Exploring the Limits of Large Scale Pre-training

Abst

近年の大規模機械学習の発展は、データ、モデルサイズ、学習時間を適切にスケールアップすることで、事前学習の改善がほとんどの下流タスクに有利に伝達されることを観察することができることを示唆している。本研究では、この現象を系統的に研究し、上流の精度を上げると下流のタスクの性能が飽和することを証明しました。具体的には、Vision Transformers、MLP-Mixer、ResNetsについて、パラメータ数が1000万から100億の範囲で4800回以上の実験を行い、最大規模の画像データ（JFT、ImageNet21K）で学習し、20以上の下流の画像認識タスクで評価した。その結果、飽和現象を反映し、上流と下流の性能の非線形関係を捉えた下流性能のモデルを提案した。さらに、このような現象が発生する理由を掘り下げて理解するために、私たちが観察した飽和現象は、モデルの層を通して表現が進化する方法と密接に関係していることを示しました。また、さらに極端な例として、アップストリームとダウンストリームのパフォーマンスが相反する場合を紹介します。つまり、ダウンストリームのパフォーマンスを向上させるためには、アップストリームの精度を落とす必要があるのです。

1. Intro

転移学習や少数ショット学習に関する最近の目覚ましい進歩は、モデルをスケールアップして膨大なコーパスのデータで学習することが、データが少ない、あるいは全くない下流のタスクでの性能向上に向けての主な障害になるという新たな方向性を示唆している。顕著な例として、[Brown et al., 2020]では、大規模なコーパスのデータで学習した大規模な変換モデル[Vaswani et al., 2017]であるGPT-3が、多くの自然言語処理（NLP）タスクやベンチマークにおいて、少数ショットの設定で実質的な性能を達成することを示しています。画像認識タスクでは、Instagramの画像[Mahajan et al., 2018]やJFT-300[Sun et al., 2017]でのトレーニングが、転移学習および数ショット学習の設定で非常に効果的であることが証明されています[Goyal et al., 2021, Kolesnikov et al., 2019, Pham et al., 2020, Dosovitskiy et al., 2020, Dumoulin et al.］ 例が提供されていない場合（ゼロショット）でも、インターネット上の4億個の画像-テキストペアを用いて対照的な損失で学習した画像エンコーダモデルとテキストエンコーダモデルのペアで構成されるCLIP［Radford et al.，2021］は、驚くべき性能を発揮します。

上記のすべての開発は、暗黙のうちに2つの一貫した見解を促しています。

1. モデルとデータのサイズをスケールアップすることで、性能が大幅に向上すること
2. 性能向上が望ましい形で下流のタスクに移行すること。

Kaplanら[2020]は、1つ目の見解を支持する、より焦点を絞った実証研究において、言語モデリングタスクにおけるモデルサイズ、データ、および計算を適切にスケールアップすることで、性能が飽和しない形で戻ってくることを示しています。Belloら[2021]、Tan and Le[2019]は、画像認識タスクにおいても好ましいスケーリングが実現できることを示している。第2の見解もまた、最近の焦点を当てた研究の対象となっている。Hernandezら[2021]は、[Kaplanら, 2020, Tayら, 2021b]と同様の有利なスケーリング法則が、NLPタスクの転送および少数ショットの設定で成立することを示している。おそらく我々に最も近い先行研究であるKornblithら[2019]は、ImageNet[Russakovskyら、2015]でのパフォーマンスと下流の画像認識タスクとの間に線形関係1があることを観察している。

