言語モデルの再利用のスケーリングに対する悪影響

リュウ　センペイ [1]　加藤　拓也 [1]
[1]SB Intuitions 株式会社
{sengpei.liew,takuya.kato}@sbintuitions.co.jp

概要

既存の学習済みモデルを継続学習やモデル拡張に再利用することは、ゼロからモデルを学習するコストを削減する手法として期待されています。しかし、特に「過学習」のベースモデルを再利用した場合の有効性は不明でした。本研究では、モデル再利用のスケーリング特性を経験的に調査し、スケーリング効率が予測可能な形で減衰することを発見しました。具体的には、第 2 段階の学習トークン数に対するスケーリング指数は、ベースモデルの学習トークン数の対数に比例して減少します。この飽和効果は、シンプルなスケーリング則によって正確にモデル化できます。この知見は、ベースモデルを学習させればさせるほど、追加学習の恩恵が少なくなるというトレードオフを浮き彫りにし、効率的な言語モデル学習に向けた実践的な指針を提供します。

1　はじめに

大規模言語モデル（LLM）のゼロからの学習には莫大な計算資源と時間が必要です。この課題に対し、既存の学習済みモデルを「再利用」する戦略が注目されています。これには、特定のドメイン性能を高める継続学習（CPT）や、既存モデルのパラメータを流用してモデルサイズを拡大するモデル拡張（Model Growth）が含まれます。

しかし、ベースモデルが「過学習（過剰なトークンで学習）」されている場合、その高度に最適化されたパラメータが、拡張後の探索や新データへの適応を困難にし、第 2 段階の学習速度を低下させることが直感的に予想されます。本論文では、このスケーリング特性を定量化しました。技術的には、本研究の主な貢献は以下の通りです。

- 広範な実証実験の実施: さまざまなモデルサイズ、学習トークン数、データセットを組み合わせ、計 450 回以上の実行（ラン）に及ぶ大規模な実験を行いました。これにより、継続学習（CPT）やモデル拡張（Model Growth）といった既存モデルを再利用した事前学習のスケーリング挙動を詳細に調査しました。
- 新しいスケーリング則の提唱: 幅広い設定において、以下の経験的なスケーリング関係式が成立することを発見しました。

$$L(D_1, D_2) = AD_1^{-\alpha_1} D_2^{-\alpha_2+\alpha_3 \log D_1} + E \quad (1)$$

ここで、L は第 2 段階学習後のバリデーションロス、D_1 および D_2 はそれぞれ第 1 段階と第 2 段階の学習トークン数、$A, \alpha_1, \alpha_2, \alpha_3, E$ は正の定数です。特に $\alpha_3 \log D_1$ を「相互作用項」と呼び、D_2 に関するスケーリング指数がこの項によって変化することで、モデル再利用における「飽和効果」を定量化できることを示しました。また、他の関数形と比較しても、この式が最も実測データに適合することを証明しました。
- 飽和効果のメカニズム解明と実用的指針: 過学習（Overtrained）気味のモデルにおける勾配ノルム（Gradient Norms）を分析することで、飽和効果が生じるメカニズムを説明しました。さらに、このスケーリング則を用いることで、飽和効果による効率低下を避けるために、「いつ既存モデルの再利用をやめて、ゼロから学習（Train from scratch）すべきか」という実用的な判断基準を提示しました。
- モデルサイズへの拡張: スケーリング挙動の分析をモデルサイズ N にまで広げ、データセットサイズとモデルサイズを統合して扱う「統合スケーリング則」へと拡張できることを示しました。

関連研究　我々は、最終的なバリデーションロス L が、学習トークン数やモデルサイズといった単一の注目変数に対して、広く観察されているべき乗則スケーリング（Power-law scaling）に従うと仮定しま

