Training ultra-deep CNNs with critical initialization

Lechao Xiao[*], Yasaman Bahri [*], Sam Schoenhol & Jeffrey Pennington
Google Brain
{xlc, yasamanb, schsam, jpennin}@google.com

Abstract

In recent years, state-of-the-art methods in computer vision have utilized increasingly deep convolutional neural network architectures (CNNs), with some of the most successful models employing 1000 layers or more. Optimizing networks of such depth is extremely challenging and has up until now been possible only when the architectures incorporate special residual connections and batch normalization. In this work, we demonstrate that it is possible to train vanilla CNNs of depth 1500 or more simply by careful choice of initialization. We derive this initialization scheme theoretically, by developing a mean field theory for the dynamics of signal propagation in random CNNs with circular boundary conditions. We show that the order-to-chaos phase transition of such CNNs is similar to that of fully-connected networks, and we provide empirical evidence that ultra-deep vanilla CNNs are trainable if the weights and biases are initialized near the order-to-chaos transition.

1 Introduction

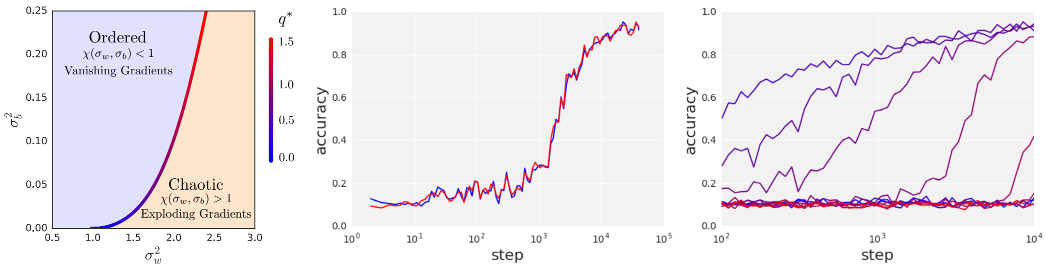
Deep convolutional neural networks (CNNs) are one of the pillars of modern deep learning, enabling unprecedented accuracy in domains ranging across computer vision [1], speech recognition [2], natural language processing [3, 4, 5], and recently even the board game Go [6, 7].

As computational resources continue to increase, so too does the network depth of the state-of-the-art models. Some of the best performing models on ImageNet [8], for example, have employed hundreds or even a thousand layers [9, 10]. But up until now, ultra-deep architectures such as these have been trainable only in conjunction with techniques like residual connections [9] and batch normalization [11]. It is an open question, therefore, whether these techniques are improving the model or merely improving our ability to train them. In this work, we investigate the training of ultra-deep vanilla CNNs in order to begin shedding light on this question.

We study the problem of training ultra-deep CNNs in the framework of signal propagation, following closely the work of [12, 13]. Those works focused exclusively on the fully-connected setting, and our main contribution is to extend the analysis to convolutional architectures. Owing to their complicated structure, it is not clear *a priori* whether CNNs will be amenable to the same type of analysis that fully-connected networks enjoyed. Indeed, in the convolutional neural networks, there are at least two additional length scales that are relevant: the input size and the kernel size. In principle, these two parameters can affect the way signals propagate through the network, even as the number of filters become infinite. In light of that observation, it is perhaps surprising that we find that these two quantities are actually irrelevant if the boundary conditions of the convolution are periodic. In this case, the mean field theory of signal propagation for CNNs turns out to be strikingly similar to that of fully-connected networks.

To test our theory, we train very deep CNNs on MNIST with a large grid of initialization parameters, and we find excellent agreement between our theoretical predictions and empirical observations of which configurations are trainable. In particular, we find that extremely deep (1500-layer) vanilla

[*]Work done as part of the Google Brain Residency program (g.co/brainresidency).



(a) Phase Diagram (b) Trainability (c) Trainability

Figure 1: (a). Phase diagram showing the ordered (with vanishing gradients) and chaotic phases (with exploding gradients) separated by a transition. The variation in value of q^* along the critical line is shown in color. (b). Training (blue curve) and test accuracy of a 1503-layer CNN initialized critically with $\sigma_b^2 = 2 \times 10^{-5}$ and $\sigma_w^2 = 1.05$. (c). For fixed bias variance $\sigma_b^2 = 2 \times 10^{-5}$, we examine the trainability of different weight variances. The five trainable curves from left to right: $\sigma_w^2 = 1, 1.05, 1.1, 1.15, 1.20$. The bottom ones are untrainable with $\sigma_w^2 = 0.85, 0.90, 0.95, 1.25, 1.30, 1.35$.