上記の見解を採用することは、今後大きな意味を持ちます。これらの見解によると、**1つの巨大なコーパスのパフォーマンスを向上させるために計算機や研究の労力を費やすことは、多くの下流のタスクをほぼ無料で解決できるため、有益であると考えられます。 また、上流の性能を向上させる一方で、下流のタスクについては、その向上が直線的な傾向に基づいて予測できるため、心配する必要がないということです。** 前述の研究は説得力のあるストーリーを提供していますが、大きな欠点があります。計算機の制限により、ハイパーパラメータ値の異なる選択に対するパフォーマンスは報告されていません。スケーリングプロットは，各スケールで選択されたハイパーパラメータが固定されているか，単純なスケーリング関数で決定されている場合には，より好ましいと思われます．さらに、多くの場合、目的は最先端の結果を改善することであり、したがって、ハイパーパラメータの選択における努力のほとんどは、当然のことながら、より高いスケールに集中しており、これはスケーリング・プロットを著しく偏らせている。しかし、**スケーリングの研究では、ハイパーパラメータのすべての可能な値が与えられたときに、モデルの下流側での最高のパフォーマンスに関心があります。さらに、ほとんどのスケーリング研究では、限られた範囲での挙動が報告されており、スケーリングのダイナミクスをさらに理解することなく、そのスケーリングを単純に外挿すると、研究した範囲の外でスケーリングが保持される理由が先験的に存在しないため、有害になる可能性があります。**

本論文では、大規模な上流タスクでの改善点を、少数ショットと転移学習の両方のシナリオで、広範囲の下流タスクに転移できるかどうかを体系的に調査する。上記の欠点を解決するために、我々の研究の一部は、4800以上のVision Transformer [Dosovitskiy et al., 2020]、MLP-Mixer [Tolstikhin et al., 2021]およびResNet [Dosovitskiy et al., 2020]モデルのメタ研究である。これらのモデルは、303Mの画像と18Kのクラスを持つJFT[Sun et al., 2017]または14Mの画像と21Kのクラスを持つImageNet21K[Deng et al., 2009]のいずれかで事前に学習され、少数ショットおよび伝達学習の設定で様々なダウンストリームデータセットで評価されます。我々の25のダウンストリームタスクは、VTAB [Zhai et al., 2019]、MetaDataset [Triantafillou et al., 2019]、Wilds [Koh et al., 2020]やメディカルイメージングなどのベンチマークに含まれる標準的なデータセットを幅広くカバーしています。

