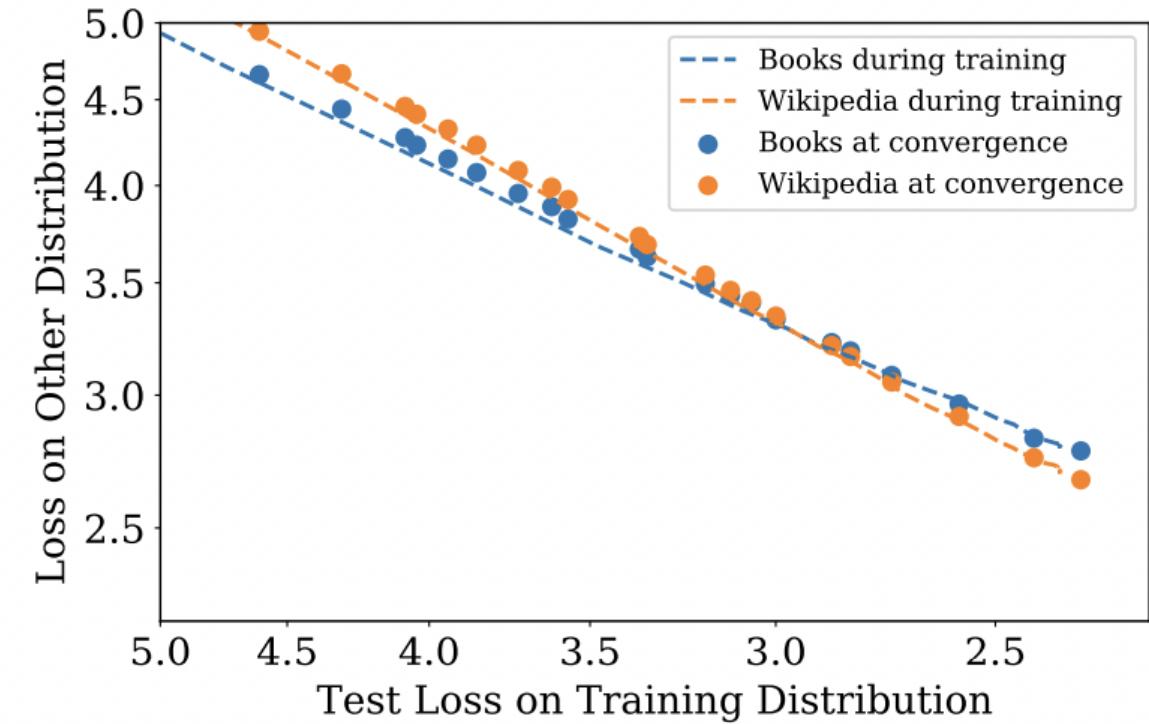
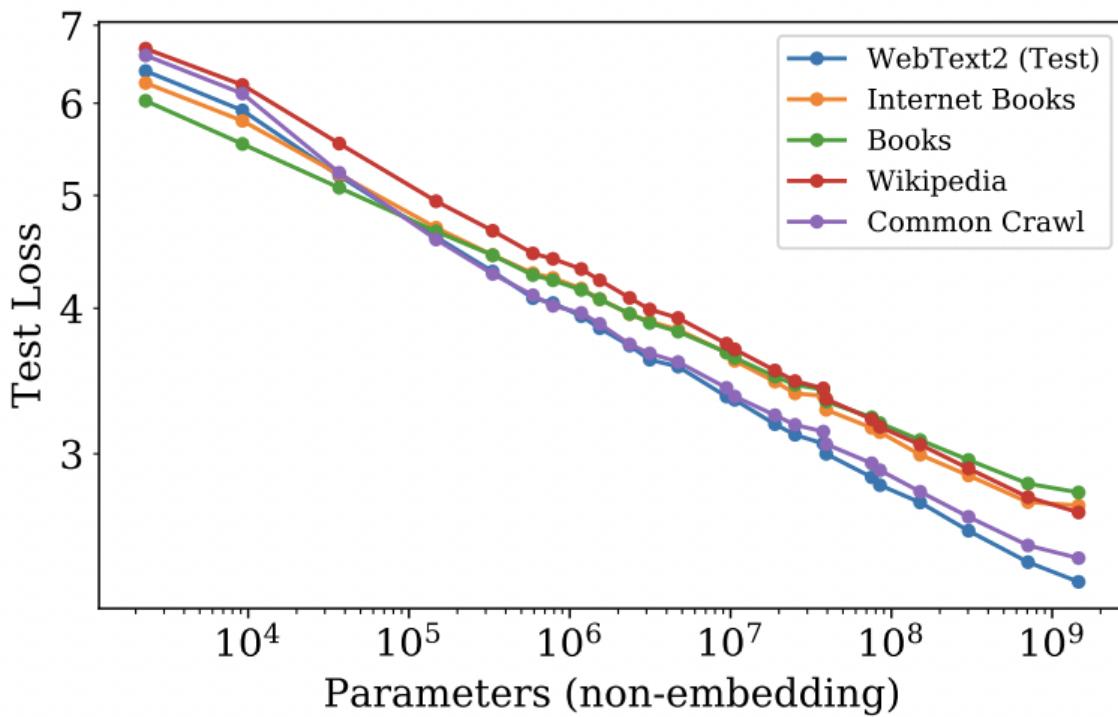


Independent with training data



“Thermodynamics” of LLM

$$L(N, D) = \left[\left(\frac{N_c}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D} \right]^{\alpha_D}$$

$$L(N, S) = \left(\frac{N_c}{N} \right)^{\alpha_N} + \left(\frac{S_c}{S_{\min}(S)} \right)^{\alpha_S}$$

- L: test loss
- D: size of training data in tokens
- N: # of parameters(non-embedding)
- B: optimal batch size
- $C \approx 6NBS$: an estimate of the total non-embedding training compute
- C_{\min} : optimally allocated compute budget
- S: # of training steps
- S_{\min} : the minimum possible number of optimization steps

With C is given

$$N \propto C^{\alpha_C^{\min}/\alpha_N}, \quad B \propto C^{\alpha_C^{\min}/\alpha_B}, \quad S \propto C^{\alpha_C^{\min}/\alpha_S}, \quad D = B \cdot S$$

