Convolutional Neural Networks for Medical ImageAnalysis: Full Training or Fine Tuning?

備考

著者

Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, Jianming Liang

掲載

``Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," IEEE Transactions on Medical Imaging, Vol. 35, No. 5, pp. 1299-1312, 2016.

Abstract

深層畳み込みニューラルネットワーク(CNN)をゼロから訓練するのは、ラベル付けされた大量の訓練データと、適切な収束を確実にするための膨大な専門知識が必要となるため困難である。 有望な代替手段としては、例えば、ラベル付けされた自然画像の大集合を用いて事前に訓練されたCNNを微調整することが挙げられる。 しかし、自然画像と医用画像の間には大きな違いがあるため、そのような知識転移は推奨されないかもしれない。この論文では，医用画像解析の文脈において，次のような中心的な疑問に答えようとしている．**事前に学習したディープCNNを十分に微調整した上で使用することで、ディープCNNを一から学習する必要がなくなるのではないか?** この疑問を解決するために、我々は3つの専門分野（放射線学、心臓病学、消化器病学）の4つの異なる医用画像アプリケーションを考慮し、3つの異なる画像モダリティからの分類、検出、セグメンテーションを行い、**スクラッチから学習したディープCNNの性能が、レイヤーごとに微調整した事前学習CNNと比較してどのようになるかを調査した.** 我々の実験では，
(1)事前に学習したCNNを十分に微調整したものを用いた場合，スクラッチから学習したCNNと同等の性能を発揮するか，最悪の場合には同等の性能を発揮すること，

(2)微調整したCNNの方がスクラッチから学習したCNNよりも学習セットの大きさに対してロバストであることが一貫して示された.
(3) 浅いチューニングも深いチューニングも特定のアプリケーションには最適な選択ではなかった
(4) 我々のレイヤー単位の微調整スキームは，利用可能なデータの量に基づいて，そのアプリケーションに最適な性能を得るための実用的な方法を提供することができた.

1. Introduction

畳み込みニューラルネットワーク（CNN）は，数十年前からコンピュータビジョンの分野で利用されている[1]-[3]． しかし，その真価が明らかにされたのは2012年のImageNetコンペティションが成功してからであり，グラフィックス処理装置（GPU）の効率的な利用，線形整流化された線形単位，新しいドロップアウト正則化，効果的なデータ増強によって革命をもたらしました[3]． 2013年のブレークスルートップ10の1つとして認められているCNN [4] は，再び人気のある学習機械となり，現在ではコンピュータビジョンのコミュニティ内だけでなく，自然言語処理からハイパースペクトル画像処理，医用画像解析に至るまで，さまざまなアプリケーションで利用されている．CNNの主な力は，その深いアーキテクチャ[5]-[8]にあり，複数の抽象度で識別特徴を抽出することができる.

しかし、ディープCNNをゼロから（あるいは完全に）学習するには複雑な問題がないわけではない[9]。 第一に，CNNは大量のラベル付き訓練データを必要とするが，専門家によるアノテーションに費用がかかり，データセットに含まれる疾患（病変など）が少ない医療分野では，この要件を満たすのは難しいかもしれない．第二に，ディープCNNの学習には膨大な計算資源とメモリ資源が必要であり，それがなければ学習プロセスは非常に時間のかかるものになってしまう．第三に、ディープCNNの学習はオーバーフィットと収束の問題によって複雑になることが多く、その解決には、すべての層が同等の速度で学習するように、ネットワークのアーキテクチャや学習パラメータを繰り返し調整する必要がある。したがって、ゼロから深層学習を行うのは面倒で時間がかかり、多くの勤勉さ、忍耐力、専門知識を必要とする。

5. Fine Tuning

式2の反復的重み更新は、ランダムに初期化された重みのセットから始まる。具体的には，訓練段階の開始前に，CNNの各畳み込み層の重みは，平均値が0で標準偏差が小さい正規分布からランダムにサンプリングされた値で初期化される．**しかし，CNNの重みの数が多く，ラベル付きデータの利用可能性が限られていることを考えると，ランダムな重みの初期化から始まる反復的な重み更新は，コスト関数の望ましくない局所的な最小値をもたらす可能性がある．** あるいは，畳み込み層の重みを，同じアーキテクチャの事前学習済みCNNの重みで初期化することもできる． 事前訓練されたネットは，別のアプリケーションからラベル付けされたデータの大規模なセットで生成される．**事前に学習**

した重みのセットからCNNを学習することを微調整と呼び，いくつかのアプリケーションで成功裏に利用されている[10]-[12].

微調整は，事前に学習したネットワークから学習したいネットワークに重みをコピー（転送）することから始まります．例外は，最後の完全接続層であり，そのノード数はデータセットのクラス数に依存する． 一般的には，事前学習したCNNの最後の完全接続層を，新しいターゲット・アプリケーションのクラス数と同じ数のニューロンを持つ新しい完全接続層で置き換えるのが一般的である． 我々の研究では、2クラスと3クラスの分類タスクを扱うので、新しい完全接続層は、研究対象のアプリケーションに応じて、2個または3個のニューロンを持つ。 最後の完全に接続された層の重みが初期化された後、新しいネットワークは層ごとに微調整することができ、最初は最後の層のみを調整し、次にCNNの全層を調整する。

最後の3層が完全に接続された層であるL層を持つCNNを考える。また，α_l をネットワークの l 番目の層の学習率とする． $l \neq L$のとき$\alpha_l = 0$とすることで、ネットワークの最後の（新しい）層だけを微調整することができる。このレベルの微調整は、$L-1$層で生成された特徴量を用いて線形分類器を学習することに相当します。同様に、ネットワークの最後の2つの層は、$l \neq L, L-1$に対して$\alpha_l = 0$を設定することで、微調整することができます。このレベルの微調整は、1つの隠れ層を持つ人工ニューラルネットワークを訓練することに相当し、これは、$L-2$層で生成された特徴を用いて非線形分類器を訓練していると見ることができます。同様に、微調整層L,$L-1$,$L-2$は、本質的には2つの隠れ層を持つ人工ニューラルネットワークを訓練することに相当する。 前の畳み込み層を更新プロセスに含めることで、事前に学習したCNNをさらにアプリケーションに適応させることができますが、オーバーフィットを避けるために、より多くのラベル付き学習データが必要になるかもしれません。

一般的に、CNNの初期の層は低レベルの画像特徴を学習し、ほとんどのビジョンタスクに適用できるが、後期の層は高レベルの特徴を学習する。したがって，通常は，最後の数層を微調整するだけで十分な伝達学習が可能である．**しかし、ソースアプリケーションとターゲットアプリケーションの間の距離が大きい場合には、初期のレイヤーも微調整する必要があるかもしれません。** したがって、効果的な微調整技術は、最後の層から始めて、所望の性能に到達するまで、更新プロセスに段階的により多くの層を含めることである。 我々は、最後の数層の畳み込み層をチューニングすることを「浅いチューニング」と呼び、全ての畳み込み層をチューニングすることを「深いチューニング」と考える。提案する微調整スキームは、ネットワークをそのまま特徴量生成器として利用する[10]や[12]とは異なり、また、ネットワーク全体を一度に微調整する[11]とは異なる点に注意が必要である。

