Spontaneous Symmetry Breaking in Deep Neural Networks: a critical analysis

Yoni Kahn

Hep-AI group, 12/21/17

Big bold claim

Deep neural networks are quantum field theories, and they learn by spontaneous symmetry breaking

My interpretation

Some very special (possibly only linear) Deep neural networks ~~are~~ have an EFT description as ~~quantum~~ statistical/thermal field theories, and ~~they learn by~~ with a particular learning rate schedule one could possibly observe some form of spontaneous symmetry breaking

Not the only one…

Difficult to parse

ICLR 2018 Conference Paper11 AnonReviewer1

01 Dec 2017 (~~modified: 02 Dec 2017~~) ICLR 2018 Conference Paper11 Official Review readers: everyone

Rating: 3: Clear rejection

Hard to follow

ICLR 2018 Conference Paper11 AnonReviewer3

26 Nov 2017 (~~modified: 02 Dec 2017~~) ICLR 2018 Conference Paper11 Official Review readers: everyone

Rating: 3: Clear rejection

Review: The paper makes a mathematical analogy between deep neural networks and quantum field theory, and claims that this explains a large number of empirically observed phenomena.

I have a solid grasp of the relevant mathematics, and a superficial understanding of QFT, but I could not really make sense of this paper. The paper uses mathematics in a very loose manner. This is not always bad (an overly formal treatment can make a paper hard to read), but in this case it is not clear to me that the results are even "correct modulo technicalities" or have much to do with the reality of what goes on in deep nets.

Do we really need quantum field theory?

ICLR 2018 Conference Paper11 AnonReviewer2

26 Nov 2017 (~~modified: 02 Dec 2017~~) ICLR 2018 Conference Paper11 Official Review readers: everyone

Rating: 3: Clear rejection

Review: In this paper, an number of very strong (even extraordinary) claims are made:

* The abstract promises "a framework to understand the unprecedented performance and robustness of deep neural networks using field theory."
* Page 8 states that this is "This is a first attempt to describe a neural network with a scalar quantum field theory."
* Page 2 promises the use of the "Goldstone theorem" (no less) to understand phase transition in deep learning
* It also claim that many "seemingly different experimental results can be explained by the presence of these zero eigenvalue weights."
* Three important results are stated as "theorem", with a statement like "Deep feedforward networks learn by breaking symmetries" proven in 5 lines, with no formal mathematics.

These are extraordinary claims, but when reaching page 5, one sees that the basis of these claims seems to be the Lagrangian of a simple phi-4 theory, and Fig. 1 shows the standard behaviour of the so-called mexican hat in physics, the basis of the second-order transition. Given physicists have been working on neural network for more than three or four decades, I am surprise that this would enough to solve all these problems!

However, there are germs of good ideas, so let's see where they lead

Compact description of DNN's

$$\mathbf{y}_t = \left(\prod_{n=0}^{t-1} R_{t-n} \mathbf{W}_{t-n} \right) \mathbf{x}_1$$

Output (layer t)

Activation (generally nonlinear)

Weights/biases (affine matrix)

Input

If R = 1 this is a linear network

Symmetries

$$\mathbf{y}_t = \left(\prod_{n=0}^{t-1} R_{t-n} \mathbf{W}_{t-n} \right) \mathbf{x}_1$$

$$\mathbf{x}_1 \to Q_1 \mathbf{x}_1$$

$$\mathbf{W}_t \to Q_t \mathbf{W}_t Q_t^{-1}$$

If [R,Q] = 0, single layer is covariant:

$$\mathbf{y} \to RQ\mathbf{W}Q^{-1}Q\mathbf{x} = Q\mathbf{y}$$

However…

what the hell is this thing?

$$\mathbf{y}_t(Q_t) = \left(\prod_{n=0}^{t-1} R_{t-n} Q_{t-n} \mathbf{W}_{t-n} Q_{t-n}^{-1} \right) Q_1 \mathbf{x}_1$$

Lagrangian description

Given N training input/output pairs $\mathbf{z}_i = (\mathbf{X}_i, \mathbf{Y}_i)$

Average loss: $L = \frac{1}{N} \sum_{i=1}^{N} L_i(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{W}, \mathbf{Q})$

why are symmetry transformations explicitly included?