$$\alpha_C^{\min} = 1 / (1/\alpha_S + 1/\alpha_B + 1/\alpha_N)$$

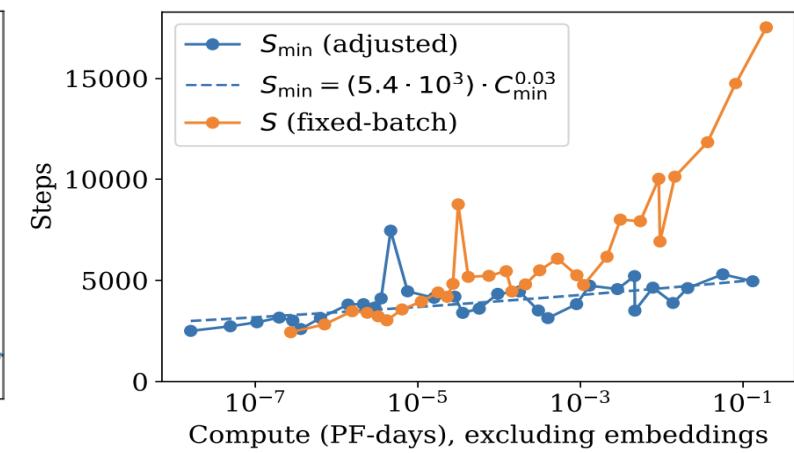
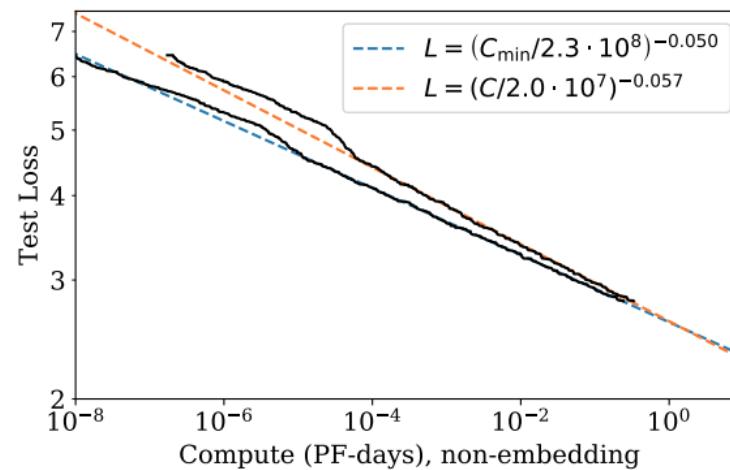
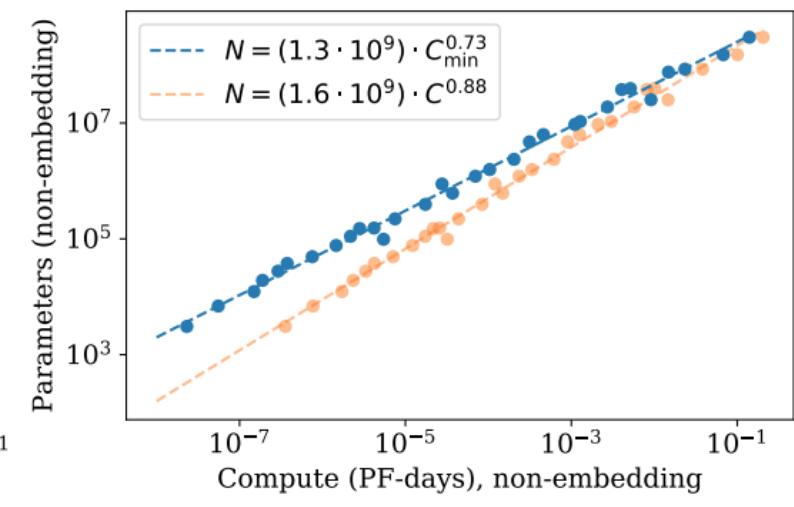
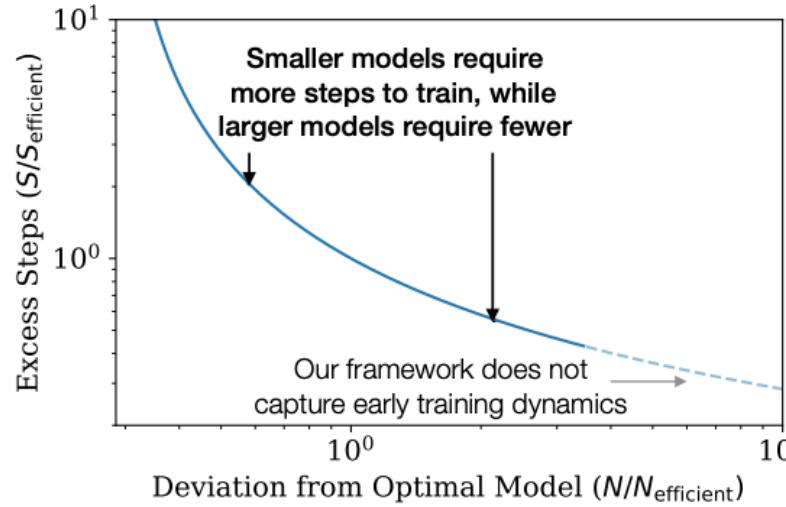
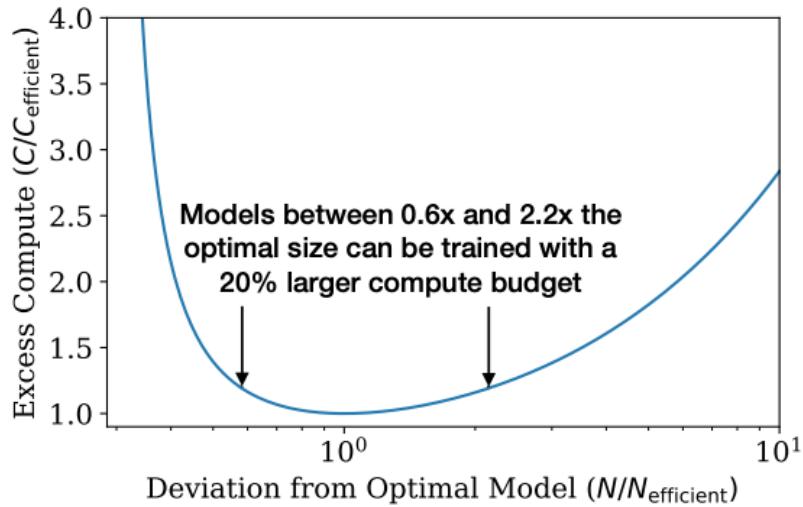
$$B_{\text{crit}}(L) \approx \frac{B_*}{L^{1/\alpha_B}}$$

$$B_* \sim 2 \cdot 10^8 \text{ tokens}, \quad \alpha_B \sim 0.21$$

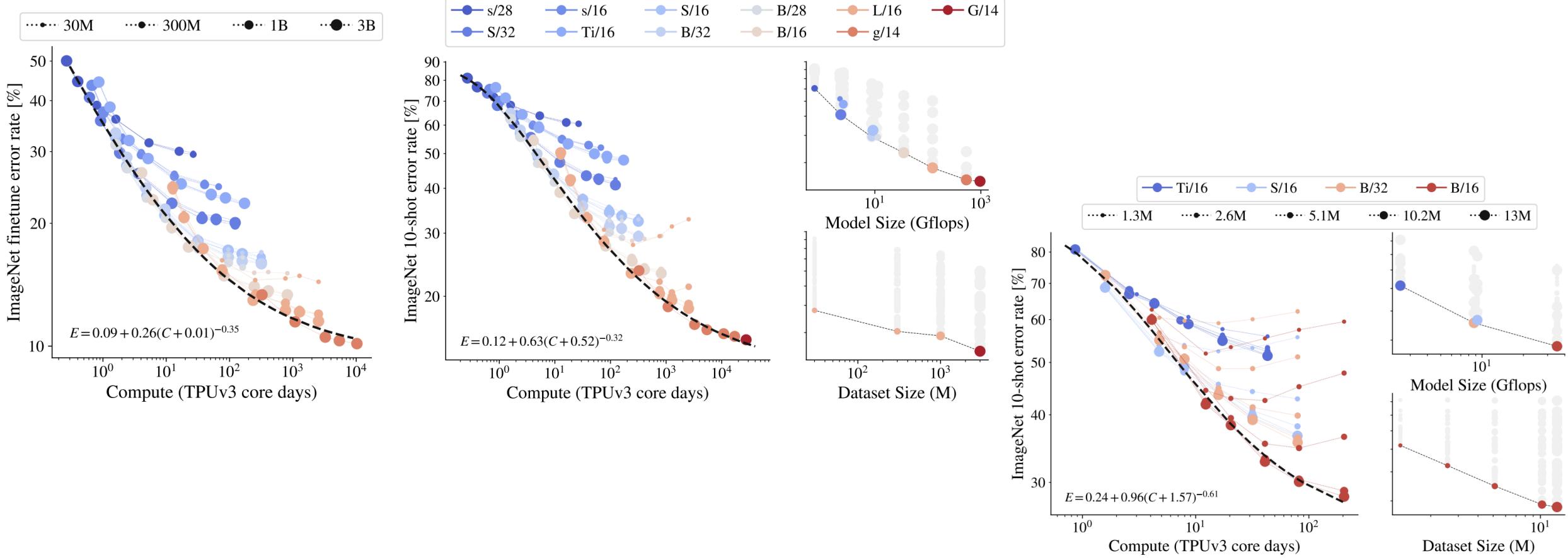
$$S_{\min}(S) \equiv \frac{S}{1 + B_{\text{crit}}(L)/B}$$

$$C_{\min}(C) \equiv \frac{C}{1 + B/B_{\text{crit}}(L)}$$

Optimal Allocation of the Compute Budget



Scaling laws of vision transformer



Emergent abilities

Published in Transactions on Machine Learning Research (08/2022)