6. APPLICATIONS AND RESULTS

本研究では、3つのイメージングモダリティシステムから4つの異なる医用画像アプリケーションを検討した。自由応答動作特性（FROC）解析によるポリープ検出とPE検出の性能、ROC解析によるフレーム分類の性能、ボックスプロット解析による境界細分化の性能を評価した。統計的な比較を行う

ために、[38]で提案された方法に従って、ROC曲線とFROC曲線の95%信頼区間に対応するエラーバーを計算しました。 エラーバーにより、統計的観点から複数の動作点での性能曲線の各ペアを比較することができます。具体的には、ペアの曲線のエラーバーが一定の偽陽性率で重ならない場合、2つの曲線は所定の操作点で統計的に異なる(p<.05)。 この統計的分析の魅力的な特徴は私達が全体としてカーブを比較するよりもむしろ臨床的に受諾可能な作動ポイントで性能のカーブを比較してもいいことである。 論文全体で統計的比較を議論してきたが、補足資料の中のいくつかの表にさらに要約したもので、補足ファイル/マルチメディアタブにある。

CNNの学習と微調整にはCaffeライブラリ[39]を使用した． 一貫性と比較を容易にするために，研究対象の4つのアプリケーションにはAlexNetアーキテクチャを用いた．各AlexNetの訓練と微調整には，訓練セットのサイズにもよるが，約2-3時間かかった．各CNNの適切な収束を保証するために，受信機の動作特性曲線の下の領域をモニターした．具体的には，各実験において，訓練セットを訓練データの８０%の小さな訓練セットと，残りの２０%の訓練データの検証セットに分割し，検証セットの曲線下面積を計算した． 検証セットで最高の精度が観測されたときに、トレーニングプロセスを終了しました。すべての訓練は、NVIDIA GeForce GTX 980TI (6GBオンボードメモリ)を使用して行われた。完全に訓練されたCNNは、ガウス分布から抽出したランダムな重みで初期化した。[40]や[41]で提案されているような他の初期化手法も実験したが、これらの初期化手法を用いて収束の速度に差があることに気づいたにもかかわらず、収束後の性能の有意な向上は観察されなかった。

フルトレーニングと微調整シナリオの両方について、我々は、ポジティブクラスとネガティブクラスが等しく存在する画像パッチの層別トレーニングセットを使用しました。 この目的のために、少数クラス（ポジティブ）は変更せずに、多数派（ネガティブ）クラスをランダムにダウンサンプリングしました。微調整シナリオでは、Caffeライブラリで提供されている事前学習済みのAlexNetモデルを使用しました。事前学習されたAlexNetは、畳み込み層の約500万個のパラメータと完全に接続された層の約5500万個のパラメータから構成されており、1000の意味クラスでラベル付けされた120万枚の画像を用いて学習されています。 本研究で使用したモデルは、36万回の繰り返し学習を行った後のスナップショットである。 [表1]に示すように、AlexNetは、227x227の入力画像を13x13の特徴マップにマッピングした2組の畳み込み層とプーリング層から始まります。 このアーキテクチャは、次に、９×９カーネルを有する畳み込み層を効率的に実装しながらも、より大きな非線形性を有する３つの畳み込み層のシーケンスで進行する。 畳み込み層のシーケンスは、次に、プーリング層と3つの完全に接続された層が続く。最初の完全に接続された層は、６×６カーネルを有する畳み込み層として見ることができ、他の２つの完全に接続された層は、１×１カーネルを有する畳み込み層として見ることができる。

表I：実験で使用したAlexNetアーキテクチャ。注目すべきは、Cはクラス数であり、内膜-メディアインターフェースのセグメンテーションでは3、大腸内視鏡のフレーム分類、ポリープ検出、肺塞栓症検出では2である。

layer	type	input	kernel	stride	pad	output
data	input	3x227x227	N/A	N/A	N/A	3x227x227
conv1	convolution	3x227x227	11x11	4	0	96x55x55
pool1	max pooling	96x55x55	3x3	2	0	96x27x27
conv2	convolution	96x27x27	5x5	1	2	256x27x27
pool2	max pooling	256x27x27	3x3	2	0	256x13x13
conv3	convolution	256x13x13	3x3	1	1	384x13x13
conv4	convolution	384x13x13	3x3	1	1	384x13x13
conv5	convolution	384x13x13	3x3	1	1	256x13x13
pool5	max pooling	256x13x13	3x3	2	0	256x6x6
fc6	fully connected	256x6x6	6x6	1	0	4096x1
fc7	fully connected	4096x1	1x1	1	0	4096x1
fc8	fully connected	4096x1	1x1	1	0	Cx1

表2：AlexNetの学習と微調整に使用された学習パラメータ μ は運動量，α は各畳み込み層の重みの学習率であり，エポック毎にどのように減少するかを決定する．バイアス項の学習率は，対応する重みの学習率の2倍に設定されています．なお，"fine-tuned AlexNet: layer1-layer2 "は，この2つの層の間とそれを含むすべての層が微調整を受けていることを示している．

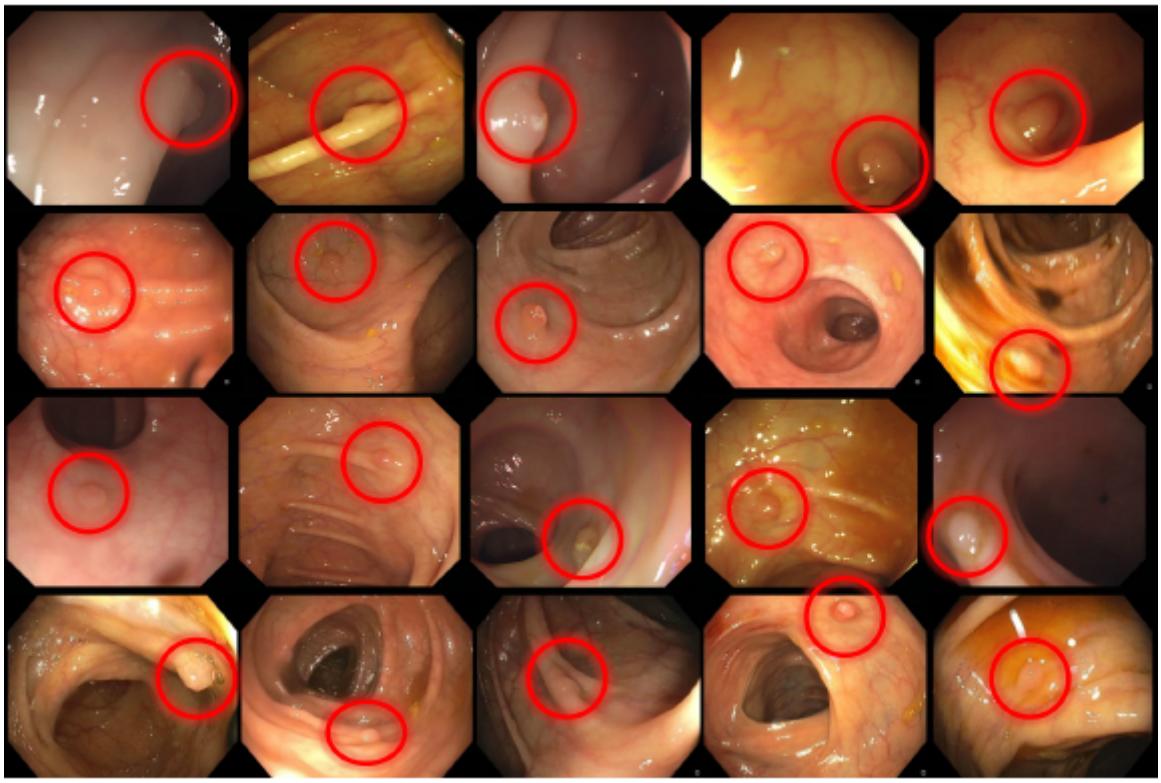
CNNs	Parameters									
	μ	α_{conv1}	α_{conv2}	α_{conv3}	α_{conv4}	α_{conv5}	α_{fc6}	α_{fc7}	α_{fc8}	γ
Fine-tuned AlexNet:conv1-fc8	0.9	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:conv2-fc8	0.9	0	0.001	0.001	0.001	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:conv3-fc8	0.9	0	0	0.001	0.001	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:conv4-fc8	0.9	0	0	0	0.001	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:conv5-fc8	0.9	0	0	0	0	0.001	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:fc6-fc8	0.9	0	0	0	0	0	0.001	0.001	0.01	0.95
Fine-tuned AlexNet:fc7-fc8	0.9	0	0	0	0	0	0	0.001	0.01	0.95
Fine-tuned AlexNet:only fc8	0.9	0	0	0	0	0	0	0	0.01	0.95
AlexNet scratch	0.9	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.95