す [8, 7, 6, 3, 12, 14, 16, 11, 1, 9, 4]。

$$L = AX^{-\alpha} + E$$

ここで、X は注目する変数、α はスケーリング指数、A はスケーリング係数、そして E はデータ分布固有のエントロピーに起因する削減不能なロス（既約損失）です。注釈: 厳密には、$X \to 0$ での値が有限であることを保証するために $L = \frac{A}{(X+1)^{\alpha}} + E$ という形式を想定しています。しかし、実用上は $X \gg 1$（通常 10^6 以上）であるため、表記を簡略化して $L \approx AX^{-\alpha} + E$ と近似します。また、記号 A, α, E は、変数が異なれば値も異なりますが、混同の恐れがない限り、便宜上同じ記号を使用します。スケーリング則の背景本研究では、ニューラルネットワークに限らず自然界や人工的な現象においても一般的に観察される「最終ロス」のべき乗則的な挙動のモデリングに焦点を当てています。なお、最近のスケーリング則の研究では、より複雑な関数形を用いて学習曲線全体をフィットさせようとする試みもありますが、我々の焦点はトークン数に対する汎用的なスケーリング挙動にあります。

2 実験

本研究では、LLaMA ライクなデコーダーのみの Transformer アーキテクチャを採用し、以下の 2 段階で学習を行いました。

- 第 1 段階 (D_1): インターネットデータ（Slimpajama-DC）を用いたベースモデルの学習 [17]。
- 第 2 段階 (D_2): 以下の手法によるモデルの再利用。
  - 継続学習 (CPT): コード（StarCoder）や数学（OpenWebMath）データへの適応 [13, 15, 10]。
  - モデル拡張 (Model Growth): 層を重ねる「Stacking」や、隠れ層を広げる「Width Expansion」によりパラメータ数を拡大 [2, 5]。

モデルサイズは 15M から 1.1B まで幅広く検証しました。

2.1 実験結果

ここでは、スケーリング則の定式化の動機となった実験的な観察結果を示します。Figure 1 は、0.1B（1 億パラメータ）サイズのベースモデルを使用し、コードデータでの継続学習（CPT）と、成長因子 2（層を 2 倍にする Stacking）での実験結果を示したものです。第 1 段階のトークン数 D_1 と第 2 段階のトークン数 D_2 を 5×5 のグリッド状に変化させて学習を行いました。Figure 1 の左パネルには、第 2 段階終了後のバリデーションロスを D_2 の関数としてプロットしています（D_1 の値ごとに色分け）。ここから以下の観察が得られました。

観察 1 ：ロスは第 2 段階のトークン数 D_2 に対してべき乗則（Power Law）に従うことが観察されました。これは典型的なニューラルスケーリング則と一致しており、「固定された初期状態から学習を始める場合、トークンを追加すれば予測可能な形で性能が向上する」という期待を裏付けるものです。さらに分析を進めると、べき乗則のスケーリング指数（グラフの傾き）が、D_1 が大きくなるにつれて減少することがわかりました。この傾向は、特にモデル拡張（Model Growth）において顕著でした。この関係を定量化するため、スケーリング指数のマイナス値 $-\alpha(D_1)$ を D_1 の関数としてプロットしたところ、明確な対数依存性が明らかになりました。

$$-\alpha(D_1) = \gamma \log D_1 + E'. \tag{2}$$

この関係をべき乗則の式に代入すると、$D_2^{-E'+\gamma \log D_1}$ という項が導かれます。これが、私たちが提唱する乗法的スケーリング則における「相互作用項」の直接的な根拠となります。また、係数 A についても $A \propto D_1^{-\alpha_1}$ という乗法的な依存関係が良好に保持されていることが確認されました。

以上の観察から、次の結論が導かれます。

観察 2 ：第 2 段階のロスが D_1 と D_2 に共同で依存する挙動は、相互作用項を持つ乗法的なスケーリング則によって捉えることができます。

スケーリング則の解釈 ：このスケーリング則からは、定量的・定性的にいくつかの重要な洞察が得られます。

1. 初期状態の改善: D_1 を固定すると、D_2 に対する実効的なスケーリング係数は $AD_1^{-\alpha_1}$ となります。D_1 が増える（ベースモデルをより長く学習させる）ほどこの値は小さくなるため、第 2 段階の学習開始時点での初期ロスが低くなります。これは「より良く学習されたベースモデルは、追加学習の強力な出発点になる」という従来の知見と一致します。
2. 収穫逓減（飽和効果）: 一方で、D_2 に対する実効的なスケーリング指数は $\alpha_2 - \alpha_3 \log D_1$ とな