Emergent Abilities of Large Language Models

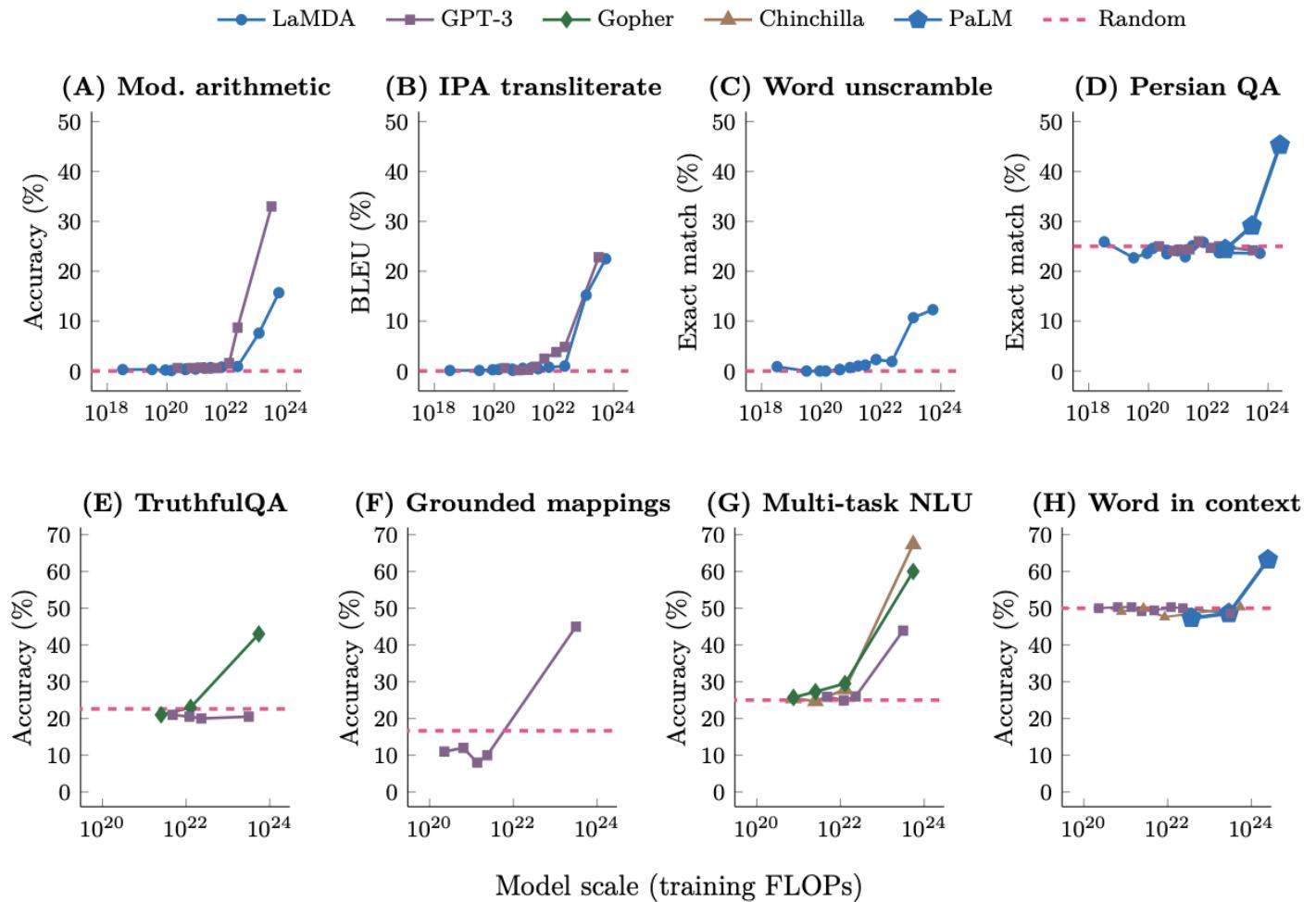
Jason Wei¹
Yi Tay¹
Rishi Bommasani²
Colin Raffel³
Barret Zoph¹
Sebastian Borgeaud⁴
Dani Yogatama⁴
Maarten Bosma¹
Denny Zhou¹
Donald Metzler¹
Ed H. Chi¹
Tatsunori Hashimoto²
Oriol Vinyals⁴
Percy Liang²
Jeff Dean¹
William Fedus¹

¹ Google Research ²Stanford University ³UNC Chapel Hill ⁴DeepMind

Reviewed on OpenReview: <https://openreview.net/forum?id=yzkSU5zdwD>

Abstract

Scaling up language models has been shown to predictably improve performance and sample efficiency on a wide range of downstream tasks. This paper instead discusses an unpredictable phenomenon that we refer to as *emergent abilities* of large language models. We consider an ability to be emergent if it is not present in smaller models but is present in larger models. Thus, emergent abilities cannot be predicted simply by extrapolating the performance of smaller models. The existence of such emergence raises the question of whether additional scaling could potentially further expand the range of capabilities of language models.



In-context learning - 类比思维

YM

如果“abc”对应的是“abd”，那么“ijk”对应什么？



根据题目中的信息，我们可以推断出“abc”到“abd”的转换是一种替换（substitution）操作，



即将原本的“c”替换成了“d”。因此，“ijk”的对应关系应该也是替换操作，即将其中一个字符替换成另一个字符。但是，题目并没有给出“ijk”对应的具体结果，因此无法确定“ijk”会转换成什么。