[表2] は，我々の実験で AlexNet の学習と微調整に使用した学習パラメータをまとめたものである．これらのパラメータは，広範な試行錯誤を経て調整された．探索的実験によると，学習率とスケジューリング率はCNNの収束に大きく影響することがわかった． しかし、学習率が0.001であれば、4つのアプリケーションのすべてで適切な収束が保証された。 学習率を小さくすると収束が遅くなり、学習率を大きくすると収束に失敗することが多い。 また、探索的な実験から、 γ の値は収束の速度に依存することがわかった。高速収束時には、数エポック後に学習率を安全に下げることができ、スケジューリング率を小さくすることができます。 しかし、収束が遅い場合には、比較的大きな学習率を

維持するために、より大きなスケジューリング率が必要となります。4つのアプリケーションすべてにおいて、$\gamma = 0.95$ が妥当な選択であることがわかりました。

A. Polyp detection

図1：大腸内視鏡検査ビデオにおけるポリープの形状と外観のばらつき。



大腸内視鏡検査は、大腸がん検診や予防のために好まれる検査法です。大腸内視鏡検査の目的は、大腸がんの前兆である大腸ポリープを発見し、除去することである。ポリープは、図1に示すように、色、形、大きさに大きな違いがあります。ポリープの出現が困難であることから、しばしば誤検出につながることがあります。

大腸内視鏡検査のビデオからポリープを自動検出するために，いくつかのコンピュータ支援検出（CAD）システムが提案されてきた． 初期のシステム[49]-[51]では，ポリープの色とテクスチャに依存した検出が行われていた． しかし，ポリープ表面のテクスチャの視認性が限られていたことや，ポリープ間の色のばらつきが大きかったことが，このようなシステムの適用性の妨げとなっていた．最近のシステム[52]-[56]では、時間情報と形状情報を利用してポリープの検出を強化している。この点では、色やテクスチャよりも形状特徴の方が効果的であることが証明されていますが、これらの特徴は、ポリープが発見された状況を考慮しなければ、誤解を招く可能性があります。 我々は、[42]を頂

点とする以前の研究[57]-[59]で、ポリープの形状のみに基づくアプローチの限界を克服することを試みた。具体的には、ポリープ境界付近の形状と文脈情報を組み合わせるための手作りのアプローチを提案し、このアプローチが他の最先端の手法よりも優れていることを実証した。

トレーニングと評価のために、40個の短い大腸内視鏡ビデオのデータベースを使用した。 **我々のデータベースの各大腸内視鏡検査フレームには、2値の基底真実画像が含まれている。 大腸内視鏡ビデオを、ポリープのある3,800フレームとポリープのない15,100フレームを含むトレーニングセットと、ポリープのある5,700フレームとポリープのない13,200フレームを含むテストセットにランダムに分割した。** 我々は、対応するバウンディングボックスを持つポリプ候補のセットを得るために、トレーニングフレームとテストフレームに我々の手作りのアプローチ[42]を適用しました。各候補の位置で、利用可能なバウンディングボックスが与えられた場合、データ拡張を用いて画像パッチのセットを抽出した。**具体的には、各候補について、対応するバウンディングボックスを1.0x、1.2x、1.5xのファクタで拡大し、3つのスケールでパッチを抽出した。 各スケールでは、候補の位置をリサイズしたバウンディングボックスを水平方向と垂直方向に10%ずつ移動させてからパッチを抽出した。さらに、得られた各パッチを水平・垂直方向のミラーリングと反転で8回回転させました。そして、基礎となる候補がポリープの基底真理値の内側にある場合には、パッチを正とラベル付けし、そうでない場合には、その候補を負とラベル付けしました。負のパッチが比較的多いので、CNNの訓練と微調整のために10万個の訓練パッチの層別セットを収集した。 テスト段階では、ポリープ候補から抽出されたすべてのテストパッチを訓練されたCNNに供給した。** その後、候補レベルでのテストパッチの確率的出力を平均化し、性能評価のためにFROC分析を行った。

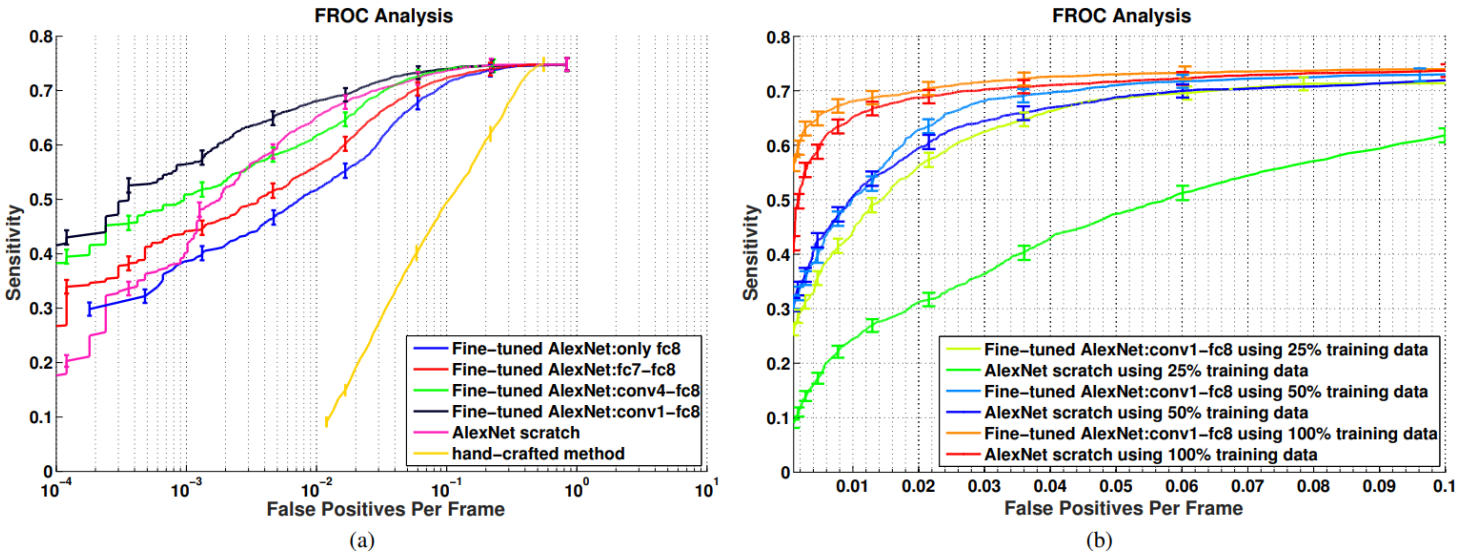


図2: ポリープ検出のためのFROC分析。(a) インクリメンタル・ファインチューニング，スクラッチからのトレーニング，ハンドクラフト・アプローチ[42]の比較。(b) 訓練データの減少が，スクラッチから訓練したCNNと深く微調整したCNNの性能に及ぼす影響.