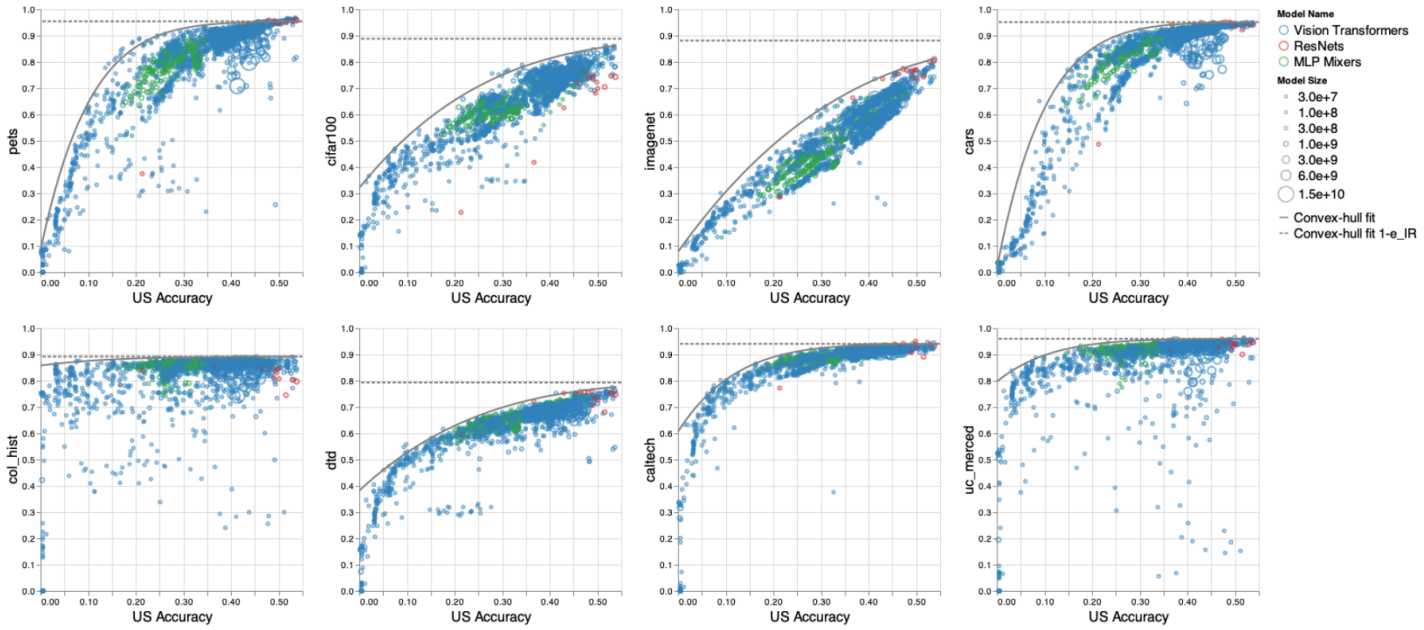

図1: 1500種類以上のVision Transformers、1400種類のMLP-Mixer、16種類の最も性能の高いResNetに基づいた、異なる下流 (DS) タスクと上流 (US) タスクの性能 (ResNetのサンプル数は少ないが、これは我々の調査に支障はない。詳細はセクション1.1を参照してください) を用いて，様々な設定を行いました．これらのモデルは，JFTで事前に学習され，数ショット (25ショット) の設定で評価されています．図2は同じプロットですが、JFTとImageNet21Kの2つの異なる上流タスクを含む4800回以上の実験を1ショットと25ショットで行っています。点の凸包も考慮しています。これは、これらのモデルを異なる確率で選択して作られたランダムな分類器の性能を捉えているからです。上流側の性能が向上すると、下流側の性能が飽和していきます。US精度が100%に達しても、DS精度は100%よりもかなり低い値に飽和してしまうのです。上流側と下流側の精度には非線形の関係があり，これをべき乗則関数でモデル化することで，USの精度が与えられたときのDSの性能を予測した。水平線は、上流側の精度が100%に達した場合に予測される下流側の精度です。ここでは，ハイパーパラメータの選択の影響を把握し，スケーリングがUS性能を通じてDS性能に影響を与えるという事実を考慮して，通常のDS-scaleプロットの代わりにDS-s-USプロットを調査した．図13は、多くの関連研究で行われているように、精度の対数スケーリングを行った場合のプロットです。図14は、上流がImageNet21Kの場合の同じプロットです。

我々は、画像認識タスクにおける少数ショット学習と転移学習のパフォーマンスにおけるスケールの役割を研究し、スケーリング (およびハイパーパラメータのチューニング) がワンモデル・フィッツ・オールのソリューションにつながらないという強力な経験的証拠を提供する。 しかし、まだ多くの未解決の課題が残されており、その中心となるのが、下流のタスクにおけるデータの多様性の問題です。我々は、この現象を初めて大規模かつ体系的に調査し、その理由を考察した。図1では、さまざ

まなモデルとダウンストリームタスクを対象に、ダウンストリーム（DS）とアップストリーム（US）のパフォーマンスを比較しています。USの精度を上げると、ほとんどの場合、DSの精度は100%を大幅に下回る値で飽和することがわかる。また、この飽和現象は例外ではなく、一般的な傾向であり、ショット数やUSタスクの選択に対しても頑健であることが分かりました (図2参照)。このギャップは、ノイズなどのDSタスクのみに依存する要因ではなく、USタスクとDSタスクの関係に依存することを立証した。さらに、同じようなUS精度を持つモデル群があったとしても、DSタスクによって最適なモデルは異なります。

