学習手法が紹介されている．

1. 最初の手法は，元のCNNの重み (ImageNetの重みのようなもの) を凍結し，元の完全に接続されたレイヤーを削除して，別の分類器を追加することです．
2. 第二の手法は、底層が非常に汎用的で、どんな種類の画像データセットにも使えるという前提のもと、非常に小さな学習率で元のCNNの上層を微調整し、下層を凍結するというものである[16]。
3. 第三の手法は、元の重みを失わないように非常に小さな学習率を使ってネットワーク全体の重みを微調整し、最後に完全に接続された層を削除し、ターゲットのデータセットに合わせて別の層を追加するというものです。

4. 第四の手法は、重みを一切インポートせずにCNN独自のアーキテクチャを使う、つまりゼロから重みを初期化するというものです。この手法のポイントは、挑戦的なデータセットで実験され、良いことが証明されているよく知られたアーキテクチャを使うことです。異なる転移学習手法を図4に示す。

この図は，異なる転移学習手法を示している．
(a)ImageNetデータセット上で学習された汎用CNN
(b) 最初の手法では，元の重みが固定され，元の分類器レイヤがターゲットデータセットに適合するように新しいレイヤに置き換えられる．
(c) 第2の手法では，最上層を微調整し，最下層の重みを固定し，最後に完全に連結された層を入れ替え，ソースの重みを固定し，元の分類器層をターゲットのデータセットに合わせて新しい層に入れ替える．
(d) 第3の手法では，元の分類器層を配置し，ネットワーク全体を微調整する．
(e)第4の手法では，重みを使わずにオリジナルアーキテクチャを使用する．
緑はImageNetデータセットから学習した重み，青はターゲットデータセットを用いてImageNetの重みを微調整したもの，白は重みをゼロから初期化することを意味する．

3.3. CNN Architectires

AlexNetアーキテクチャは2012年のImageNetチャレンジでエラー率16%で1位を獲得したことから，画像の分類にCNNを用いたアーキテクチャが多く導入された． 2014年にはAlexNetアーキテクチャ[12]よりもかなり深いVGG[24]アーキテクチャが導入され、同年にはAlexNetよりも深く広いとされるGoogLeNetアーキテクチャ[28]も導入された。 その後、2015年にはResNetアーキテクチャ[29]が導入され、より深く、残余接続を含んだものとなり、2016年にはDenseNet[26]が導入されました。 これらの最先端のCNNはすべてImageNetデータセットで学習され、その重みは公開されている。本研究では，VGG16，VGG19，InceptionV3アーキテクチャの3つのCNNを考慮に入れることにした．この選択は、これら3つのネットワークがこのデータセットに関連したKaggle競争で一般的に使われているものであり、さらに重要なことに、他の競争相手よりも優れた性能を発揮していたという事実に関連している。 さらに、考慮すべきアーキテクチャの選択は、微調整が手元の組織病理学データセットを分析するのに適したアプローチであるかどうかを理解することに焦点を当てた研究を展開するための基本的なものではありません。以下、本論文で使用したCNNについて簡単に説明する。

3.3.1. VGG Architectures

VGGアーキテクチャ[24]は、オックスフォードのVisual Geometry Groupによって2014年に導入され、ILSVRCコンペティションに参加し、7.3%というトップ5のエラー率を達成しました。 VGG16とVGG19の2つのネットワークが導入されたが、2つのネットワークの違いは使用されている畳み込み

層の数だけである。VGG16は13の畳み込み層で構成されており、VGG19は16の畳み込み層で構成されているため、VGG16よりも深いと考えられる。著者らは、フィルタサイズの大きい畳み込み層を用いる代わりに、フィルタサイズの小さい2つの層を連結することで、パラメータ数を28%削減した。VGGネットワークは5つの畳み込みブロックで構成されており、最初の2つのブロックはそれぞれ3×3のフィルタサイズを持つ2つの畳み込み層で構成され、最初のブロックの畳み込み層はそれぞれ64個のフィルタを持ち、2番目のブロックの畳み込み層はそれぞれ128個のフィルタを持つ。ＶＧＧ１６の第３のブロックは３つの畳み込み層からなり、ＶＧＧ１９では４つの畳み込み層からなり、全ての層が３×３のサイズの２５６個のフィルタを有している。第４及び第５の畳み込みブロックは、ＶＧＧ１６では３つの畳み込み層からなり、ＶＧＧ１９では４つの畳み込み層からなり、全ての層がサイズ３×３の５１２個のフィルタを有している。5つのブロックは最大プーリング層で区切られている。完全に接続された2つの層は、4096個のニューロンを持つネットワークの分類器として使用されている。VGG16は23層の深さを持つ1億3800万個のパラメータを持ち、VGG19は26層の深さを持つ1億4300万個のパラメータを持つ。ＶＧＧのアーキテクチャを図５に示す。