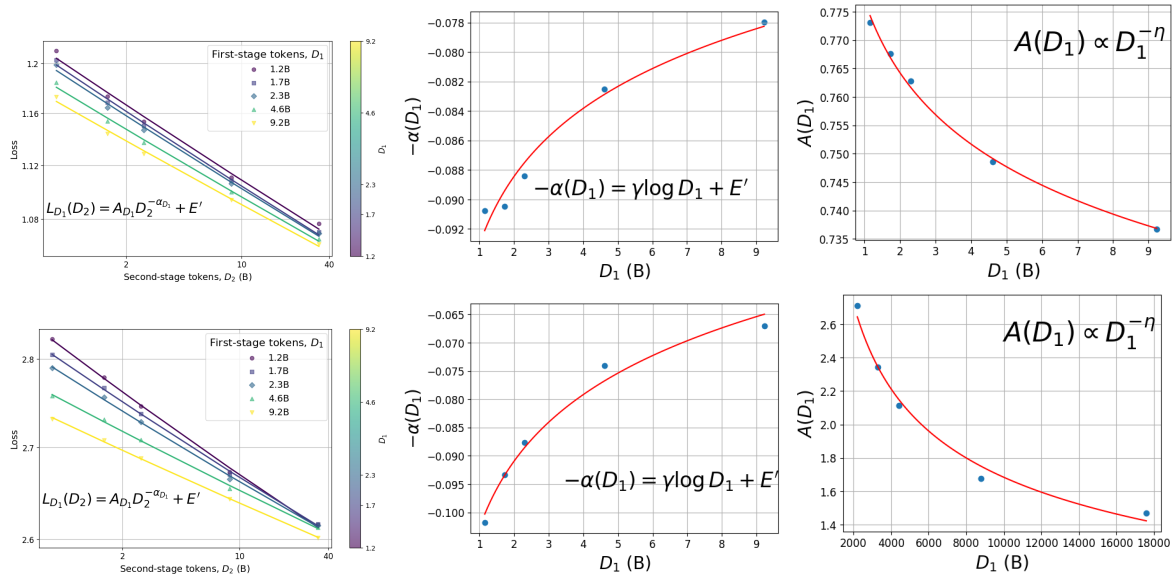

図 1　model reuse with overtrained base models leads to saturation in scaling behavior. Left: D_2 has power-law scaling. We show scaling behavior of second-stage training tokens (D_2) for different values of first-stage tokens (D_1), trained on a 0.1B base model. **Middle: Interaction term explains decreasing exponents.** The fitted exponents in the left plots are used to fit Equation 2 as a function of D_1, and are shown to agree well with the functional form. **Right: Scaling factor has power-law scaling w.r.t. D_1. Top: Continual pretraining (CPT) on code data. Bottom: Model growth from 0.1B to 0.2B by stacking.**

ります。これは、ベースモデルが過学習（D_1 が増大）するほど、第 2 段階でトークンを追加した際のロス改善効率がますます悪くなることを意味します。言い換えれば、ベースモデルを学習させればさせるほど、追加学習から得られるリターンが減少し、スケーリングの飽和（Saturation）として現れるのです。

3　実験結果の分析

異なる関数形の比較　加法的な形式や相互作用項のない乗法形式などと比較した結果 [14]:

$$\textbf{Multiplicative: } L(D_1, D_2) = AD_1^{-\alpha_1} D_2^{-\alpha_2 + \alpha_3 \log D_1} + E. \tag{3}$$

$$\textbf{Additive: } L(D_1, D_2) = AD_1^{-\alpha_1} + FD_2^{-\alpha_2} + E. \tag{4}$$

$$\textbf{Hybrid: } L(D_1, D_2) = \left(AD_1^{-\alpha_1} + F\right) D_2^{-\alpha_2} + E. \tag{5}$$