図2(a)は、我々の手作りのアプローチのFROC曲線を、微調整されたCNNとスクラッチから訓練されたCNNのFROC曲線と比較したものである。図の乱雑さを避けるために、代表的なFROC曲線のサブセットのみを示している。3つの動作点における各組のFROC曲線間の統計的比較も表[S1]に示す。**ハンドクラフトアプローチは、すべてのCNNベースのシナリオに比べて有意に性能が優れている（p<.05）。 この結果は、我々のハンドクラフトアプローチが偽陽性候補を除去するために幾何学的情報のみを使用したためであろう。 微調整については，AlexNetの最後の層のみを大腸内視鏡検査データで更新した場合に最も低い性能が得られた． しかし、最後の2層（FT: fc7-fc8）を微調整した場合は、1層のみ微調整した事前学習済みのAlexNet（FT: fc8のみ）と比較して、ほぼすべての操作ポイントで有意に高い感度（p<.05）を達成しました。また，より多くの畳み込み層を微調整プロセスに含めると，性能の漸増的な向上が見られた．** 具体的には、浅い微調整層(FT:fc7-fc8)を持つ事前学習CNNは、ほとんどの動作点において、中程度の微調整層(FT:conv 5,4,3-fc8)を持つ事前学習CNNよりも有意にアウトパフォームされていた。さらに，深いチューニングを施したCNN (FT: conv 1,2-fc8)は，特に偽陽性率が低い場合には，中程度の微調整を施した事前学習CNNよりも有意に高い感度を達成した．また、図2(a)に見られるように、最後の数層の畳み込み層を微調整することで、低偽陽性率の設定では、ゼロから学習したAlexNetモデルを上回る性能を示した。

B. Pulmonary embolism detection

PEは、下肢から肺に移動する血栓で、肺動脈の閉塞を引き起こします。未治療のPEの死亡率は30%に達することもありますが[61]、早期診断と適切な治療により、死亡率は2%と低くなります[62]。CT肺動脈造影（CTPA）は、PE診断の主要な手段であり、放射線科医はPEが疑われる場合、肺動脈の各枝を慎重に追跡する。CTPAの読影は、時間のかかる作業であり、その精度は、注意力やPEの視覚的特徴に対する感度などの人的要因に左右されます。CADは、PEの診断を改善し、CTPAデータセットの読影時間を短縮する上で大きな役割を果たします。

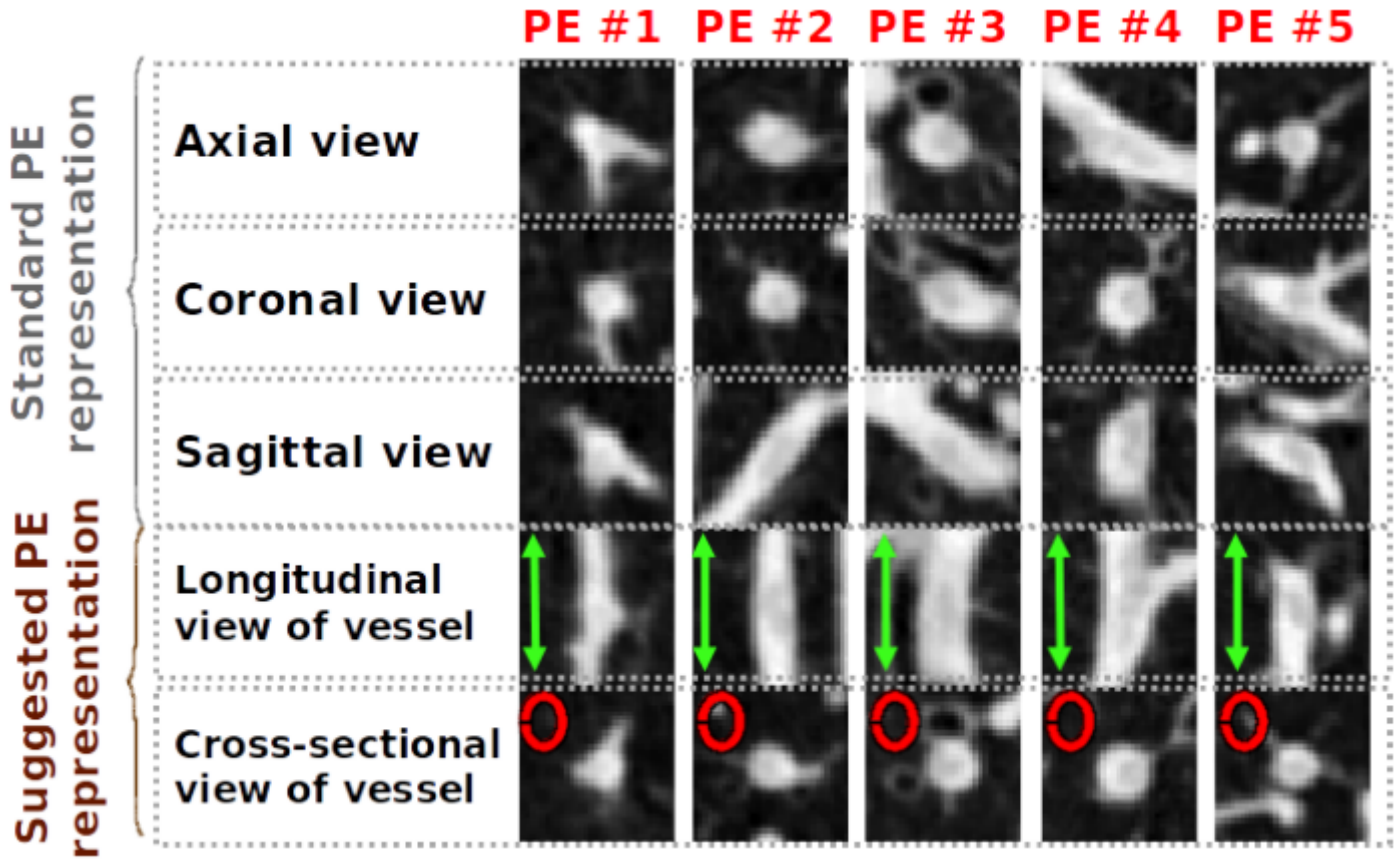


図3: 標準的な3チャンネル表現と，我々が提案する2チャンネル表現における5つの異なるPE。我々の表現では，PEがより一貫して現れている．ここで紹介する実験では，より高い分類精度を達成し，収束性を向上させることができる我々のPE表現を使用している．