YM

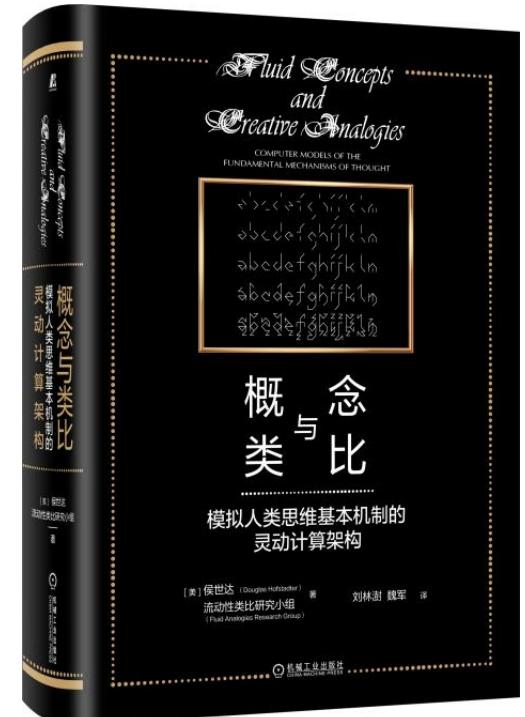
提示你一下，c的下一个字母是d



如果“abc”对应的是“abd”，那么“ijk”对应的可能并没有一个固定的结果。在这种情况下，需



要提供更多的上下文信息或规则才能确定“ijk”对应的值。



Chain of thought

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

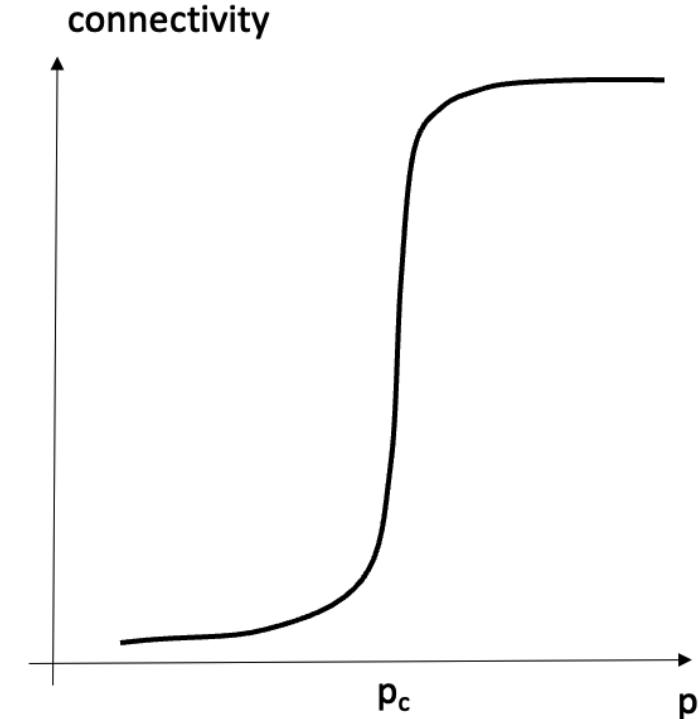
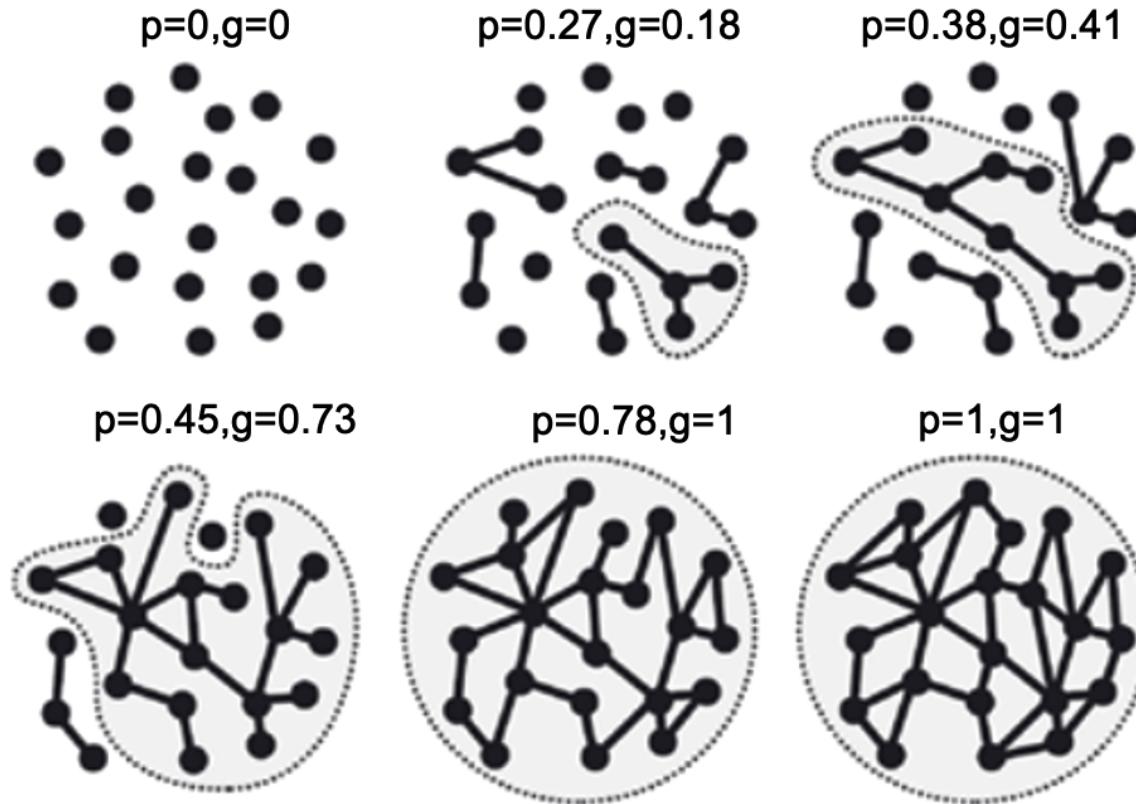
Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. 

From: Chain of thought prompting elicits reasoning in large language models
知乎 @张俊林

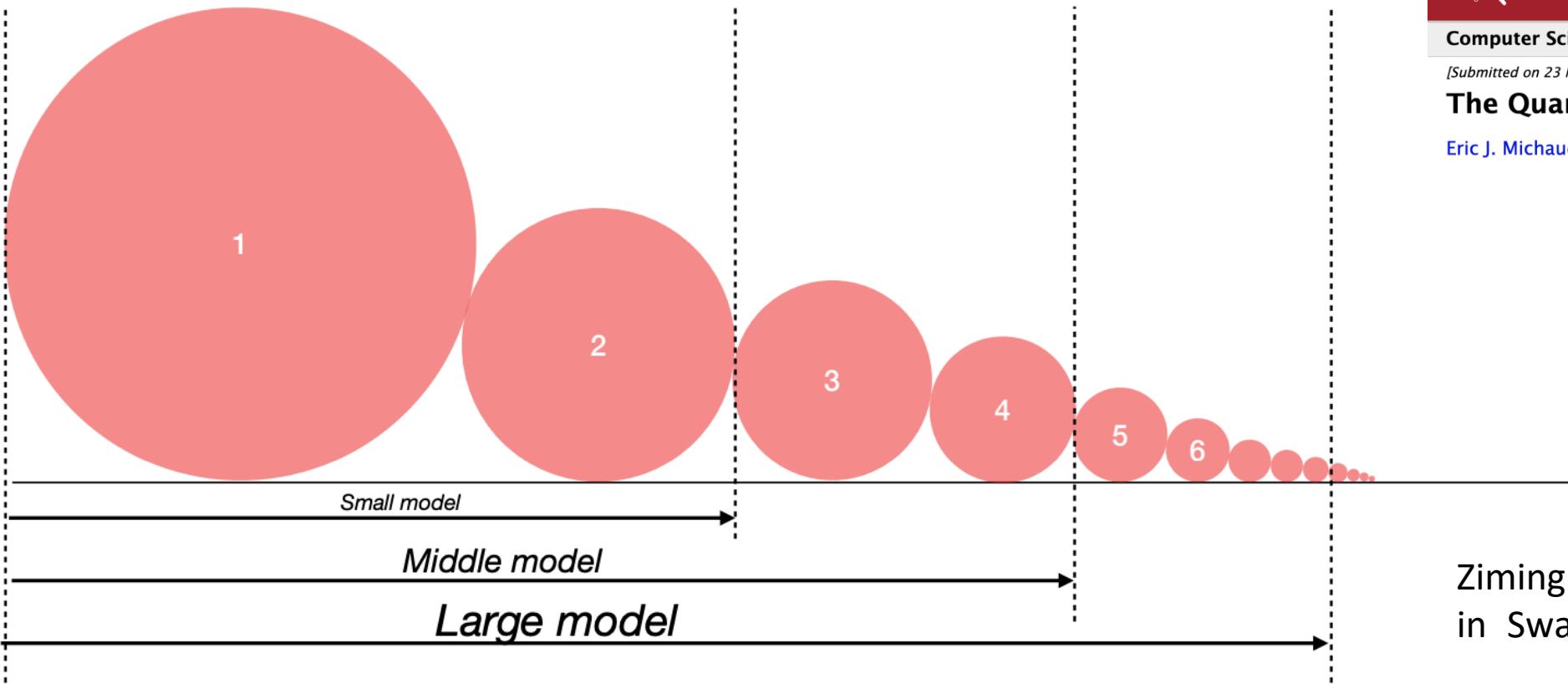


Critical Phase Transition



Knowledge Quanta

Knowledge quanta sequence



Computer Science > Machine Learning

[Submitted on 23 Mar 2023]