提案した相互作用項（$\alpha_3 \log D_1$）を持つモデルが、CPT とモデル拡張の両方において一貫して最小の誤差（RMS）を記録しました (表 1)。

デルサイズを含めた統合スケーリング則　モデルサイズ N を変数に加えた場合も、以下の式で精度高く予測できることが確認されました [9]。

$$L(D_1, D_2, N) = AD_1^{-\alpha_1} D_2^{-\alpha_2 + \alpha_3 \log D_1} + BN^{-\beta} + E$$

これにより、実験で使用したサイズ以上の大規模なモデルやデータセットにおける性能を正確に外挿（予測）することが可能になります (図 2 の左)。

3.1　メカニズムの解析

過学習モデルがなぜ学習しにくくなるのかを調査するため、勾配ノルム（Gradient Norms）を分析しました。その結果、過学習されたモデルは第 2 段階の学習開始時に勾配ノルムが非常に大きく、ロスランドスケープが急勾配になっていることが判明しました。これが、最適化を困難にする物理的な要因の一つと考えられます (図 3)。

3.2　実践的な指針

提案したスケーリング則を用いると、「既存モデルを拡張して使うべきか」それとも「最初から（Scratch）学習すべきか」の判断基準が得られます (図 2 の右)。

- 追加の学習予算 (D_2) が少ない場合は、モデル拡張が有利です。
- しかし、D_2 が一定の閾値を超えると、過学習による飽和の影響で、ゼロから学習したモデルに性能を追い抜かれます。この閾値はモデルサイズが大きくなるほど小さくなる傾向があります。

表 1　Multiplicative scaling law with interaction consistently achieves lowest error. Leave-one-out RMS error ($\times 10^{-3}$) for fitting the loss for model reuse of a 0.1B base model.

RMS ($\times 10^{-3}$)	Code	Math	CPT (rep)	CPT (sta)	Exp x2	Stk x2	Stk x4	Stk (sta)
Mul.	**1.573**	**1.737**	**0.987**	**1.371**	**2.790**	**2.845**	**2.152**	**2.957**
Mul. ($\alpha_3 = 0$)	2.095	3.147	2.222	2.366	4.837	7.818	6.557	8.757
Add.	3.913	7.443	4.646	4.315	9.855	10.323	8.866	10.559
Hyb.	2.245	3.340	2.223	2.367	4.841	7.748	6.667	8.822

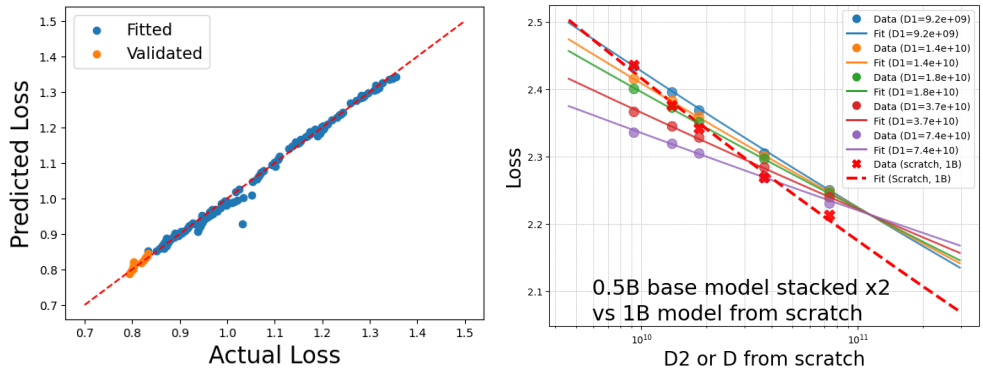

図 2　Left: Joint scaling law fit for continual pretraining on code data. Orange points indicate the 10% of data with lowest losses used for validation. **Left: Loss versus second-stage (from-scratch) training token for stacking (training from scratch) a 0.5B-to-1B model (1B model).** Data points and fitted lines using our multiplicative scaling law with different number of first-stage training tokens (power-law) are shown. It can be seen that as the number of training tokens increases, the losses from model growth saturate, and from-scratch training eventually outperforms it.