我々は，以前の研究[60]で生成したPE候補と，最近発表した研究[18]でPEに対して提案した画像表現に基づいて実験を行った．我々の候補生成法は，明るい背景に囲まれた暗い領域として塞栓を見つけることを目的としたtobogganingアルゴリズム[63]の改良版である．我々の画像表現では，血管の断面および縦断方向のPEを捉える2チャンネルの画像パッチが一貫して得られる（図3参照）．このユニークな表現により、PEの外観のばらつきが劇的に減少し、より精度の高いCNNを学習することができる。ただし，AlexNetアーキテクチャはカラー画像を入力として受け取るため，2チャンネル画像パッチをカラーパッチに変換する必要がある．このため、単純に2チャンネル目を繰り返して、3チャンネルのRGBライクな画像パッチを作成した。このパッチを用いて，AlexNetの学習と微調整を行いました．性能比較のために，我々は手作業によるアプローチ[60]を用いた．手作りのアプローチは、同じ候補生成方法[60]を利用していますが、PEの特徴付けには、血管ベースの特徴に加え、Haralick[64]とウェーブレットベースの特徴を使用し、最後に候補の分類にはマルチインスタンス分類器を使用しています。

実験には，121個のCTPAデータセットからなるデータベースを使用し，合計326個のPEを得た．まず、tobogganingアルゴリズムを適用して、PE候補の粗いセットを得た。この適用により、6,255個のPE候補が得られ、そのうち5,568個が偽陽性、687個が真陽性であった。tobogganingアルゴリズムで

は、同じPEに対して複数の候補を投じることができるため、真の陽性者の数はPEの数よりもはるかに多かった。収集した候補を患者レベルで、真陽性434件（ユニークPE 199件）、偽陽性3,406件のトレーニングセットと、真陽性253件（ユニークPE 127件）、偽陽性2,162件のテストセットに分けた。CNNの学習には，物理的なサイズが異なる3種類のパッチ（幅10mm，15mm，20mm）を使用した．さらに，パッチの物理的な大きさの20%まで，血管の方向に沿って各候補地を3回変換した．さらに，血管の縦断面と横断面を血管軸を中心に回転させることで，トレーニングデータセットを拡張し，各スケールと移動について5つのバリエーションを追加した．CNNのトレーニングと微調整のために，81,000個の画像パッチからなる層別のトレーニングセットを形成した．テストでは，各テスト候補に対して同じデータ拡張を行い，各PE候補のデータ拡張されたパッチに対して生成された確率的スコアを平均することで，全体のPE確率を計算した．

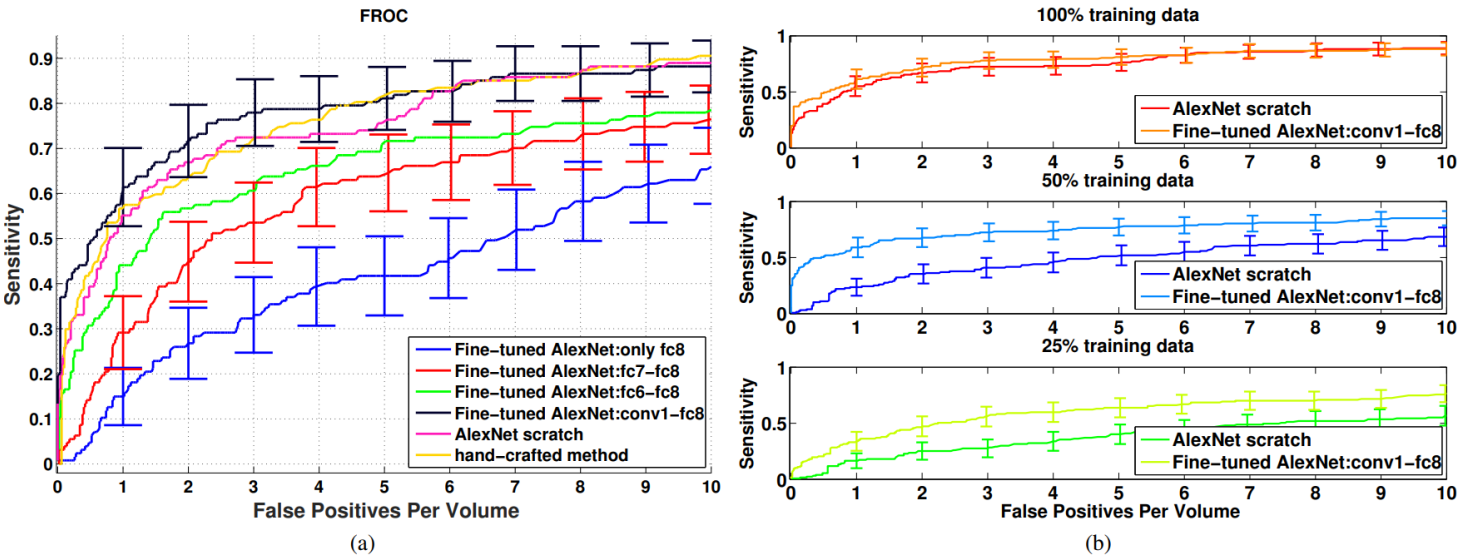


図4: 肺塞栓症検出のためのFROC分析。(a) 段階的な微調整，ゼロからのトレーニング，および手作りのアプローチ [60]の比較．図の乱れを防ぐため，エラーバーは一部のプロットのみに表示されている．より詳細な分析結果は，表S2に示されている．(b) 訓練データの減少が，スクラッチから訓練したCNNとディープ・ファイン・チューニングしたCNNの性能に及ぼす影響．

評価のために、テストPE候補に対して生成された確率的スコアの閾値を変更してFROC分析を行った。図4(a)は、ハンドクラフト手法、ゼロから学習したディープCNN、およびレイヤーごとに微調整された代表的な事前学習済みCNNのサブセットのFROC曲線を示している。さらに、各組のFROC曲線間の統計的な比較を表S2にまとめた。示したように，**2つの層を微調整したCNN (FT:fc7-fc8) は，1つの層だけを微調整したCNN (FT:only fc8) よりも，有意に高い感度を達成した (p<0.05).** この感度向上は、ほとんどの動作点で認められました。しかし、新しい層を追加して微調整しても、わずかな性能向上しか得られなかった。しかし、そのような微調整の積み重ねによって、**深く微調整したCNNと1、2、3層を微調整したCNNとの間にはかなりの差が生じた。** 特に、深くチューニングした

CNN（FT:conv1-fc8）は、図4（a）に示した動作点の大部分で、2層の微調整をしたCNN（FT:fc7-fc8）よりも有意に高い感度（$p<0.05$）を示した。また、1つのボリュームにつき3つの誤検出があった場合、ディープ・ファイン・チューニングを施したCNNは、3つのファイン・チューニングを施した層を持つ事前学習済みのCNN（FT:fc7-fc8）よりも有意に高い感度（$p<0.05$）を達成した。図4(a)から、深く微調整されたCNNは、手作りのアプローチに比べて有意ではない性能向上をもたらしたことも明らかである。これはおそらく、ハンドクラフト手法が、特定のタイプの誤検出を取り除くために基礎的な特徴が特別かつ段階的に設計された正確なシステムだからでしょう。しかし、エンド・ツー・エンドの学習マシンが、最小限のエンジニアリング努力で、このような洗練された特徴のセットを学習できることは興味深いことです。**図4(a)から、深く調整されたCNNは、ゼロから学習したCNNと同等の性能を発揮することもわかりました。**

