Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com

Abstract

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.

*Equal contribution.

Contributions: Jared Kaplan and Sam McCandlish led the research. Tom Henighan contributed the LSTM experiments. Tom Brown, Rewon Child, and Scott Gray, and Alec Radford developed the optimized Transformer implementation. Jeff Wu, Benjamin Chess, and Alec Radford developed the text datasets. Dario Amodei provided guidance throughout the project.

Contents

1 Introduction

Language provides a natural domain for the study of artificial intelligence, as the vast majority of reasoning tasks can be efficiently expressed and evaluated in language, and the world's text provides a wealth of data for unsupervised learning via generative modeling. Deep learning has recently seen rapid progress in language modeling, with state of the art models [RNSS18, DCLT18, YDY+19, LOG+19, RSR+19] approaching human-level performance on many specific tasks [WPN+19], including the composition of coherent multi-paragraph prompted text samples [RWC+19].

One might expect language modeling performance to depend on model architecture, the size of neural models, the computing power used to train them, and the data available for this training process. In this work we will empirically investigate the dependence of language modeling loss on all of these factors, focusing on the Transformer architecture [VSP+17, LSP+18]. The high ceiling and low floor for performance on language tasks allows us to study trends over more than seven orders of magnitude in scale.

Throughout we will observe precise power-law scalings for performance as a function of training time, context length, dataset size, model size, and compute budget.

1.1 Summary

Our key findings for Transformer language models are are as follows:

[2] Here we display predicted compute when using a sufficiently small batch size. See Figure 13 for comparison to the purely empirical data.