The Quantization Model of Neural Scaling

Eric J. Michaud, Ziming Liu, Uzay Girit, Max Tegmark



Ziming Liu, April 2023
in Swarma Club

<https://arxiv.org/pdf/2001.08361.pdf>



Scale is all you need

Scaling Laws for Neural Networks

Jared Kaplan *

Johns Hopkins University, OpenAI
jaredk@jhu.edu

Tom Henighan Tom B. Brown

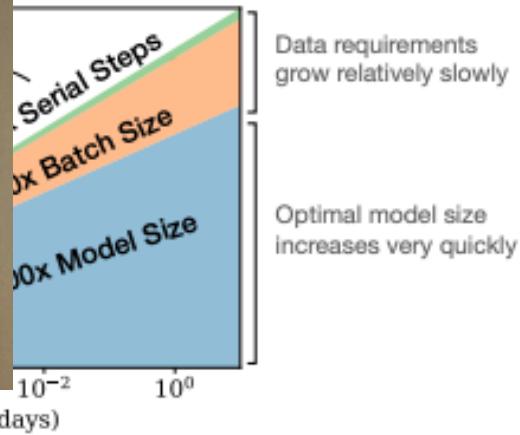
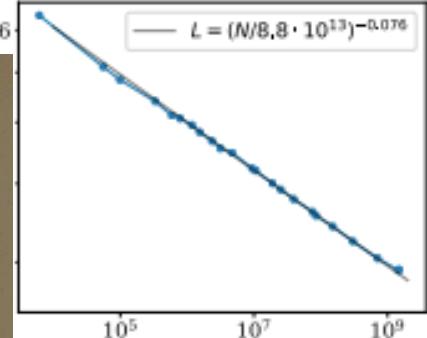
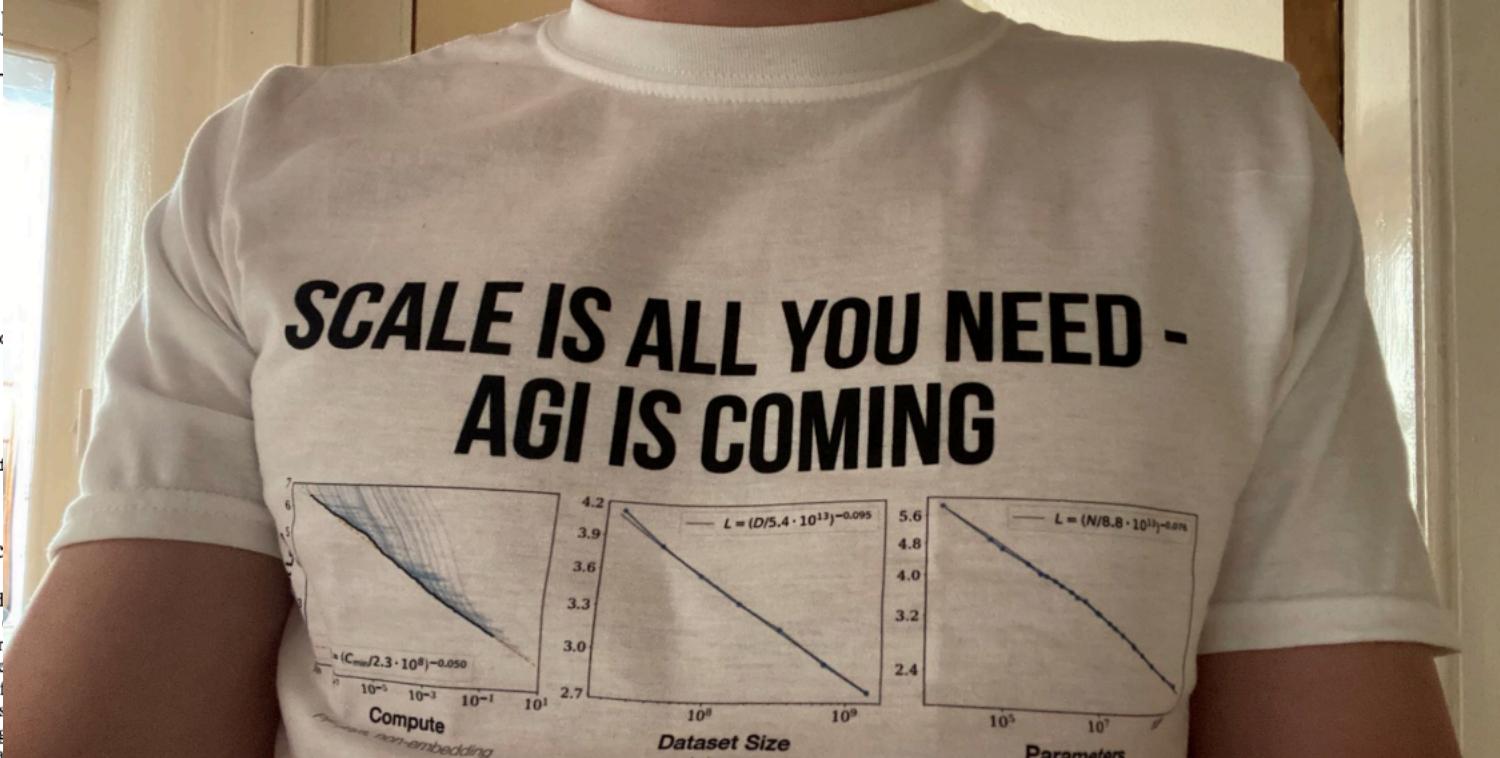
OpenAI OpenAI
henighan@openai.com tom@openai.com b...@openai.com

Scott Gray Alec Radford

OpenAI OpenAI
scott@openai.com alec@openai.com jefi...@openai.com

Abstract

We study empirical scaling laws for language models. The loss scales as a power-law with model size, used for training, with some trends spanning more architectural details such as network width or depth range. Simple equations govern the dependence of the dependence of training speed on model size. These optimal allocation of a fixed compute budget. Large efficient, such that optimally compute-efficient training on a relatively modest amount of data and stopping significantly before convergence.



Large Models & CAS

后 ChatGPT 时代：从通用人工智能到意识机器：3. 复杂适应系统视角下的ChatGPT与AI大模型



展开本季目录

这里展示正在播放的幻灯片页中的术语和资源

全屏播放

本期大纲

截图 字幕

读书会&主讲人介绍

EChatGPT读书会

复杂适应系统视角下的 ChatGPT 与 AI 大模型

00:00:01 术语

建议反馈

QR code

This block displays a video player interface showing a school of fish forming a circular pattern. The video is titled '复杂适应系统视角下的 ChatGPT 与 AI 大模型' (Viewed from the perspective of Complex Adaptive Systems, CAS, on ChatGPT and AI Large Models). It includes a QR code for access.

大纲

- 什么是复杂适应系统 (CAS) ?
 - 复杂系统与复杂科学
 - 复杂适应系统
- 为什么说AI大模型是复杂适应系统?
 - 大语言模型的涌现能力
 - 类比、上下文学习与思维链
 - 大语言模型的规模法则
- 从CAS的视角理解ChatGPT
 - 复杂视角下的神经网络与机器学习
 - 神经网络与动力系统
 - 从Holland的分类器系统到误差反传算法
 - 图灵机与自注意力机制
 - 上下文学习的可能机理
 - 从相变与临界到大模型的涌现
- 为什么说AI大模型可能助力复杂系统研究?
 - 复杂系统自动建模
 - AI可解释性问题
 - 通用大模型与通用原理

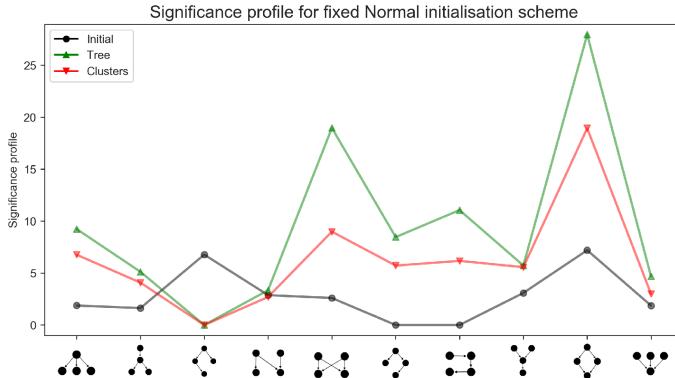
核心概念

- 复杂适应系统 (Complex Adaptive System, CAS)
- 规模法则 (Scaling Laws)
- 动力系统 (Dynamical Systems)
- Transformer
- 梯度反传 (Gradient Backpropagation)
- 语言模型 (Language Model)
- 图灵机 (Turing Machine)
- 相变与临界性 (Phase Transition and Criticality)

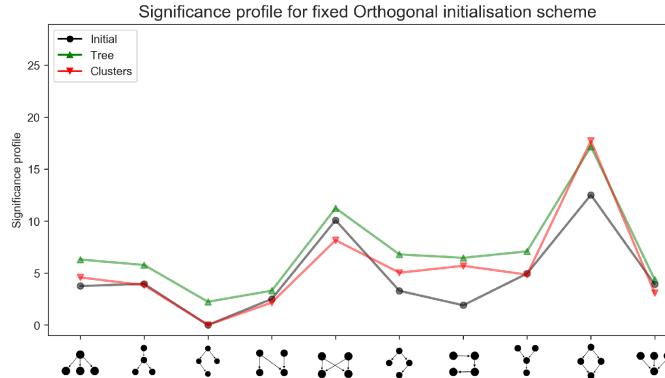
https://pattern.swarma.org/study_group_issue/412