3.3.2. Inception V3 Architecture

インセプションアーキテクチャ[28]は、2014年にReference[34]の著者らによって初めて導入され、ImageNetコンペティションに参加し、6.65%のトップ5エラー率で第1位を獲得した。これらは、同じオブジェクトの異なるスケールでも、正しく観測するためには異なるフィルタサイズが必要であるという仮説のもとに設計されました。インセプションモジュールは、同じ入力から始まり、それをカーネルサイズの異なる異なる畳み込みレイヤーと、1つの最大プーリングレイヤーに分割します。VGGモデルのように後続ではなく、これらのレイヤーを並列に配置することで、多くのメモリを節約し、深さを増すことなくモデルの容量を増加させることができます。図6にインセプションモジュールを示す。 本論文で使用したバージョンは3番目である。 インセプションアーキテクチャは、順次配置された9つのインセプションモジュールで構成されています。 InceptionV3アーキテクチャは、159層の深さを持つ2380万個のパラメータを持つ。1×1、3×3、5×5の3つのフィルタサイズが1つのインセプションモジュールで使用されているが、更新版では、Reference[24]の影響を受けて、5×5を2つの3×3畳み込みレイヤーに置き換えている。インセプションのアーキテクチャを図７に示す。

3.6.4. Image Augumentation

オーバーフィッティングを減らすための主な方法の1つは、可能な限りすべての画像でモデルを訓練するために膨大な訓練データセットを持つことですが、現実的にはそれは不可能であり、そこで画像拡張が導入されました。そこで、学習データセットを増やすために、画像オーグメンテーションを適用することができます。画像オーギュメンテーションとは、元の画像を変形、回転、明るさの変更などの一連の変更を行うことで、人工的な画像を作成するために使用できるアルゴリズムと定義されています。

4. Results

4.1. Experiment Parameters

本研究では、3つのCNNアーキテクチャをブロック単位で微調整し、各ブロックの微調整がネットワークの一般化可能性に与える影響を調べた。使用したCNNアーキテクチャはVGG16, VGG19, InceptionV3である．元の重みはImageNetデータセットの重みを使用した．使用したデータセットは公開されているPatchCamelyonデータセットであり、データセットサイズは学習用ラベル付き画像22万枚であり、正クラスが60%、負クラスが40%で構成されている。分類器の性能を評価するために、別の57,458枚のラベル化されていない画像を提供し、テストデータセットを分類した結果をKaggleのウェブサイトにアップロードして、ROC曲線のAUCを決定した。実験では、すべての実験において、Keras ライブラリを Python で使用した。提案モデルの模式図を図10に示す。

3つのCNNアーキテクチャをブロック単位で微調整した． バッチサイズは64で，10-3，10-4，10-5の3つの学習率が適用された．Adamオプティマイザ[41]をすべての実験に用いた．すべての画像は、96×96ピクセルの元の寸法に維持された。アーキテクチャの元の完全に接続された層は除去され、ネットワークの精度を高めるために50%の確率でドロップアウト層に置き換えられ、新しい完全に接続された層が分類器として使用された。 学習データセットは2つのパーティションに分割された。 トレーニング用に80%、検証用に20%とした。学習データセットのサイズを大きくし、画像の並進を克服するために、水平・垂直反転、180回転、幅・高さのシフト、剪断・ズームなどのパラメータを用いて、画像増大処理を行った。ここで注目すべきは、画像増大処理はトレーニングデータセットのみに適用され、検証データセットには適用されていないことです。最良のモデルはエポック毎に保存され、10エポックの早期停止基準が適用された。