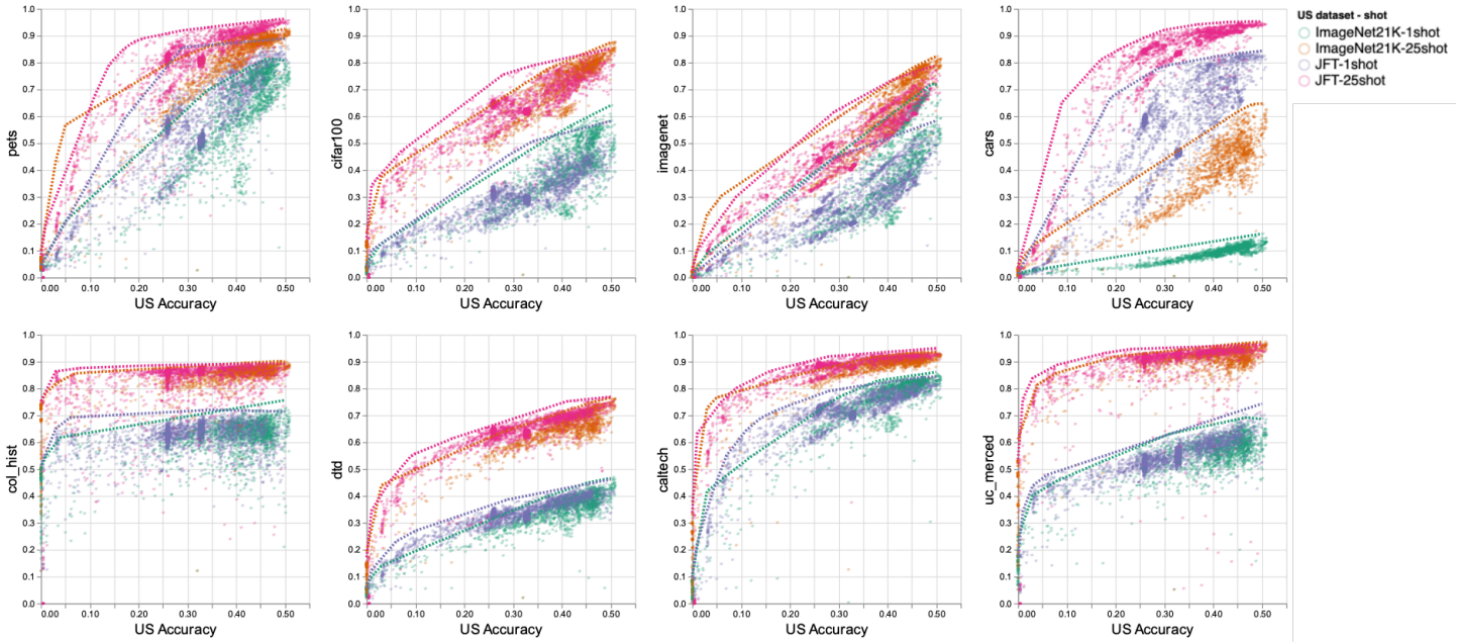

図2：4800種類以上の実験（2974個のVision Transformers、1593個のMLP-Mixer、249個のResNets）に基づく、下流（8種類のタスク）と上流のパフォーマンスの比較。実験は，上流のデータセット（JFTまたはImageNet21k）と，数ショット評価のショット数（1および25）に基づいてグループ化されている．点線は，DS-vs-USプロット上の点の凸包を示している．上流の2つのタスクで異なる値で飽和が起こるということは、飽和はDSタスクだけに依存するのではなく、むしろUSタスクとDSタスクの関係が重要であることを示唆している。

貢献度 本論文の主な貢献度は以下の通りです。

- 大規模な研究により、スケールアップやハイパーパラメータ、アーキテクチャの選択によって上流 (US) タスクの性能を向上させると、下流 (DS) タスクの性能が飽和することが明らかになりました。我々の実験では、いくつかのDSタスクが、調査した範囲内で完全に飽和状態に達しました (セクション2)。

