Deepformers: Training Very Deep Transformers via Dynamical Isometry

Thomas Bachlechner
Department of Physics, UC San Diego
tbachlechner@physics.ucsd.edu

Bodhisattwa Prasad Majumder
CSE, UC San Diego
bmajumde@eng.ucsd.edu

Huanru Henry Mao
CSE, UC San Diego
hhmao@eng.ucsd.edu

Julian McAuley
CSE, UC San Diego
jmcauley@eng.ucsd.edu

1 Motivation

Neural networks have achieved increasing performance by leveraging the exponential representational capacity of deeper models, with some using thousands of layers [1]. This trend has dominated many convolutional neural networks but has yet to dominate NLP architectural design, such as Transformers [2]. The Transformer self-attention architecture that achieves state of the art performance in many NLP tasks [3] usually has less than 24 layers and we find that trying to train deeper models lead to either convergence difficulties or slow training times.

The theoretical study of neural networks with random parameters has revealed a maximum penetration depth that depends on the initialization scheme and the model architecture, which limits the number of layers that can be effectively trained. A network that allows an infinitesimal input perturbation to propagate to the output layer unimpeded in magnitude is said to satisfy the property of dynamical isometry [4]. Fully connected [5, 6], convolutional [1] and recurrent [7] networks can be initialized to satisfy dynamical isometry, which enables effective training of deep models.

In this paper, we aim to leverage dynamical isometry to construct viable deep Transformer-architecture inspired models — Deepformers — specially targeted for generative modeling tasks. As we analyze signal propagation through Transformers, we find two components prohibit dynamical isometry in these models: (1) self-attention (unmasked) and (2) Layer Normalization [8] allowing only a low-dimensional subspace of the input signal to propagate, rendering dynamical isometry impossible. We propose a simple modification to the standard Transformer architecture [2] and show that it enables us to train much deeper networks for downstream language modeling tasks as compared to the standard ones.

2 Theoretical results

In order for a model to be trainable via gradient descent, a signal has to be able to propagate forward from the input to the output layer, as well as backward in order to evaluate the gradients. This notion has been formalized with the study of mean field theory governing deep and wide neural networks with random parameters [9]. A central quantity is the input-output Jacobian $J_{io}$ defined as

$$J_{io} \equiv \frac{\partial x^L}{\partial h^0},$$

where $x^L$ are the outputs of the $L$th layer, and $h^0$ are the pre-activations the 0th layer, i.e. the input signal. The singular values $\lambda_{io}$ quantify whether a given perturbation of an input signal is enhanced or

*denotes equal contributions

Preprint. Under review.
attenuated as it propagates through the network. For small singular values, \( \lambda_{io} \ll 1 \), a deep network approaches an ordered phase in which perturbations decay, while for large singular values \( \lambda_{io} \gg 1 \) a deep network approaches a chaotic phase in which the outputs are unpredictable. In either of those limiting scenarios, training becomes impossible. It is at the boundary between the ordered and the chaotic phase, \( \lambda_{io} \approx 1 \), where training proceeds efficiently. At this critical point the signal penetration depth diverges, signals propagate through the network with unchanged magnitude and training of arbitrarily deep networks becomes possible. This property is referred to as dynamical isometry.

We examine the singular value distribution of the input-output Jacobian for vanilla Transformer encoder networks of varying depths at initialization, illustrated in Figure 1a. While shallow Transformers exhibit a singular value distribution peaked around unity, all singular values rapidly decrease with increasing depth of the network. We observe that most singular values vanish to machine precision already for encoders of depth 64, despite the residual connections. This is consistent with the empirical observation that deep Transformer architectures are extremely challenging to train.

### 3 Experimental results

Based on this theory, we propose a novel architecture modification that is designed to enable signal propagation through the network. At each residual add operation of the vanilla Transformer, we incorporate a learnable scaling parameter that is initialized to zero. At initialization, each encoder Transformer layer performs the identity function, and so the networks trivially satisfy dynamical isometry. During training the weight of the residual connection gradually increases, allowing the encoder to model extremely complex functions while maintaining signal propagation properties close to dynamical isometry. We illustrate this behavior in Figure 1b, which shows the singular value distribution of a 64 layer Deepformer model during training.

We performed a comparison of a 64 layer Deepformer with the vanilla Transformer on language modeling for WikiText 2 [10] and find that the vanilla Transformer diverges quickly and is unable to converge while Deepformer converges well. To show the expressivity of depth, we also compared the performance of Deepformer by varying its depth and width fixed a fixed 18M parameter count, and find that deeper variants generalize better. Our preliminary results in Figure 2 indicate that Deepformers have much higher representational capacity and can increase parameter efficiency.
References