各CNNアーキテクチャは、その設計に基づいてブロックに分割されている。VGGネットワークでは、最大プール層に基づいてブロックを分割し、InceptionV3ネットワークでは、インセプションモジュールに基づいてブロックを定義した。VGGアーキテクチャでは合計5ブロック、InceptionV3ネットワークでは13ブロックとなっている。つまり、VGGネットワークで最初に微調整されたブロックは5番目、InceptionV3ネットワークで最初に微調整されたブロックは13番目ということになります。

4.2. Experiment results

3つの異なる学習率でVGG16を微調整した結果を表2に示す。これらの結果は、ラベルなしのデータセットを用いたKaggleのウェブサイトからのROC曲線のAUCである。学習したのは最初のブロックが5番目のブロックであり、3つの学習率では10-3学習率が最も高かったが、結果はほぼ同じであっ

た。2 番目のブロックは 5 層目で微調整した 4 番目のブロックであり、最も高い AUC は 10-4 学習率であり、次いで 10-3、10-5 の順であった。4層目を微調整することでもAUCは上昇した。3 回目の実験では、5 番目から 3 番目までのブロックを微調整し、前 2 回の実験よりも高い結果が得られた。最も高い結果が得られたのは、学習率10-4を用いた場合であった。第4回目の実験では、第5回目から第2回目までのブロックの微調整を行い、前2回の実験に比べて精度が低下し始めていることを示した。第5回目の実験では、ネットワーク全体を微調整した結果、前回の実験よりも高い精度が得られたが、全体としては第3回目の実験よりも低い結果となった。全体的には、最後の3ブロックを微調整し、10-4の学習率を利用することで、最高のAUCは96%となりました。

VGG19ネットワークの微調整結果を表3に示します。VGG16アーキテクチャの微調整と同様の手順で行った。表3の結果によると、最後の3ブロックを微調整し、他のブロックを凍結させ、学習率10-4で最も高いAUCが得られていることがわかる。 学習率10-5でも同等の性能が得られているが、学習率10-3を用いた場合には、最後の2層のみを微調整することで最高の性能が得られている。

InceptionV3アーキテクチャを用いて微調整を行った結果を表4に示す。InceptionV3アーキテクチャを用いた最初の実験は、ネットワーク全体をフリーズさせ、最後のブロック（13番目のブロック）を3回、3つの学習率で微調整するというものであった。 結果は3つの学習率でほぼ同じであり、最高の学習率（10-3）で最高、中程度の学習率（10-4）で最低となった。 この手順（最後のブロックを微調整し、残りのブロックを凍結することからなる）を[2,13]の範囲内で反復し、各反復の際に3つの異なる学習率を考慮した。表 4 の結果によると、10-4 と 10-5 の学習率を考慮した場合、アーキテクチャ全体を微調整した場合（つまり n=13）が最も良い性能が得られた。一方、学習率10-3では、最後の8ブロック（ブロック6からブロック13まで）を微調整し、最初の5ブロックをフリーズさせることで最高の性能が得られた。

微調整アプローチの可能性を示すために、特にパッチカメレオンデータセットのためにスクラッチから訓練されたCNNを用いて得られたものと、その性能を比較することにした。

アーキテクチャをゼロから学習させた結果を表5にまとめた。 最初の実験では，3つの異なる学習率を用いて3つのCNNをゼロから学習させた．VGG16アーキテクチャでは、最も高い学習率を用いた場合、我々が課した停止基準に対してモデルは全く収束しなかった。中程度の学習率を用いた場合，モデルは許容可能な性能に収束し，この実験では最高の性能であるが，ネットワークを微調整した結果よりも低かった．学習率が低い場合は満足のいく性能が得られず、中程度の学習率の値を考慮した場合にはネットワークの性能が低下することがわかった。VGG19アーキテクチャについては，学習率が最も高い場合にはVGG16と同様の動作をした． 中程度の学習率を用いた場合の性能は、低学習率を用いた場合の性能よりも低い。 InceptionV3アーキテクチャは、学習率が小さいほど収束し、中程度の学習率を用いた場合に最も良い性能が得られた。全体的には、ネットワークをゼロから学習させても、ネットワークを微調整するよりも良い結果は得られなかった。