- 同様のUS精度を持つモデル群が与えられた場合、あるDSタスクTDS1に対するベストモデルは、別のDSタスクTDS2に対するベストモデルと比較して、はるかに悪いパフォーマンスを示すことが実証されました (図6)。
- 実験の規模を考えると、提案モデルがDS-vs-USプロットのポイントの密度に影響されないことは非常に重要である。我々は、実験の凸包にべき乗則を当てはめることで、下流の精度の予測に対するサンプリングバイアスの影響を回避し、サンプルサイズの変化に対するモデルの頑健性を示すことができると主張し、実証しました (2.2項)。
- 上流側の精度と下流側の精度の間に非線形の関係があることを確認した後，与えられた上流側の精度に対する下流側の性能を予測するために，これらの関係をべき乗則曲線でモデル化し，サンプル数が少なくてもその挙動をよく捉えていることを確認しました (セクション2.2).
- モデルサイズ，データサイズ，計算量の拡大がDSの性能にどのように影響するかを調べ，これらのパラメータが主にUSの性能を通じてDSの性能に影響することを示した (セクション2.3).
- DSの性能が飽和する理由を調査し，この挙動は，学習済みモデルの高次層における特徴表現の有用性によって捉えられることを示す (セクション3).
- さらに，上流側と下流側の性能の不一致を調査し，ハイパーパラメータの選択によっては，両者が相反する可能性があることを示した．特に、事前学習 (上流タスク) で使用したヘッドの最適なハイパーパラメータが、USとDSで異なることを紹介します。そして、この不一致の理由を明らかにします (セクション4)。すなわち、重み減衰や学習率などの頭部のハイパーパラメータを変更することで、頭部で圧縮された情報を下層に押し下げることができ、上流タスクでの性能低下と、上流タスクに関連する下流タスクでの性能向上をもたらすことができます。これはレイヤーマージンや重みのL2ノルムで捉えることができます。
- 最後に，上流データのサイズ，精度の一般的なスケーリングの選択，ショット数，転移と数ショットの設定，アーキテクチャなどのいくつかの選択に対して，我々の観察結果がどのように頑健であるかを示します（セクション5）．
  関連する仕事 我々に最も近い研究は、Kornblithら[2019]の研究である。彼らは、数ショット、転送、ランダムな初期化のシナリオについて、12のデータセットでImageNet [Russakovsky et al., 2015]の事前学習が画像分類のパフォーマンスに与える影響を調査しています。彼らは，ImageNetでのパフォーマンスが，DSタスクでのパフォーマンスに (ロジットスケーリングで) 線形に変換されることを示している．しかし、彼らはその値の外挿を考慮していません。両者とも、様々な実験を通じて事前学習の効果を調べていますが、「上流のパフォーマンスの向上が下流のパフォーマンスの向上につながるか」という問いに対する回答には、2つの大きな違いがあります。**まず、DSとUSのパフォーマンスを比較した場合、明らかな「飽和」現象が存在することを確認しました。** 図1を見ると、AとBの2つのモデルを比較したときに、モデルAのUS精度がはるかに高く、DS精度が低いというさまざまなケースがありますが、これは例外ではなく、むしろ大多数のケースです。基本的には、DS-USプロットにおいて、一方が右にあり、他方が下にある2つの点は、このようなケースの例です。次に、各DSタスクにおいて、式1のように最適なモデルがべき乗則でスケーリングされていることがわかりますが、各アーキテクチャにおいて、最適なモデルはDSタスクごとに異なり、これはトレーニングのハイパーパラメータに依存しま

す (図6参照)。つまり、2つのDSタスク (TDS1 , TDS2) を考慮した場合、モデルAはUSとTDS1では優れた性能を発揮するが、DS2では優れた性能を発揮するとは結論づけられないケースが多数あります。この結論の違いは、以前の研究では考慮する精度値の範囲が限られていたためではないかと考えています。このような結論の違いに加えて、我々はこの飽和挙動の背景にある理由を調査する。さらに、セクション4では、USとDSの性能が相反する場合、つまり、USの性能低下がDSの性能向上につながるシナリオを検討します。[Zhai et al., 2021]が、事前学習時に頭部の重み減衰を大きくすると、DSの性能が向上する一方で、USの性能が低下することを指摘していることにヒントを得て、頭部のハイパーパラメータ（重み減衰と学習率の両方）をさらに調査し、これらの操作が頭部に蓄積された情報を下層に押し下げることに着目することで説明できることを示しています。その他の関連研究については、付録Aをご覧ください。