C. Colonoscopy frame classification

画質評価は、大腸内視鏡検査の客観的な品質評価に大きな役割を果たします。一般的に、大腸内視鏡検査の映像には、大腸の可視性が低く、大腸の検査や治療行為を行うのに適していない、情報量の少ない画像が多く含まれています。映像中の非情報画像の割合が大きいほど、大腸の可視化の質が低くなり、その結果、大腸内視鏡検査の質が低下する。したがって、大腸内視鏡検査の品質を測定する1つの方法は、撮影された画像の品質を監視することである。このような品質評価は、低品質の検査を制限するために実技中に使用したり、品質監視を目的として後処理の設定で使用することができます。

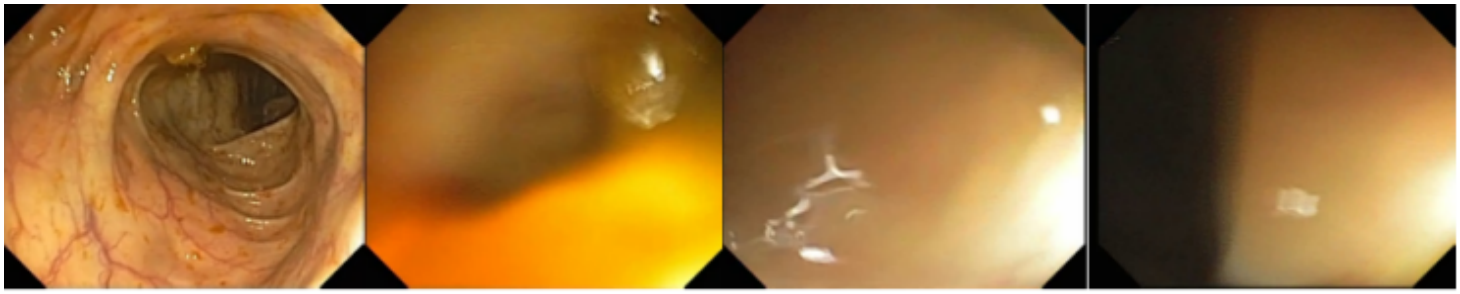


図5: (a) 情報量の多い大腸内視鏡フレーム。(b,c,d) 情報量の少ない大腸内視鏡画像の例。情報量の少ないフレームは、通常、スコープの急速な動きや壁との接触の際に撮影される。

大腸内視鏡検査の画質評価は，技術的には，入力画像を「情報あり」と「情報なし」に分類する画像分類タスクと捉えることができます．図5は，情報量の少ない大腸内視鏡フレームと情報量の多い大腸内視鏡フレームの例である．以前の研究[65]では，画像再構成誤差からプールされた局所特徴量と

大域特徴量に基づいて，手作りのアプローチを提案した．その結果，手作りの手法は，他の主要な手法[66]，[67]に比べて，大腸内視鏡検査映像の品質評価に優れていることがわかった．今回の取り組みでは，慎重に作られた手法の代わりに，ディープCNNの使用を検討した．具体的には，手作りの手法と，ゼロから学習したディープCNN，およびラベル付きの大腸内視鏡フレームを用いてレイヤーごとに微調整した事前学習済みのCNNの性能を比較した．

実験では，6つの完全な大腸内視鏡映像を使用しました．すべてのビデオフレームにアノテーションを施すにはコストがかかるため，各ビデオの5秒ごとに1フレームを選択して各大腸内視鏡ビデオをサンプリングし，類似の大腸内視鏡フレームを多数削除した．その結果，4,000枚の大腸内視鏡画像からなるバランスのとれたデータセットが得られた．このデータセットでは，情報量の多いクラスと少ないクラスが均等に表現されている．次に，訓練を受けた専門家が，収集した画像に手動で「有益」または「非有益」のラベルを付けました．さらに，消化器内科医がラベル付けされた画像を確認し，修正を加えた．ビデオレベルでラベル付けされたフレームをトレーニングセットとテストセットに分け、それぞれに約2,000個の大腸内視鏡フレームを入れた。データ増強のために，500×350の大腸内視鏡フレームのランダムな位置から227×227ピクセルのサブ画像を200枚抽出し，約40,000枚のサブ画像を含む層別トレーニングセットを作成した．テスト段階では，各フレームが有益である確率は，ランダムに切り取られた部分画像に割り当てられた平均確率として計算された．

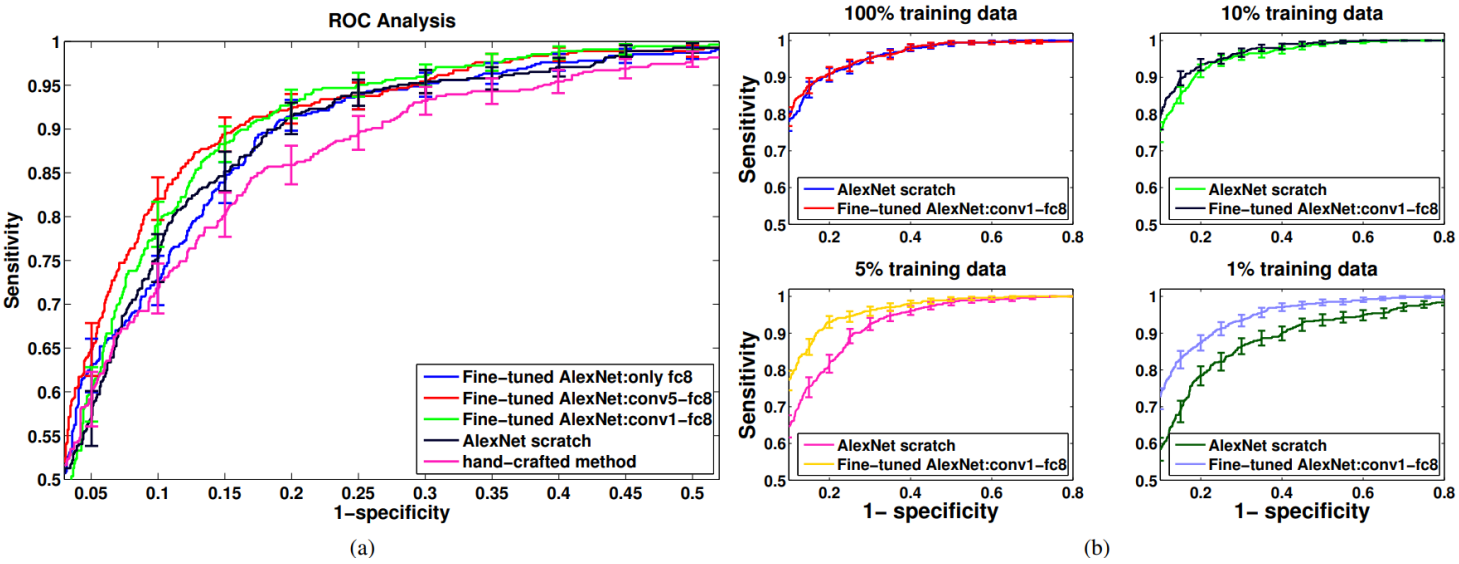


図6: 画質評価のためのROC分析。(a) インクリメンタル・ファインチューニング，スクラッチからのトレーニング，ハンドクラフト・アプローチの比較 [65]．(b) 訓練データの減少が，スクラッチから訓練した畳み込みニューラルネットワーク（CNN）と深く微調整したCNNの性能に及ぼす影響．