CNNs are trainable if the initial parameters are chosen to be "critical", i.e. exactly such that the network is neither in the chaotic nor ordered phase. We find empirically that our results also apply to standard zero-padded boundary conditions.

2 Mean field theory for CNNs

Consider an L-layer 1D[2] periodic CNN with filter size $2k+1$, channel size c, spatial size n, per-layer weight tensors $\omega \in \mathbb{R}^{(2k+1)\times c \times c}$ and biases $b \in \mathbb{R}^c$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be the activation function and let $h_j^l(\alpha)$ denote the pre-activation at layer l, channel j, and spatial location α. The forward-propagation dynamics can be described by the recurrence relation,

$$h_j^{l+1}(\alpha) = \sum_{i=1}^{c} \sum_{\beta=-k}^{k} x_i^l(\alpha+\beta)\omega_{ij}^{l+1}(\beta) + b_j^{l+1}\,, \quad x_j^l(\alpha) = \phi(h_j^l(\alpha)). \tag{1}$$

At initialization, we take the weights ω_{ij}^l to be drawn i.i.d. from the Gaussian $\mathcal{N}(0, \sigma_\omega^2/(c(2k+1)))$ and the biases b_j^l to be drawn i.i.d. from the Gaussian $\mathcal{N}(0, \sigma_b^2)$. Note that $x_i^l(\alpha) = x_i^l(\alpha+n) = x_i^l(\alpha-n)$ since we assume the circular boundary conditions. We wish to understand how signals propagate through these networks as the number of channels $c \to \infty$. For this purpose, we need to understand how the covariance matrices of the pre-activations evolve as the depth increases.

For each j and l, the central limit theorem implies that the pre-activation vectors h_j^l are i.i.d. Gaussian with mean zero and covariance $\Sigma^l = \{q_{\alpha,\alpha'}^l\}_{\alpha,\alpha'}$, where $q_{\alpha,\alpha'}^l = \mathbb{E}\left[h_j^l(\alpha)h_j^l(\alpha')\right]$. Moreover, since $\{h_j^l(\alpha)\}_j$ are also i.i.d. Gaussian,

$$q_{\alpha,\alpha}^{l+1} \approx \frac{1}{c}\sum_j (h_j^{l+1}(\alpha))^2 = \sigma_b^2 + \frac{\sigma_w^2}{(2k+1)} \sum_{\beta=-k}^{k} \mathbb{E}\left[(x_j^l(\alpha+\beta))^2\right]. \tag{2}$$

Next,we define an average operator, $\mathcal{A}(f(\alpha)) = \sum_{\beta=-k}^{k} f(\alpha+\beta)/(2k+1)$, and introduce the $\mathcal{Q}$-map defined as the variance propagation operator of fully-connected networks [12],

$$\mathcal{Q}(q) = \sigma_\omega^2 \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\phi(\sqrt{q}z)^2\right] + \sigma_b^2. \tag{3}$$

In terms of these operators, the above equation is characterized by the "average Q-map", i.e. the $\mathcal{A}Q$-map, $q_{\alpha,\alpha}^{l+1} = \mathcal{A} \circ \mathcal{Q}(q_{\alpha,\alpha}^l)$. Similarly, the correlation between two pixels within the same channel can be obtained by the "average C-map" or the $\mathcal{A}C$-map, $q_{\alpha,\alpha'}^{l+1} = \mathcal{A} \circ \mathcal{C}(q_{\alpha,\alpha'}^l)$, where the $\mathcal{C}$-map is the iterative correlation operator defined in the fully-connected context [12], and is given by

$$\mathcal{C}(q_{\alpha,\alpha'}) = \sigma_\omega^2 \mathbb{E}\left[\phi(z_1)\phi(z_2)\right] + \sigma_b^2, \tag{4}$$

[2]For notational simplicity, we consider 1D convolutions, but the 2D case proceeds identically.