1.1. Experimental Setup

本論文の議論と分析は、画像認識タスクに関する膨大な数の大規模実験の研究と、我々の設定を解消し、研究された現象の理解を深めるために行った一連の制御実験に基づいています。Vision Transformers、MLP-Mixer、ResNetsを用いた4800以上の実験 (Vision Transformers 2974、MLP-Mixer 1593、ResNets 2493) を調査し、大量のデータを用いて教師付きで事前学習を行い、数ショットの学習と微調整を経て、いくつかの下流の画像認識タスクで評価しました。**これらの実験は、上流のデータセット(303Mの画像と18kのクラスを持つJFT-300M [Sun et al., 2017]または14Mの画像と21kのクラスを持つImageNet21K [Deng et al., 2009]のいずれか)、モデルのサイズと形状 (アーキテクチャの異なるハイパーパラメータ)、最適化(例：異なる学習率値と学習率スケジュール、異なる重み減衰、異なるオプティマイザ)、計算(例：エポック数)、および研究者が様々な目的のためにモデルの開発中に変更した他のノブの点で異なります。**

我々が調査した大規模な実験セットは、本論文の目的のために訓練されたものではなく、異なる研究者が異なる目的のために訓練した異なるViT、Mixer、ResNetモデルを集約して、それらについてのメタ研究を行っていることを強調します。このことは、このメタ研究がユニークな位置にあることを示しています。第一に，特定の現象を研究する目的で，これほど多くの大規模な試験を行うことは，経済的にも環境への影響の点でも実行不可能である可能性がある。第二に、これらの実験では、その後に行った分析の種類に関して、暗黙的または明示的な仮定がなされていないため、結果に含まれる分析プロセスの系統的なバイアスを最小限に抑えることができます。ただし、これ以外にもバイアスがある可能性があることに留意してください。例えば、**研究者は通常、特定のタスク (通常はImageNet) でSOTAを改善するためにハイパーパラメータのチューニングに焦点を当てていますが、これは、すべての可能なハイパーパラメータの高次元空間でのグリッド検索を行わないことにつながり、プロットに影響を与える可能性があります。セクション2.3では、これを調査し、このケースでは、観察された傾向は、グリッド検索を実行することに似ていることを議論します。**

実験では，主にViT-B／32を使用しています。ViT-B／32は，32×32のパッチサイズを持つ基本モデルです。JFTで7エポックの事前学習を行い，20以上のタスクで評価している．下流の評価では，主に，数ショットの学習設定 (1，5，10，20ショット) と，一部のアブレーションの微調整に焦点を当てている。これは、下流のデータポイントの数が増えると転移学習の効果が消失するという事実から動機づけられている[Kornblithら、2019年、Zophら、2020年、Mensinkら、2021年]。それゆえ、我々は転移学習が最も輝く設定に焦点を当てます。集約実験と制御実験の両方において、少数ショット設定では、クラスごとに固定数の訓練例のみが与えられ、凍結した事前訓練モデルからの表現の上に線形分類器が訓練される。微調整セットアップでは、VTAB標準[Zhai et al., 2019]に従い、下流タスクから1000個のトレーニングサンプルを使用し、下流ヘッド以外のモデルのすべてのパラメータを更新します。上流および下流タスクのベンチマークとトレーニングセットアップの詳細は、付録Dに記載されています。

本文では、紙面の都合上、8つの下流タスクでの結果を報告し、20以上の下流タスクを含む結果とプロットを付録Cに記載しています。また、JFTでの事前学習に関連するプロットを本編に掲載し、ImageNet21Kでの事前学習に対応するプロットを付録Cに掲載しました。