CNNベースのシナリオとハンドクラフトのアプローチの性能比較にはROC分析を用いた。その結果を図6(a)に示します。図の乱れを防ぐために、代表的なROC曲線のサブセットのみを示しています。

しかし、誤検出率が10%、15%、20%のときのすべてのROC曲線の統計的な比較を表S3にまとめている。すべてのCNNベースのシナリオは、上記の3つの動作点のうち少なくとも1つにおいて、手作りのアプローチを有意に上回ることが確認された。また、**事前に学習したCNNをネットワークの途中で微調整する方法（FT:conv4-fc8およびFT:conv5-fc8）は、浅く微調整する方法よりも有意に優れているだけでなく、深く微調整したCNN（FT:conv1-fc8）よりも10%および15%の偽陽性率で優れていることが確認された。** これは、CNNの初期層で学習したカーネルが画質評価に適していたため、微調整が不要だったためと考えられる。さらに、**スクラッチから学習したCNNは、浅い微調整（FT:only fc8）では事前学習したCNNを上回ったが、中程度の微調整（FT:conv5-fc8）では事前学習したCNNを上回った。** したがって、ファインチューニングスキームは、ゼロからのフルトレーニングスキームよりも優れていました。

トレーニングデータのサイズによってCNNの性能がどのように変わるかを調べるために，**トレーニングサンプルの数を1／10，1／20，1／100倍に減らした．** 他のアプリケーションと比較すると、訓練データセットのサイズを適度に減らしてもCNNの性能に大きな影響はないので、さらに減らすことを検討した。図6（b）に示すように、深く微調整されたCNNも完全に訓練されたCNNも、元の訓練セットの10%を使用しても性能の低下は顕著ではなかった。しかし、訓練セットのサイズをさらに小さくすると、完全に訓練されたCNNの性能は大幅に低下し、ディープ・ファイン・チューニングされたCNNの性能もほぼ低下した。**限られた学習セットでも、深く微調整されたCNNが比較的高い性能を示したことは、ImageNetから学習したカーネルが大腸内視鏡のフレーム分類に有用であることを示している。**

7. Discussion

本研究では、本研究で得られた知見の汎用性を確保するために、3種類の画像モダリティシステムから4つの一般的な医用画像問題を検討した。具体的には、3次元画像におけるコンピュータ支援病変検出の代表としてPE検出、2次元画像におけるコンピュータ支援病変検出の代表としてポリープ検出、機械学習を用いた医用画像分割の代表として内膜境界分割、医用画像分類の代表として大腸内視鏡検査画像の品質評価を選んだ。これらのアプリケーションは、異なる画像スケールで問題を解決する必要がある。例えば、内膜境界のセグメンテーションやPEの検出では、画像内の小さな領域の検査が必要となるが、ポリープの検出やフレームの分類では、はるかに大きな受容領域が必要となる。したがって、我々は、選択されたアプリケーションは、医療画像の分野に関連する様々なアプリケーションを包含していると信じています。

医用画像解析の文脈において、一からディープCNNを訓練する代わりに、微調整されたCNNの可能性を徹底的に調査した。**大規模なトレーニングセットと縮小されたトレーニングセットの両方を用いて解析を行った。完全なデータセットを用いた場合，事前に学習したCNNを浅くチューニングした場合には，スクラッチから学習したCNNよりも劣る性能になることが多かったが，より深くチューニングした場合には，スクラッチから学習したCNNに匹敵する，あるいはそれ以上の性能を得ることができ**

た. 深く微調整したCNNとスクラッチから学習したCNNとの性能の差は，学習セットのサイズを小さくすると拡大し，利用可能な学習セットのサイズにかかわらず，常に微調整したCNNが望ましい選択肢であると結論づけた.