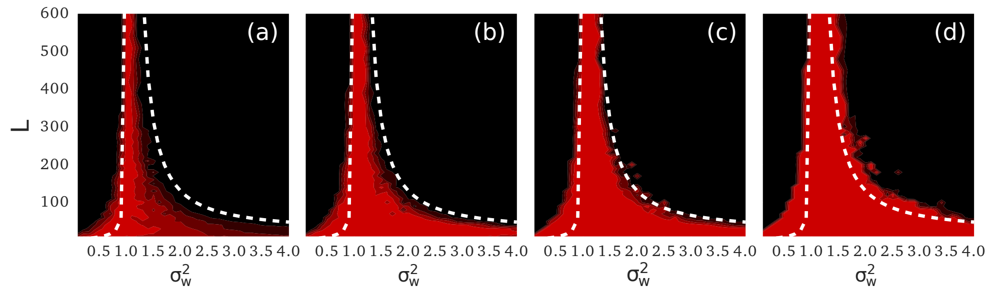


Figure 2: For fixed bias variance $\sigma_b^2 = 2 \times 10^{-5}$, we examine the training accuracy obtainable for a given depth L network and weight variance σ_w, after (a) 500, (b) 2500, (c) 10000, and (d) 100000 training steps. Also plotted (white dashed line) is a multiple ($6\xi_c$) of the characteristic depth scale governing convergence to the fixed point.

with $(z_1, z_2)^T$ drawn from a mean zero Gaussian with covariance matrix $[[q_{\alpha,\alpha}, q_{\alpha,\alpha'}], [q_{\alpha,\alpha'}, q_{\alpha',\alpha'}]]$. Since the $\mathcal{Q}$-map is a special case of the $\mathcal{C}$-map when $\alpha = \alpha'$ [12], the recurrence relation for the covariance matrices can be written compactly as,

$$\Sigma^{l+1} = \{q^{l+1}_{\alpha,\alpha'}\}_{\alpha,\alpha'} = \{\mathcal{A} \circ \mathcal{C}(q^l_{\alpha,\alpha'})\}_{\alpha,\alpha'} = \mathcal{A} \circ \mathcal{C}(\Sigma^l). \tag{5}$$

Note that this equation completely characterizes the mean-field dynamics of convolutional networks. We will now be concerned with understanding its properties.

Since equ. (5) involves both the $\mathcal{A}$ operator and the $\mathcal{C}$ operator, we will study their effect on the mean-field dynamics separately. The $\mathcal{C}$ operator was also found in the fully-connected setting and we will therefore begin by briefly summarizing its properties. For any choice of σ_w and σ_b with monotonically bounded ϕ, the $\mathcal{Q}$-map has a fixed point q^* and the $\mathcal{C}$-map has a fixed point $c^* q^*$ with $c^* = 1$. Here we define c^* as the fixed point correlation between two inputs. It is defined as the fixed point of the the normalized $\mathcal{C}$-map, referred to as the $\mathcal{C}^*$-map,

$$\mathcal{C}^*(c) = (\sigma_\omega^2 \mathbb{E}\left[\phi(z_1)\phi(z_2)\right] + \sigma_b^2)/q^*, \tag{6}$$

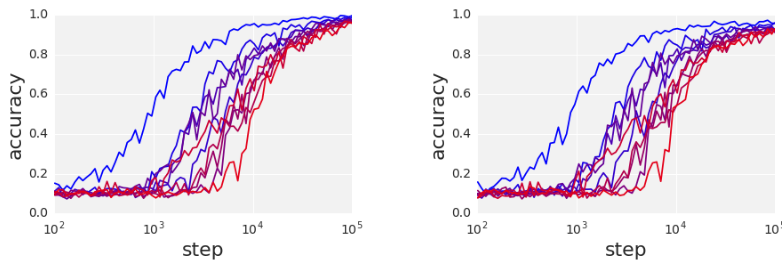
with $(z_1, z_2)^T$ drawn from a mean zero Gaussian with covariance matrix $[[q^*, cq^*], [cq^*, q^*]]$.

To analyze the stability of the $c^* = 1$ fixed point, we define χ_1 to be the derivative of $\mathcal{C}^*$ at c. When $\chi_1 < 1$ the $c^* = 1$ fixed point is stable and asymptotically, dissimilar inputs become increasingly similar until they eventually become indistinguishable. This is known as the ordered phase and in it gradients vanish exponentially with depth [13]. By contrast, when $\chi_1 > 1$, the $c^* = 1$ fixed point is unstable and similar inputs become increasingly dissimilar as they evolve through the network. This is known as the chaotic phase and here gradients explode exponentially with depth. Thus, the curve $\chi_1 = 1$ acts as a boundary between the ordered and chaotic phases. See Fig. 1a for a schematic.

In both the ordered and chaotic regimes it has been shown that information cannot travel through the network as $L \to \infty$ and that the ability of a network to propagate information dictates whether or not it may be trained. This was made precise in[13], who showed that information can flow through a network up to a depth ξ_c, where $\xi_c^{-1} = -\log \chi_c$ Note that by construction, $\xi_c \to \infty$ along the critical line and so information can propagate at all depths in critical networks.