2. 転移学習におけるスケールアップの利益逓減について

転移学習の顕著な目標は、下流のタスクで良いパフォーマンスを発揮することです。 我々が最初に取り組む問題は、上流のタスクでのパフォーマンス向上が、異なる下流のタスクでのパフォーマンスにどのように影響するかということである。この効果をモデル化して、下流のパフォーマンスを予測することに興味がある。そのために、セクション1.1で説明した大規模な実験セットについて、DSとUSのパフォーマンスを調査した。前述したように、これらの実験は、モデルのサイズや形状、最適化手法、計算量など、研究者が視覚タスクにおいて最先端の結果を追いかけるなど、さまざまな目的でモデルを開発する際に変更したハイパーパラメータによって異なります (2.2項)。次に、**モデルサイズ、USデータサイズ、計算量の3つの軸でスケールアップし、ショット数を変化させた場合のDS性能への影響を調べるコントロール実験を行います (2.3節)。**

2.1. Recap: ランダムな分類法

DSとUSの性能比較に入る前に，ランダム化分類器の概念をおさらいしておきましょう (このセクションではランダム化分類器を多用します).

上流と下流の精度を持つ2つの分類器 $a1 = (a_1^{US}, a_1^{DS}), a2 = (a_2^{US}, a_2^{DS})$ が与えられたとき，それぞれの入力に対して，確率p_1の第1の分類器の出力と，確率$1 - p_1$の第2の分類器の出力を独立に選

ぶことで，ランダム化された分類器を作ることができます．すると、ランダム化分類器は、$p_1 a_1 + (1 - p_1) a_2$ の精度を示します。つまり，ランダム化分類器の精度は，2つの分類器の精度を凸状に組み合わせたものになります．p_1の値を掃引することで、この凸結合の経路上のすべての点を達成することができます。2つ以上の分類器がある場合、この概念を拡張することができます。次のレンマにあるように、このようなランダム化された分類器の精度は、その端点の精度の凸結合であることを示すことは難しくありません。

Lemma 2.1.
あるタスクのペア (US, DS) で精度 $a_j = (a_j^{US}, a_j^{DS}), j \in [N]$に達するモデル群 $\theta_j, j \in [N]$ を考える。各入力 x_i に対して、確率 p_j でモデル θ_j を選び、$\theta_j(x_i)$ を出力する、というように、ランダム化モデル $\bar{\theta}$ を構築する。すると、ランダム化モデルは精度 $\sum_{j=1}^{N} p_j a_j$ を示す。

証明については、付録Bを参照してください。

したがって，学習したモデルのDS対USの精度の凸包上のすべての点が達成可能であり，それに到達するための前述の方法があることになります．これにより，手元の学習済み分類器の性能の凸包に相当する精度を示すランダム化分類器が得られます．

以上の考察から，実験結果に対応する点に加えて，モデル性能の凸包の上側の包 (与えられたUSの精度ごとに最も高いDSの精度を表す) を分析に含めることにした。これにより，プロット内の点の密度に依存しない，DSとUSの関係のモデルを得ることができる．この点についてはセクション2.2で詳しく説明します。

2.2. ダウンストリーム精度のスケーリング法則

図1は，異なるアーキテクチャをJFTで事前学習し，数ショット (k = 25) のDSタスクのセットで評価した3,000回以上の実験のDS-vs-US性能を示している．図2は、4800回の実験（JFTまたはImageNet21Kで事前学習したもの）を対象に、1ショットまたは25ショットの両方で同様のプロットを示しています。

我々のモデルの性能を考慮して、**我々は、USのパフォーマンスを向上させるために、DSタスクのパフォーマンスがどのように変化するかを予測することに興味がある。** そのために、DS-US性能プロットに曲線を当てはめました。今回の分析は、**DS精度対データセットサイズ、モデルサイズ、計算量ではなく、DS精度対US精度を分析している点で、スケーリング則を分析した以前の研究[Kaplan et al., 2020, Hernandez et al., 2021, Zhai et al., 2021]とは異なることを強調しておきます。** ほとんどの場合、USでの性能向上はスケーリング (データセットサイズ、モデルサイズ、計算量) によって達成されるため、このアプローチはスケーリングの影響を間接的に捉えていることになります。この点については、セクション2.3で説明します。