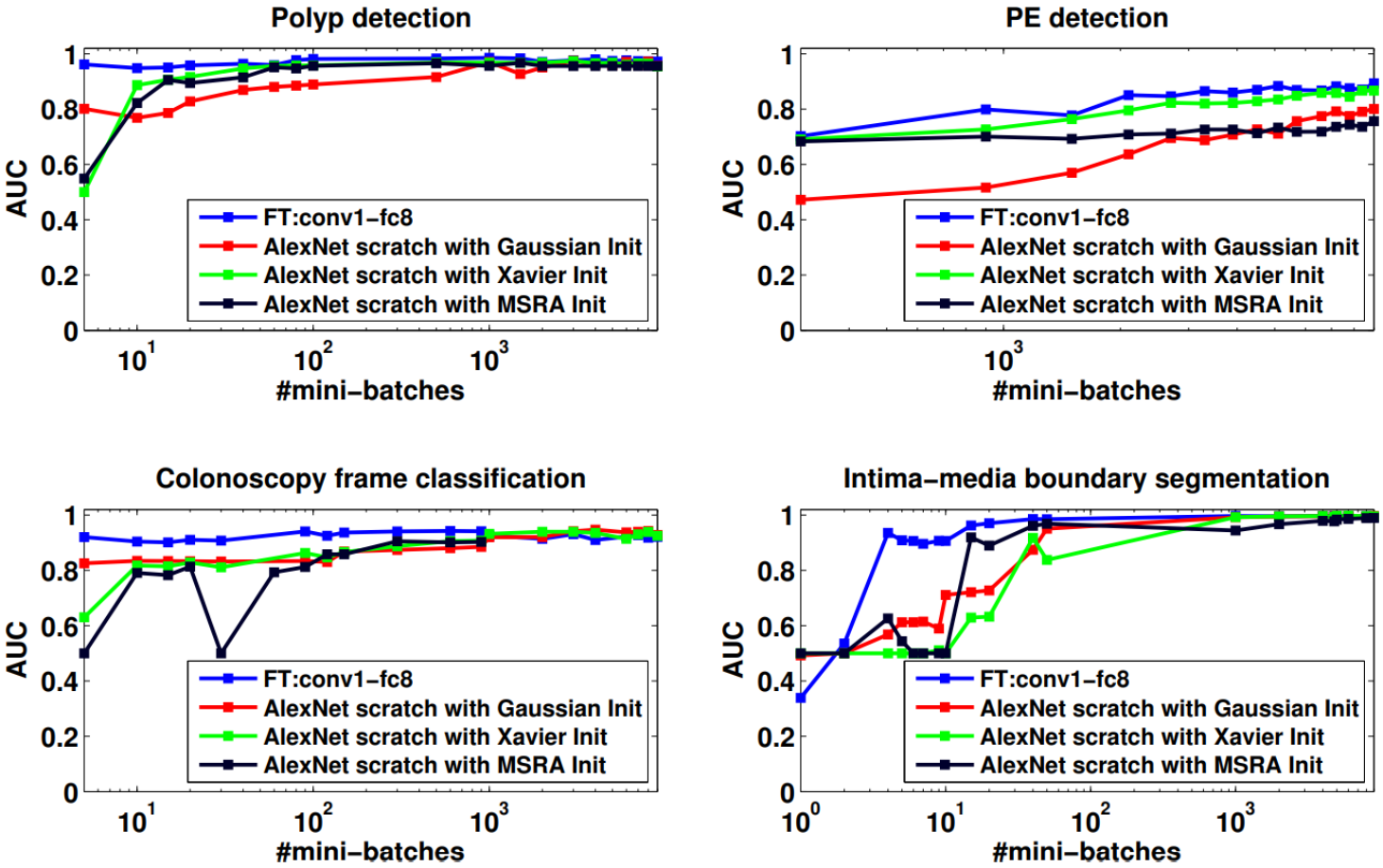


図10: 深く微調整したCNNと、3種類の初期化手法でゼロから学習したCNNの収束速度。

微調整されたCNNのもう一つの利点は、収束の速さである。この利点を実証するために、図10では、深く微調整されたCNNとスクラッチから学習されたCNNの収束速度を比較している。 完全に比較するために，完全に訓練されたCNNの重みを初期化するために3つの異なる手法を用いた． 1) [40]で提案された一般的に知られているXavierと呼ばれる手法，2) [41]で提案されたMSRAと呼ばれるXavierの改訂版，そしてガウス分布に基づく基本的なランダム初期化手法である．今回の解析では、収束の指標として検証データのAUCを計算した。 具体的には、モデルの各スナップショットをバリデーションセットのパッチに適用し、ROC分析を用いて分類性能を評価した。 私たちは、intimia-media境界セグメンテーションの質問に対して3クラスの分類問題を扱ったので、2つのインターフェースクラスを1つの正のクラスにマージして、結果として得られた2値分類（インターフェース対バックグラウンド）のAUCを計算しました。 示されているように，**微調整されたCNNはすぐに最大性能に達するが，ゼロから学習されたCNNは最高性能に達するまでに長い学習時間が必要である． さらに、異な**

る初期化手法を用いることで、収束の傾向は異なるが、完全な収束後にはPE検出を除いて有意な性能向上は見られなかった。

我々は、高精度な画像分類器を実現するためには、微調整の深さが重要であることを明らかにした。
コンピュータビジョン分野の多くのアプリケーションでは、浅いチューニングや最後の数層の畳み込み層を更新するだけで十分な性能が得られるが、医療画像アプリケーションでは、より深いレベルのチューニングが不可欠であることを発見した。 例えば，特にポリープの検出や内膜境界のセグメンテーションでは，深くチューニングされたCNNを用いた場合に顕著な性能向上が見られたが，これは，これらのアプリケーションと事前に学習されたCNNが構築されたデータベースとの間に大きな違いがあるためであろう． しかし、大腸内視鏡のフレーム分類では、これほど大きな性能向上は見られなかったが、これはImageNetと我々のデータベースの大腸内視鏡フレームの相対的な類似性に起因するものである。具体的には、どちらのデータベースも同様の低レベルの画像情報を持つ高解像度画像を使用しているため、アプリケーション固有の特徴を持つ後期畳み込み層を微調整することで、大腸内視鏡フレーム分類の高レベルの性能を達成するのに十分である。

我々の実験はAlexNetアーキテクチャに基づいて行われたが，これはCaffeライブラリで事前に学習されたAlexNetモデルが利用可能であったことと，このアーキテクチャが十分に深く，微調整の深さが事前に学習されたCNNの性能に与える影響を調べることができたからである． あるいは、VGGNetやGoogleNetのようなより深いアーキテクチャを使うことも可能であった。より深いアーキテクチャは最近、困難なコンピュータビジョンタスクに対して比較的高い性能を示しているが、医用画像アプリケーションでより深いアーキテクチャを使用しても大きな性能向上は期待できない。我々は、この作業の目的が、多くの異なる医用画像処理タスクに対して最高の性能を達成することではなく、ゼロからの訓練スキームと比較して微調整の能力を検討することであったことを強調している。これらの目的のために、AlexNetは合理的なアーキテクチャの選択です。

異なるモデルやアプリケーションについて報告された性能曲線は，それぞれの実験で達成できた最高のものではないかもしれないことを認めたい． この最適でない性能は、モデルの収束速度や最終的な精度に影響を与えるCNNのハイパーパラメータの選択に関係している。これらのパラメータの動作値を見つけようとしたが，論文で研究したCNNの数が多いことと，ハイエンドGPUでも各CNNの学習に時間がかかることを考えると，最適な値を見つけることは不可能であった． とはいえ，比較に用いたCNNの大部分は事前に学習したモデルであり，ゼロから学習したCNNよりもハイパーパラメータの選択の影響が少ないかもしれないので，この問題は全体的な結論を変えるものではないかもしれない．

本研究では、スペースの都合上、すべての医用画像モダリティをカバーすることはできませんでした。例えば、ゼロからCNNを完全に訓練することで有望な性能を示したMR画像や病理組織画像の微調整性能については研究していない。しかし，自然画像からCT，超音波，内視鏡などへの知識転移に成功していることを考えると，微調整は他の医療分野でも成功するのではないかと推測される．さらに、本研究では、事前に訓練された教師付きモデルの微調整に焦点を当てている。しかし、1000の意味クラスからの数百万のラベル付き画像を持つImageNetデータベースが利用可能であるため、事前学

習済み教師付きモデルの使用が微調整のための自然な選択となるかもしれないが、制限付きボルツマンマシン（RBM）や畳み込みRBM [74]によって得られたような事前学習済み教師なしモデルも考慮することができた。それでも、教師なしモデルは、ラベル付けされた1次元信号の大規模なデータベースがないため、1次元信号処理に有用であることに変わりはない。 例えば、教師なしモデルの微調整は、[75]では音響音声認識に、[76]では脳波記録におけるてんかんの検出に使用されている。

8. Conclusion

本論文では、医用画像解析の文脈において、次のような中心的な問題に取り組むことを目的とした。事前に訓練されたディープCNNを十分に微調整した上で使うことで、ディープCNNを一から訓練する必要がなくなるのではないか？3つの異なる画像モダリティシステムからの4つの異なる医用画像アプリケーションに基づいた我々の広範な実験により、微調整されたCNNが医用画像解析に有用であり、完全に訓練されたCNNと同等の性能を発揮し、訓練データが限られている場合には後者よりも優れた性能を発揮することが実証された。 この結果は、自然画像から医用画像への知識転移が可能であることを示している点で重要であるが、元のデータベースと対象のデータベースの間に比較的大きな差があるため、そのような応用は不可能である可能性がある。 また、**アプリケーションによって必要とされる微調整のレベルが異なることも観察された。** 具体的には、PE検出では、完全に接続された後期の層を微調整することで性能が飽和し、大腸内視鏡のフレーム分類では、後期と中期の層を微調整することで最高の性能を達成し、インターフェースのセグメンテーションとポリープ検出では、事前学習されたCNNの全層を微調整することで最高の性能を観測した。この結果は，特定のアプリケーションに対しては，浅いチューニングも深いチューニングも最適な選択ではないことを示唆している． **レイヤー単位の微調整を行うことで，効果的なチューニングの深さを知ることができる．** 層別微調整は、利用可能なデータ量に応じて、そのアプリケーションに最適な性能を実現するための実用的な方法を提供することができます。 我々の実験は、深く微調整されたCNNと完全に訓練されたCNNの両方が、対応する手作業で作られた代替品よりも優れた性能を示したことから、医用画像アプリケーションにおけるCNNの可能性をさらに確認するものである。