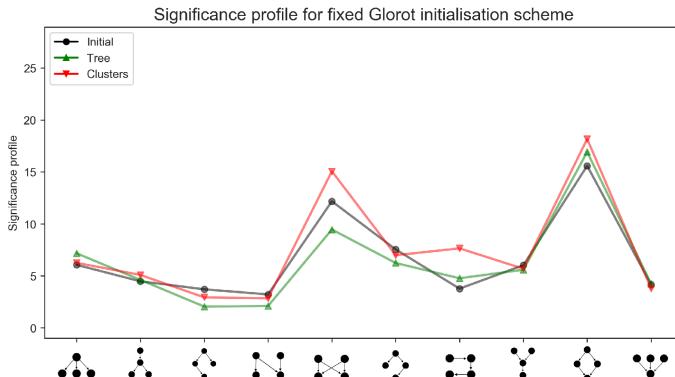
Neural networks as complex networks



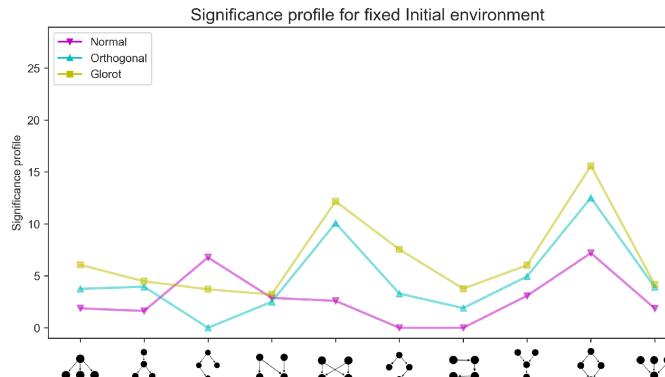
(a) Normal initialization scheme.



(b) Orthogonal initialization scheme.



(c) Xavier initialization scheme.



(d) Initial profile for schemes.

arXiv:1912.12244v1 [nlin.AO] 27 Dec 2019

Emergence of Network Motifs in Deep Neural Networks

Matteo Zambra
Department of Civil, Environmental
and Architectural Engineering
University of Padova
Padova, IT 35131
matteo.zambra@studenti.unipd.it

Alberto Testolin
Department of General Psychology and
Department of Information Engineering
University of Padova
Padova, IT 35131
alberto.testolin@unipd.it

Amos Maritan
Department of Physics and Astronomy and
Istituto Nazionale di Fisica Nucleare
University of Padova
Padova, IT 35131
amos.maritan@unipd.it

Abstract

Network science can offer fundamental insights into the structural and functional properties of complex systems. For example, it is widely known that neuronal circuits tend to organize into basic functional topological modules, called *network motifs*. In this article we show that network science tools can be successfully applied also to the study of artificial neural networks operating according to self-organizing (learning) principles. In particular, we study the emergence of network motifs in multi-layer perceptrons, whose initial connectivity is defined as a stack of fully-connected, bipartite graphs. Our simulations show that the final network topology is primarily shaped by learning dynamics, but can be strongly biased by choosing appropriate weight initialization schemes. Overall, our results suggest that non-trivial initialization strategies can make learning more effective by promoting the development of useful network motifs, which are often surprisingly consistent with those observed in general transduction networks.

1 Introduction

The topological structure of complex networks can be characterized by a series of well-known features,

Training dynamics and diffusion dynamics

Published as a conference paper at ICLR 2021

A DIFFUSION THEORY FOR DEEP LEARNING DYNAMICS: STOCHASTIC GRADIENT DESCENT EXPONENTIALLY FAVORS FLAT MINIMA

Zeke Xie^{1,2}, Issei Sato^{1,2}, and Masashi Sugiyama^{2,1}

¹The University of Tokyo

²RIKEN Center for AIP

xie@ms.k.u-tokyo.ac.jp

{sato,sugij}@k.u-tokyo.ac.jp

ABSTRACT

Stochastic Gradient Descent (SGD) and its variants are mainstream methods for training deep networks in practice. SGD is known to find a flat minimum that often generalizes well. However, it is mathematically unclear how deep learning can select a flat minimum among so many minima. To answer the question quantitatively, we develop a density diffusion theory to reveal how minima selection quantitatively depends on the minima sharpness and the hyperparameters. To the best of our knowledge, we are the first to theoretically and empirically prove that, benefited from the Hessian-dependent covariance of stochastic gradient noise, SGD favors flat minima exponentially more than sharp minima, while Gradient Descent (GD) with injected white noise favors flat minima only polynomially more than sharp minima. We also reveal that either a small learning rate or large-batch training requires exponentially many iterations to escape from minima in terms of the ratio of the batch size and learning rate. Thus, large-batch training cannot search flat minima efficiently in a realistic computational time.

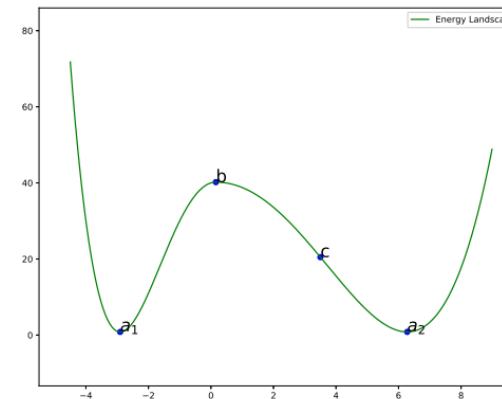
SGD Dynamics. Let us replace η by dt as unit time. Then the continuous-time dynamics of SGD (Coffey & Kalmykov, 2012) is written as

$$d\theta = -\frac{\partial L(\theta)}{\partial \theta} dt + [2D(\theta)]^{\frac{1}{2}} dW_t, \quad (2)$$

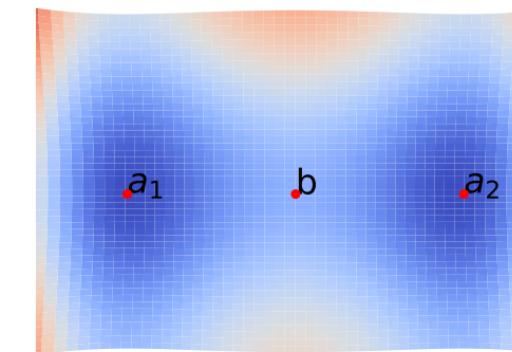
where $dW_t \sim \mathcal{N}(0, Idt)$ and $D(\theta) = \frac{\eta}{2} C(\theta)$. We note that the dynamical time t in the continuous-time dynamics is equal to the product of the number of iterations T and the learning rate η : $t = \eta T$. The associated Fokker-Planck Equation is written as

$$\frac{\partial P(\theta, t)}{\partial t} = \nabla \cdot [P(\theta, t) \nabla L(\theta)] + \nabla \cdot \nabla D(\theta) P(\theta, t) \quad (3)$$

$$= \sum_i \frac{\partial}{\partial \theta_i} \left[P(\theta, t) \frac{\partial L(\theta)}{\partial \theta_i} \right] + \sum_i \sum_j \frac{\partial^2}{\partial \theta_i \partial \theta_j} D_{ij}(\theta) P(\theta, t), \quad (4)$$



(a) 1-Dimensional Escape



(b) High-Dimensional Escape

Neural network and renormalization group

An exact mapping between the Variational Renormalization Group and Deep Learning

Pankaj Mehta
Dept. of Physics, Boston University, Boston, MA

David J. Schwab
Dept. of Physics, Northwestern University, Evanston, IL

Deep learning is a broad set of techniques that uses multiple layers of representation to automatically learn relevant features directly from structured data. Recently, such techniques have yielded record-breaking results on a diverse set of difficult machine learning tasks in computer vision, speech recognition, and natural language processing. Despite the enormous success of deep learning, relatively little is understood theoretically about why these techniques are so successful at feature learning and compression. Here, we show that deep learning is intimately related to one of the most important and successful techniques in theoretical physics, the renormalization group (RG). RG is an iterative coarse-graining scheme that allows for the extraction of relevant features (i.e. operators) as a physical system is examined at different length scales. We construct an exact mapping from the variational renormalization group, first introduced by Kadanoff, and deep learning architectures based on Restricted Boltzmann Machines (RBMs). We illustrate these ideas using the nearest-neighbor Ising Model in one and two-dimensions. Our results suggest that deep learning algorithms may be employing a generalized RG-like scheme to learn relevant features from data.