DSとUSの比較を行う際には、適切なスケーリングを選択することが重要です。Kornblithら[2019]は、ImageNetで事前学習されたモデルのDS-vs-US曲線を調査し、ロジットスケーリングで精度をプロットした場合、線形のDS-vs-US性能を報告している5。 上流タスクと下流タスクの関係に関する先行研究では、ロジットスケーリングを使用している[Rechtら, 2018, 2019]。ロジットスケーリングが誤差0.5付近で対称的な挙動を示すことがこれらの問題では不自然であることを考えると、スケーリング則の文献で用いられているlogスケーリングの方が適切であると主張する。logスケーリングにおけるUSとDSの性能の線形関係は以下のように捉えることができる。

$$e_{DS} = a(e_{US})^b$$

図1を見ると、その挙動は直線的ではないことがわかります。むしろ、DSタスクのパフォーマンスはある時点で飽和し、その時点はDSタスクごとに異なる。

2.2.1. パフォーマンスの飽和

図1および図2の観測結果にヒントを得て、飽和点を定義します。以下では
飽和値をさらに数学的にモデル化して調べます。

定義 2.2 (飽和値)
下流側と上流側の精度を考慮し、下流側タスクのT_{DS}において、上流側の精度が 1.0 に達したときの下流側の精度の値を飽和値と定義する。

定義2.2を考慮すると、性能飽和とは、ある上流の精度値が存在し、それを超えると下流での性能向上が非常に小さくなることを意味します。したがって、下流の精度への影響は無視できるため、データサイズ、計算、モデルサイズを拡大してUSの精度を向上させる価値はありません。

この関係は線形ではないため、DSの性能を予測するためには、プロットに適合する関数形式が必要です。スケーリング・ローに関する最近の研究[Kaplan et al., 2020, Hernandez et al., 2021]に触発されて、我々は以下の関数形を提案する。

$$e_{DS} = k(e_{US})^\alpha + e_{IR}$$

ここで、e_{DS}、e_{US}はそれぞれ下流側と上流側の誤差 $1 - accuracy$ を表し、k、αは定数、e_{IR}は不可逆誤差を表します。

救いようのないエラー (e_{IR}) は，US のエラーがゼロになった場合の DS のエラーの値を表しており，したがってバイアス項に似た働きをする．e_{IR} 項は，US と DS の精度の間の非線形性の傾向を表している．つまり、式1をログ・スケーリングでプロットすると、eIRがゼロのときにのみ依存関係が線形になる。

図1のDS-US精度プロットの $1 - e_{IR}$ に対応する線をスケッチし、多くの下流タスクでは1.0に近づかず、US精度が高くてもUSのパフォーマンスがDSのパフォーマンスに反映されないことに注意してください。このように、一般的に考えられていることとは異なり、飽和現象は例外ではなく、DSタスクでは典型的な現象であることがわかりました。

2.2.2. デザインの選択がパワー・ロー・パラメータに与える影響

図2を見ると、DSタスクによって飽和値が異なり、USタスクが変わるとこの値も変化することがわかります。さらに，ショット数を変えると e_{IR} も変化する．さらに、図2では、異なるUSデータセット (ImageNet21KとJFT) で学習したモデルと、転送に使用するショット数を変えたモデルの、同じセットのDSタスクに対するDS対USの精度を比較しています。その結果、**飽和時のDS精度は、USデータセットに依存することがわかりました。**

上記の観察結果をより明確に示すために，様々な選択がパワー則 (式1) のパラメータにどのように影響するかを図3，15，16に示します．USタスクとDSタスクの選択は、すべてのパラメータに影響を与えるが、ショット数は主に k と e_{IR} に影響を与えることがわかる。具体的には、ショット数を増やすと、e_{IR} が低くなる。

要するに，モデルと学習アルゴリズムの特定の選択に対して，以下のような関数f1(-)，f2(-)，f3(-)が存在するということである．