4.3. Experiment Results on a Different Histopathology Dataset

Patch Camelyonデータセットで得られた結果を裏付け、我々の知見を強化するために、我々は別の組織病理学のデータセット、すなわちBreakHis [18]データセットを使用して2番目の実験セットを実行しました。BreakHisデータセットは7909枚の画像からなり、良性画像2480枚と悪性画像5429枚に分割されている。 実験を行うために、データセットを80%に分割してモデルを学習、10%に分割して学習中のモデルを検証、10%に分割して未見データでのテストを行った。データセットの大きさの関係で、アーリーストップを50エポックに増やした。VGG16アーキテクチャを用いてBreakHisデータセットを微調整した結果を表6に、VGG19を用いた結果を表7に、最後にInceptionV3を用いた結果を表8に示す。

表6の解析からわかるように、BreakHisデータセットを用いてVGG16を微調整した結果は、PatchCamelyonデータセットで得られた結果と一致している。特に、最上位層の微調整を行い、学習率が10-4と10-5の最小値を用いることで、最も高い結果が得られた。具体的には、学習率10-3では5番目のブロックのみを微調整することで最高の性能が得られ、学習率10-4では5番目から3番目までのブロックを微調整することで最高の性能が得られました。最後に，学習率10-5，5番目から3番目までのブロックの微調整で総合的に最良の性能が得られた.

また、BreakHisデータセットを用いてVGG19を微調整した結果（表7）についても、Patch Camelyonデータセットを用いた解析と同様の挙動が見られます。より詳細には、10-3の学習率で考えた場合、5番目のブロックのみを微調整することで最も良い結果が得られました。 一方、残りの学習率10-4と10-5については、5番目から3番目までのブロックを微調整することで最高の性能を得ることができました。

5. Discussion

そこで，別の非常に大きなデータセットで学習した別のCNNの重みを利用して，別の非常に大きなデータセットで学習したCNNの重みを利用する伝達学習が，医療分野では非常に重要になる． このようにCNNをターゲットのデータセット上で再訓練するプロセスを微調整と呼ぶ。 CNN全体の微調整は非常に時間がかかり、最高の性能を保証するものではない。Yosinskiら[16]が指摘しているように，自然な画像データセットの場合，下層は主にすべての画像データセットに共通する円や辺などのより一般的な特徴を学習し，上層は元のデータセットの非常に特殊な特徴を学習することができる.

本研究では、2つの組織病理学データセットの画像を分類するために、3つの最新のCNNを微調整した場合のブロックごとの効果を、3つの学習率を用いて比較した。この研究の結果、VGGアーキテクチ

ャでは、ネットワーク全体を微調整しても最高の性能は得られず、代わりにトップブロックのみを微調整することで最高の性能が得られることが示された。InceptionV3アーキテクチャでは、ネットワーク全体を微調整することで性能が向上した。主な議論は、ボトム層の一般化可能性に関するもので、ボトム層の重みを凍らせることができるという意味で、あらゆる種類のデータセットに使用することができます。

この結果から、ネットワークの元々の重みを台無しにしないためには学習率を低くする必要があり、また、学習率を非常に低くしても性能は向上せず、学習プロセスが遅くなることがわかりました。実験の結果、学習率を0.0001にすることで、学習時間と性能を完璧に両立できることがわかりました。また、アーキテクチャをゼロから学習させた場合と微調整技術を用いた場合を比較した結果、ゼロから学習させた方が微調整よりも高い結果は得られないことが分かりました。また、学習率を高くすることで、VGGのような浅いアーキテクチャは非常に不安定になり、収束を妨げていました。

References

14. Tajbakhsh, N.; Shin, J.Y.; Gurudu, S.R.; Hurst, R.T.; Kendall, C.B.; Gotway, M.B.; Liang, J. Convolutional
    Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Trans. Med. Imaging 2016,
    35, 1299–1312. CrossRef
15. Chollet, F. Deep Learning with Python, 1st ed.; Manning Publications Co.: Greenwich, CT, USA, 2017.
16. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks?
    In Proceedings of the Advances in Neural Information Processing Systems 27, Montreal, QC, Canada,
    8–13 December 2014; pp. 3320–3328.