A central goal of modern machine learning research is to learn and extract important features directly from data. Among the most promising and successful techniques for accomplishing this goal are those associated with the emerging sub-discipline of deep learning. Deep learning uses multiple layers of representation to learn descriptive features directly from training data [1, 2] and has been successfully utilized, often achieving record breaking results, in difficult machine learning tasks including object labeling [3], speech recognition [4], and natural language processing [5].

In this work, we will focus on a set of deep learning algorithms known as deep neural networks (DNNs) [6]. DNNs are biologically-inspired graphical statistical models that consist of multiple layers of “neurons”, with units in one layer receiving inputs from units in the layer below them. Despite their enormous success, it is still unclear what advantages these deep, multi-layer architectures possess over shallower architectures with a similar number of parameters. In particular, it is still not well understood theoretically why DNNs are so successful at uncovering features in structured data. (But see [7–9].)

One possible explanation for the success of DNN architectures is that they can be viewed as an iterative coarse-graining scheme, where each new high-level layer of the neural network learns increasingly abstract higher-level features from the data [1, 10]. The initial layers of the DNN can be thought of as low-level feature detectors which are then fed into higher layers in the DNN

cerned solely with the unsupervised aspect of training DNNs, and the use of DNNs for compression [6].) In what follows, we make this explanation precise.

This successive coarse-graining procedure is reminiscent of one of the most successful and important tools in theoretical physics, the renormalization group (RG) [11, 12]. RG is an iterative coarse-graining procedure designed to tackle difficult physics problems involving many length scales. The central goal of RG is to extract relevant features of a physical system for describing phenomena at large length scales by integrating out (i.e. marginalizing over) short distance degrees of freedom. In any RG sequence, fluctuations are integrated out starting at the microscopic scale and then moving iteratively on to fluctuations at larger scales. Under this procedure, certain features, called relevant operators, become increasingly important while other features, dubbed irrelevant operators, have a diminishing effect on the physical properties of the system at large scales.

In general, it is impossible to carry out the renormalization procedure exactly. Therefore, a number of approximate RG procedures have been developed in the theoretical physics community [12–15]. One such approximate method is a class of variational “real-space” renormalization schemes introduced by Kadanoff for performing RG on spin systems [14, 16, 17]. Kadanoff’s variational RG scheme introduces coarse-grained auxiliary, or “hidden”, spins that are coupled to the physical spin systems through some unknown coupling parame-

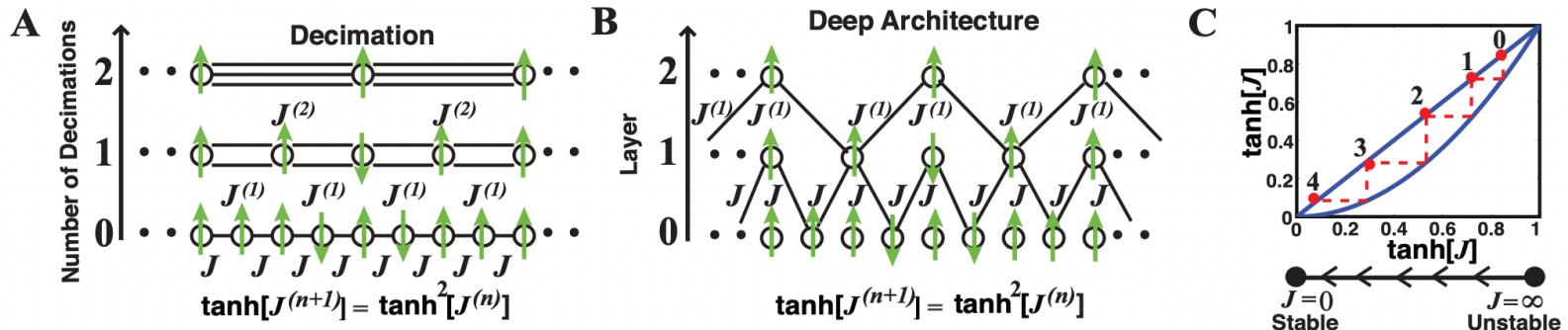


FIG. 2. RG and deep learning in the one-dimensional Ising Model. (A) A decimation based renormalization transformation for the ferromagnetic 1-D Ising model. At each step, half the spins are decimated, doubling the effective lattice spacing. After, n successive decimations, the spins can be described using a new 1-D Ising models with a coupling J^n between spins. Couplings at a given layer are related to couplings at a previous layer through the square of the hyperbolic tangent function. (B) Decimation-based renormalization transformations can also be realized using the deep architecture where the weights between the $n + 1$ and n -th hidden layer are given by J^n . (C) Visualizing the renormalization group flow of the couplings for 1-D Ferromagnetic Ising model. Under four successive decimations or equivalently as we move up four layers in the deep architecture, the couplings (marked by red dots) get smaller. Eventually, the couplings are attracted to stable fixed point $J = 0$.

Neural network and renormalization group

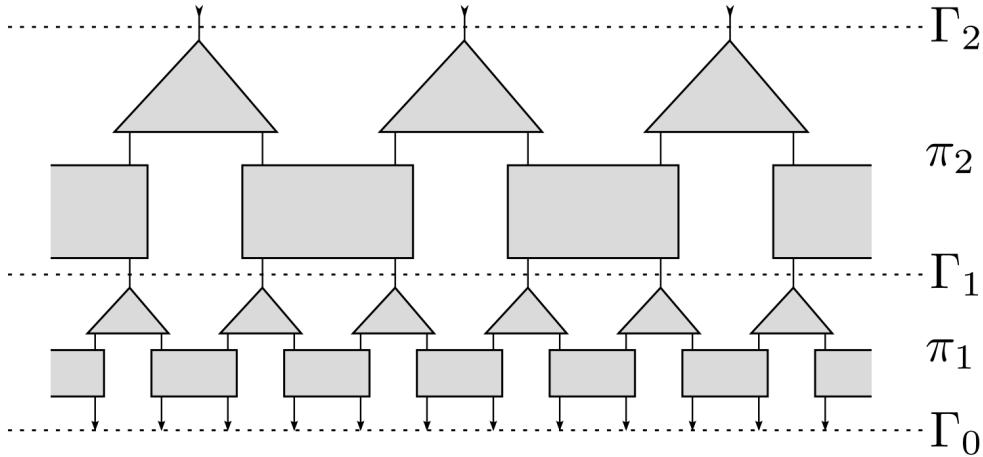


Figure 1: Example of a MERA for a one-dimensional lattice, where only two layers are represented. The boxes and triangles are arbitrary isometric maps between Hilbert spaces represented by the vertical lines. We obtain a CORA by replacing the boxes and triangles by stochastic maps between random variables. This diagram can then be read as an operational recipe to produce the state on the lattice Γ_0 , from that of Γ_2 : each vertical line represents some data, or random variable, and a box or triangle is a specific stochastic map which must be applied to the joint probability distribution of all its input(s) (the lines coming in from above) in order to produce the joint distribution of its outputs (represented by the lines coming out below it). This can be equivalently represented by the Bayesian network of Figure 2. The parameters of the model are the components of all the local stochastic maps (boxes and triangles), together with the initial state of the last layer $\Gamma_{j_{\max}}$.