Continuum limit: $L = \int p(\mathbf{X}, \mathbf{Y}) L_{\mathbf{X}}(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{Q}) d\mathbf{X}\, d\mathbf{Y}$

write as loss per layer as number of layers $t \to \infty$

$$L_{\mathbf{X}} = L_{\mathbf{X}}(t=0) + \int_0^T \frac{dL_{\mathbf{X}}(\mathbf{X}, \mathbf{Y}, \mathbf{W}(t), Q(t))}{dt}\, dt$$

value of loss before training (??)

can a general loss function be split up layer-by-layer like this?

Assumption: $L_{\mathbf{X},t}$ invariant under Q

Concrete example?
None given in this paper…

Lagrangian description

$$S[\mathbf{W}, Q] = \int p(\mathbf{X}, \mathbf{Y}) L_{\mathbf{X},t}(\mathbf{X}, \mathbf{Y}, \mathbf{W}(t), Q(t)) d\mathbf{X}\, d\mathbf{Y}\, dt$$

Claim: minimizing $L_{\mathbf{X},t}$ minimizes $L_{\mathbf{X}}$.

Let W* be minimizer, shift weights to minimum:

$$w^i(\mathbf{z}, Q(t), t) = R(t) W^i(\mathbf{z}, Q(t), t) - R(t) W^{*i}(\mathbf{z}, Q(t), t)$$

why are weights a function of input/output?

Define Lagrangian in terms of these shifted weights:

$$\mathcal{L} = \mathcal{T}[\partial_t \mathbf{w}, \partial_{\mathbf{z}} \mathbf{w}, Q(t)] - p(\mathbf{z}) L_{\mathbf{x},t}(\mathbf{z}, \mathbf{W}(t), Q(t))$$

This seemed like a ruse to end up with a Lagrangian with loss as potential. Is there a better-motivated way to add kinetic terms?

If symmetry group is O(N):

what loss function/network architecture has O(N) invariance?

EFT is phi4 model:

$$\mathcal{L} = \frac{1}{2}(\partial_t w)^2 - \frac{1}{2}(\partial_{\mathbf{z}} w)^2 - \frac{m^2}{2}w^2 - \frac{\lambda}{4}(w^2)^2$$

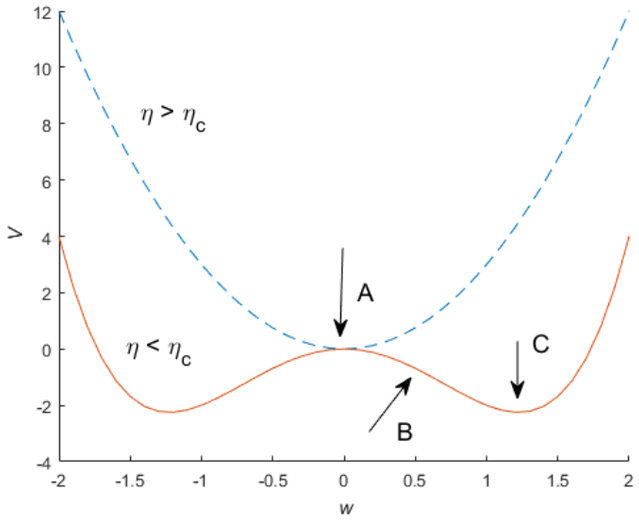
Claim: “to account for the effect of the learning rate, we employ results from thermal field theory and identify the temperature with the learning rate”