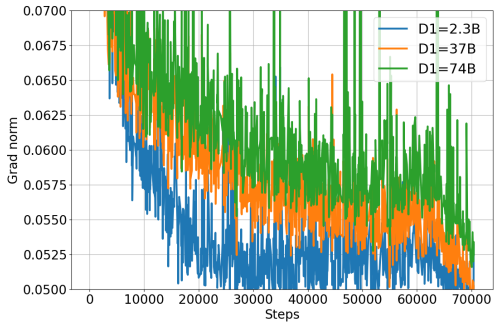

図 3　Overtrained models have larger gradient norms when undergoing model reuse. We show the gradient norm curves with respect to number of training steps in the second stage for base models trained with different values of first-stage tokens (D_1). We see that overtrained models have larger gradient norms, indicating that the loss landscape is sharper.

4　終わりに

本研究では、2 段階の事前学習に関する広範なスケーリング調査を実施しました。今後さらに深く追求すべき潜在的な研究方向がいくつかあります。以下にその主要なものを挙げます。

他の要因のスケーリング則への組み込み 本研究では、継続学習（CPT）におけるリプレイ比率や、モデル拡張における拡張係数（Growth Factor）といった手法特有の要因に関するスケーリング挙動については考慮しませんでした。しかしながら、これらの要因は、べき乗則に基づいた我々のスケーリング則の中に極めて自然に組み込むことができると考えています。

飽和効果の解消 本研究は、ブートストラップ・プリトレーニング（モデル再利用による事前学習）で見られる飽和効果に対し、メカニズム的な説明と実用的な指針を提供しましたが、これらの効果をより効果的に「軽減」する方法という根本的な問いは依然として解決されていません。LLM の学習で一般的に採用されている極めてスケーラブルな正則化技術（レイヤー正規化など）では、この問題を完全には解消できないようです。これを実現するための、より高度でスケーラブルな技術の開発は、本研究で提示したスケーリング則に基づくガイドラインを補完する極めて重要なソリューションとなるでしょう。

参考文献

[1] Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. **arXiv preprint arXiv:2404.10102**, 2024.

[2] Tianqi Chen, Ian Goodfellow, and Jonathon Shlens.

Net2net: Accelerating learning via knowledge transfer. **arXiv preprint arXiv:1511.05641**, 2015.

[3] Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. pages 4057–4086. PMLR, 2022.

[4] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. **SIAM review**, 51(4):661–703, 2009.

[5] Wenyu Du, Tongxu Luo, Zihan Qiu, Zeyu Huang, Yikang Shen, Reynold Cheng, Yike Guo, and Jie Fu. Stacking your transformers: A closer look at model growth for efficient llm pre-training. **arXiv preprint arXiv:2405.15319**, 2024.

[6] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. **arXiv preprint arXiv:2010.14701**, 2020.

[7] Joel Hestness, Newsha Ardalani, and Gregory Diamos. Beyond human-level accuracy: Computational challenges in deep learning. pages 1–14, 2019.

[8] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. **arXiv preprint arXiv:1712.00409**, 2017.

[9] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. **Proceedings of the 36th International Conference on Neural Information Processing Systems**, pages 30016–30030, 2022.

[10] Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. **Transactions on Machine Learning Research**, 2024.

[11] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. **arXiv preprint arXiv:2001.08361**, 2020.

[12] Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, et al. Scaling laws for fine-grained mixture of experts. **arXiv preprint arXiv:2402.07871**, 2024.

[13] Raymond Li, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, LI Jia, Jenny Chim, Qian Liu, et al. Starcoder: may the source be with you! **Transactions on Machine Learning Research**, 2023.

[14] Seng Pei Liew, Takuya Kato, and Sho Takase. Scaling laws for upcycling mixture-of-experts language models. In **Forty-second International Conference on Machine Learning**, 2025.

[15] Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text. **arXiv preprint arXiv:2310.06786**, 2023.

[16] Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. **Advances in Neural Information Processing Systems**, 37:100535–100570, 2024.

[17] Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, et al. Slimpajama-dc: Understanding data combinations for llm training. **arXiv preprint arXiv:2309.10818**, 2023.