To extend the fully-connected results to the CNN setting, note that eq. (1) implies that if q^* and $c^* q^*$ are fixed points of the $\mathcal{Q}$-map and the $\mathcal{C}$-map respectively, then they are also fixed points of the $\mathcal{A}\mathcal{Q}$-map and $\mathcal{A}\mathcal{C}$-map, respectively. This implies the $\mathcal{A}\mathcal{C}$-map of the covariance matrices has a fixed point $\Sigma^* = \{q^*_{\alpha,\alpha'}\}$, where $q^*_{\alpha,\alpha} = q^*$ and $q^*_{\alpha,\alpha'} = c^* q^*$ for all $\alpha \neq \alpha'$. Note that this covariance matrix has the diagonal-and-offdiagonal structure, which is a consequence of the circular boundary conditions and which is not expected to hold for open boundary conditions.

It is also clear that this fixed point Σ^* is stable given c^* is a stable fixed point of the $\mathcal{C}^*$-map. Thus, we expect the order-to-chaos phase diagram of CNNs to contain a sub-diagram that is quite similar to that of the fully-connected neural networks, though the full diagram of CNNs might be more complicated. Surprisingly, these results are independent of both the kernel size $(2k+1)$ $(k \geq 1)$ and the spatial size n so long as the channel size c is sufficiently large. Thus we conjecture that, similar to the fully-connected setting, a random CNN is trainable so long as $L = O(\xi_c)$.



(a) Training (b) Test

Figure 3: (a). and (b). Training and test accuracy for a 1003-layer networks with varying initialization near the critical line. From the blue to the red curves, the variance of the bias increases. Note that the learning speeds up as q^* decreases.

3 Experiments

The preceding analysis predicts that extremely deep convolutional neural networks ought to be trainable provided they are initialized critically. To test this prediction, we trained a 1503-layer vanilla convolutional network on MNIST with $\phi = \tanh$, $\sigma_w^2 = 1.05$, $\sigma_b^2 = 2 \times 10^{-5}$, using ADAM with a learning rate of 1.2×10^{-7}. The training curve for this network is shown in fig. 1b. We see that, as expected, this ultradeep network is trainable with critical initialization.

Additionally, we expect these very deep networks to quickly become untrainable as the network is initialized further from the critical point. We present evidence supporting this prediction in fig. 1c where we attempt to train networks of depth 1003 varying σ_w^2 for fixed σ_b^2 chosen so that the critical line crosses $\sigma_w^2 = 1.05$. We see that for σ_w^2 sufficiently close to the critical point we are able to train these ultra deep networks relatively quickly and the accuracy increases over the course of training. However, as we move even slightly away from the critical point, the networks quickly cannot be trained and the accuracy stays fixed at 10% over 10^4 training steps.

Our analysis gives a prediction for precisely when this crossover between trainability and untrainability will occur. In particular, we predict that the network ought to be trainable provided $L \lesssim \xi_c$. To test this, we train a large number of convolutional neural networks whose depth varies between $L = 10$ and $L = 600$ and whose weights are initialized with $\sigma_w^2 \in [0, 4]$. In fig. 2 we plot - using a heatmap - the training accuracy obtained by these networks after different numbers of steps. Additionally we overlay the depth scale predicted by our theory, ξ_c. We find strikingly good agreement between our theory of random networks and the results of our experiments.

Finally, we show evidence that there is interesting behavior that occurs beyond the level of mean field theory. To do this we train a number of different 1003-layer convolutional neural networks for different (σ_w, σ_b) pairs near the critical line. The training curves for these different initializations are shown in fig. 3. From our mean-field analysis we know that the average magnitude of gradients will be constant along this line. However, we see that different critical initializations display relatively large differences in their learning dynamics. Understanding this behavior will likely take us beyond mean field theory and we leave it as an interesting avenue for future work.

4 Discussion

In this short note, we derive a mean field theory for the dynamics of signal propagation in random CNNs with circular boundary conditions. We show that the order-to-chaos phase transition of such CNNs is similar to that of the fully-connected networks and we verify the trainable depth scale empirically. We also provide empirical evidence that ultra-deep (1500-layer) vanilla CNNs are trainable if the weights and biases are initialized critically. However, there is still much work to be done beyond trainability. For instance, how the initialization scheme also influences other aspects, such as learning dynamics and generalization, is an interesting subject we wish to address in the near future.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a: Overview of mini–batch gradient descent.

[3] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[4] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[5] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[6] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 01 2016.

[7] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[12] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. *NIPS*, 2016.

[13] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep Information Propagation. *ICLR*, 2017.