$$m^2(\eta) = -\mu^2 + \frac{1}{4}\lambda\eta^2$$

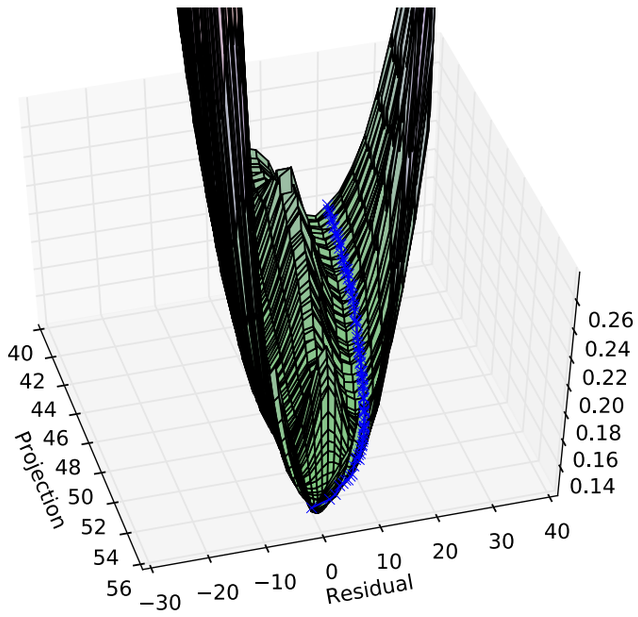
This is probably BS, but something like this seems true. (why does learning rate have dimension 1? would be good to develop a power-counting scheme)

Claim: SSB occurs at end of training

"Proof":



looks like



[Goodfellow, 1412.6544]

Not especially convincing

Can we calculate η_c from first principles given the loss?

Does this phenomenon really require a schedule for the learning rate? DNN's seem to work even with vanilla gradient descent…

From here, a lot of pretentious garbage

Theorem 1: Deep feedforward networks learn by breaking symmetries *Proof*: Let A_i be an operator representing any sequence of layers, and let a network formed by applying A_i repeatedly such that $x_{out} = (\prod_{i=1}^{M} A_i)x_{in}$. Suppose that $A_i \in \text{Aff}(D)$, the symmetry group of all affine transformations. We have $L = \prod_{i=1}^{D} A_i$, where $L \in \text{Aff}(D)$. Then $x_{out} = Lx_{in}$ for some $L \in \text{Aff}(D)$ and x_{out} can be computed by a single affine transformation L. When A_i contains a non-linearity for some i, this symmetry is explicitly broken by the nonlinearity and the layers learn a more generalized representation of the input. □

not SSB…

Theorem (Goldstone) For every spontaneously broken continuous symmetry, there exist a weight π with zero eigenvalue in the Hessian $m_\pi^2 = 0$. □

etc. etc.

Provocative statement: information bottleneck is due to SSB

- Observation: variance of weight gradients grows at end of training

- Claim: this is due to considering two populations of weights, Higgs modes and Goldstone modes, as the same distribution

- Obvious thing to check: correlation functions of weights

Goldstone weights and overfitting

- Claim: zero-eigenvalue weights are robust to overfitting, related to “implicit regularization” [Zhang 1611.03530]
- Seems weird, would have thought you want to gap out the Goldstone modes otherwise they can just slosh around in the potential without changing loss value

Is there any evidence that real DNN’s have degenerate loss minima? What is the role of stochasticity?

Is this a quantum field theory?

No. But trying to model a network during training as a statistical/thermal field theory (e.g. Landau-Ginzburg) seems like an idea worth pursuing.

Good ideas that need fleshing out

- Adding a kinetic term to the loss function to get a Lagrangian (penalty for weights changing too fast between layers?)
- EFT of weights near loss minimum
- Identify learning rate with thermal potential
- If network has O(N) symmetry, EFT is a phi4 model with SSB
- If SSB, some weights are goldstone modes

Questions for us

- What aspects of this seem most promising?
- Who wants to do some numerical experiments?