Deep learning and the renormalization group

Cédric Bény
Institut für Theoretische Physik
Leibniz Universität Hannover
Appelstraße 2, 30167 Hannover, Germany
cedric.beny@gmail.com

Abstract

Renormalization group (RG) methods, which model the way in which the effective behavior of a system depends on the scale at which it is observed, are key to modern condensed-matter theory and particle physics. We compare the ideas behind the RG on the one hand and deep machine learning on the other, where depth and scale play a similar role. In order to illustrate this connection, we review a recent numerical method based on the RG—the multiscale entanglement renormalization ansatz (MERA)—and show how it can be converted into a learning algorithm based on a generative hierarchical Bayesian network model. Under the assumption—common in physics—that the distribution to be learned is fully characterized by local correlations, this algorithm involves only explicit evaluation of probabilities, hence doing away with sampling.

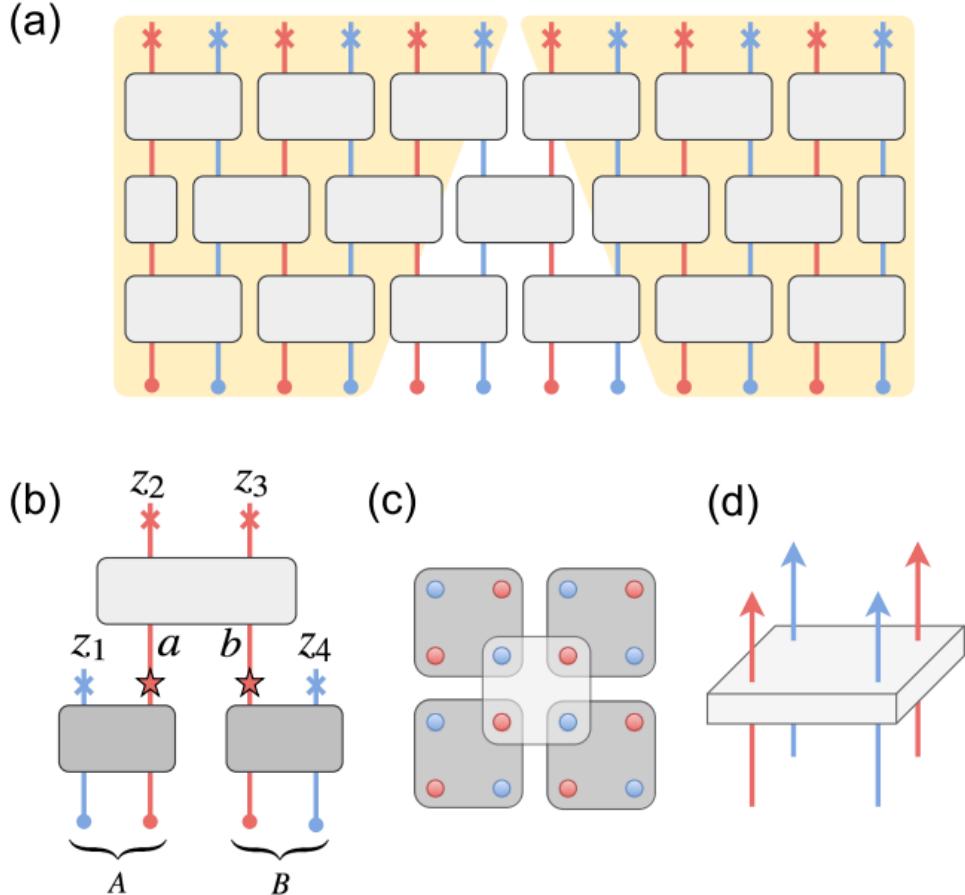
Natural sciences extend the process by which we intuitively discover patterns in our sensory data to domains beyond our natural sensory abilities, but also beyond our intuitive reasoning abilities. Already, machine learning methods are becoming important in fields where the patterns are too complex to be modeled with simple equations, and too far removed from intuition to be comprehended without formal tools.

Physics may seem an unlikely candidate for application of machine learning, given that it specifically focuses on systems which can be understood with simple laws. Nevertheless, these very compact laws have to be compared to experiments, and hence require a “decompression” process which rapidly leads to overwhelming levels of complexity.

The *renormalization group* (RG) was designed to handle some of these complexities and had a tremendous influence in particle physics and condensed-matter physics [1]. It models the dependence of a system’s effective behavior on a parameter which is usually thought of as *scale* or *energy*, but which we will think of here as *depth*. Although the RG idea is not recent, new approaches have emerged which have the potential to fully automatize it [2–5]. These methods are designed for simulation rather than learning, but, because they are defined through clever representations of certain classes of (quantum) states, they can be adapted to learning tasks as well [6, 7].

In this paper, I will explain how one of these methods (the multiscale entanglement renormalization ansatz [8]) can be made into a deep learning algorithm for classical probability distributions.

Neural network renormalization group



Neural Network Renormalization Group

Shuo-Hui Li^{1,2} and Lei Wang^{1,3,*}

¹Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³Songshan Lake Materials Laboratory, Dongguan, Guangdong 523808, China

We present a variational renormalization group (RG) approach based on a reversible generative model with hierarchical architecture. The model performs hierarchical change-of-variables transformations from the physical space to a latent space with reduced mutual information. Conversely, the neural network directly maps independent Gaussian noises to physical configurations following the inverse RG flow. The model has an exact and tractable likelihood, which allows unbiased training and direct access to the renormalized energy function of the latent variables. To train the model, we employ probability density distillation for the bare energy function of the physical problem, in which the training loss provides a variational upper bound of the physical free energy. We demonstrate practical usage of the approach by identifying mutually independent collective variables of the Ising model and performing accelerated hybrid Monte Carlo sampling in the latent space. Lastly, we comment on the connection of the present approach to the wavelet formulation of RG and the modern pursuit of information preserving RG.

The Renormalization group (RG) is one of the central schemes in theoretical physics, whose impacts span from high-energy [1] to condensed matter physics [2, 3]. In essence, RG keeps the relevant information while reducing the dimensionality of statistical data. Besides its conceptual importance, practical RG calculations have played important roles in solving challenging problems in statistical and quantum physics [4, 5]. A notable recent development is to perform RG calculations using tensor network machinery [6–18].

The relevance of RG goes beyond physics. For example, in deep learning applications, the inference process in image recognition resembles the RG flow from microscopic pixels to categorical labels. Indeed, a successfully trained neural network extracts a hierarchy of increasingly higher-level concepts in its deeper layers [19]. In light of such intriguing similarities, Refs [20–23] drew connections between deep learning and the RG. Ref. [24] proposed an RG scheme based on mutual information maximization, Ref. [25] employed deep learning to study holography duality, and Ref. [26] examined the adversarial examples from a RG perspective. Since the discussions are not totally uncontroversial [21, 23, 24, 27, 28], it remains highly desirable to establish a more concrete, rigorous, and constructive connection between RG and deep learning. Such a connection will not only bring powerful deep learning techniques into solving complex physics problems but also benefit theoretical understanding of deep learning from a physics perspective.

In this paper, we present a neural network based variational RG approach (NeuralRG) for statistical physics prob-

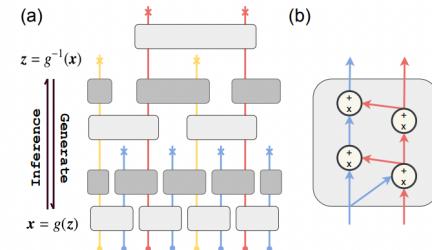


Figure 1. (a) The NeuralRG network is formed by stacking bijector networks into a hierarchical structure. The solid dots at the bottom are the physical variables x and the crosses are the latent variables z . Each block is a bijector. The light gray and the dark gray blocks are the disentanglers and the decimators respectively. The RG flows from bottom to top, which corresponds to the inference of the latent variables conditioned on the physical variables. Conversely, one can directly generate physical configurations by sampling the latent variables according to the prior distribution and passing them downwards through the network. (b) The internal structure of the bijector block consists of Normalizing Flows [32].

(variational, exact and tractable likelihood, principled structure design via information theory) and high computational efficiency. The NeuralRG approach is closer in spirit to the

Splendid future

- AI for complex systems
 - Automated modelling
 - Optimization
 - Control
 - ...
- Complexity science for AI
 - Scaling laws
 - Emergence
 - Renormalization & statistical physics
 - Complex networks
 - ...

