Criticality and Deep Learning: Generally Weighted Nets arXiv:1702.08039

Przemek Witaszczyk

Jagiellonian University

2 czerwca 2017

INTRODUCTION

Physical motivation

- The so-called criticality is a phenomenon widespread in Nature
- It arises in systems large or *complex* enough to support emergent phenomena
- It is related to phase transitions and critical points in phase diagrams
- Intuitively it is bound to the disappearance of scales in the system
- Critically is very often discovered in network systems of natural and man-made structures, like the Internet, citations, opinions and many more.

Physical motivation

It seems criticality is also present in brain structures.
(cf. works of D. Chialvo)

Can we trace critical behaviour in artificial (deep) neural networks?

Physical motivation

Mathematically criticality implies power laws of various system characteristics, like correlation lengths.

This implies strong group behaviour at large distances, much larger than microscopic interactions.

We shall seek these in neural networks properties.

SIMPLE (RELEVANT) MODEL OF A CRITICAL BEHAVIOUR

Statistical mechanics

- Since the seminal work of Hopfield it is known, that (some) neural networks can be mapped as spin-like physical systems.
- In our case we shall study N units on a lattice with various interactions.
- A particularly simple yet robust model exhibiting critical behaviour is provided by the Curie-Weiss model of magnetisation.
- Let us consider it and study how critical behaviour is described in statistical physics.
- Consider a Hamiltonian based on binary units $s_i = \{\pm 1\}$

$$H_1 = -\frac{J}{2N}\sum_{ij}^{N} s_i s_j - b\sum_{i}^{N} s_i$$

where J is called the coupling and b is the external magnetic field

Statistical mechanics

- It is possible to rewrite H_1 in the following form differing by a constant

$$H = -\frac{J}{2N}\left(\sum_i^N s_i\right)^2 - b\sum_i^N s_i$$

- The critical behaviour we are after is encoded in the partition function

$$\begin{aligned} Z &= \sum_{s_i \in \{\pm 1\}} e^{-\beta H(s)} \\ &= \sum_{s_i \in \{\pm 1\}} \exp \beta \left[\frac{J}{2N} \left(\sum_i^N s_i \right)^2 + b \sum_i^N s_i \right] \end{aligned}$$

- $s_i \in \{\pm 1\}$ represents all the 2^N possible states of the lattice.

Statistical mechanics

- Physical observables are derived from the free energy

$$F[b] = -kT \ln Z[b]$$

  by differentiating over the parameter b.

- The magnetisation $m = \mathrm{d}_b F[b]$ is governed by the average spin and leads to the mean field equation

$$m = b \tanh\big(\frac{K}{b} m + \frac{b}{T}\big), \quad K = (J/T)^{1/2}, \quad T_c = J$$

- It leads to the desired power laws for averages near T_c

$$m \equiv < s_i > \simeq \sqrt{3}\frac{(K-1)^{1/2}}{K^{3/2}} \sim |t|^{1/2}, \ \langle s_i, s_j \rangle \sim \frac{b^2}{T_c}|t|^{-1}$$

Statistical mechanics

Basically all of this can and will be repeated for a system representing feed forward neural network remapped to a spin system.

Let's do it.

STATISTICAL MODEL OF A NEURAL NETWORK

Criticality in deep learning nets

- We start with a two-layers $a_i,\ b_j \in \{0,1\}$ FF network connected with weights matrix w_{ij}, ReLU in b_j and bias h

$$H = -\frac{1}{2N}\sum_{ij}^{N} w_{ij} a_i b_j - h\sum_{i}^{N} b_i$$

- The weight matrix with biases can be represented as

$$W = \begin{pmatrix} 2Nh & 0 & \cdots & 0 \\ 2Nh & w_{11} & \cdots & w_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ -2Nh & w_{n1} & \cdots & w_{nn} \end{pmatrix}$$

acting now on any vector of the form $V_i = (1, V_1, ..., V_N)$

Criticality in deep learning nets

- This Hamiltonian can be recast in a form based on $a \in \{\pm 1\}$
- The partition function up to a constant will then be

$$Z = \sum_{a,b \in \{\pm 1\}} e^{-\frac{\beta}{2N} \sum_{ij} W_{ij} a_i b_j}$$

- The connections here are still of the bipartie graph

Criticality in deep learning nets

- However it can be shown, that the leading contribution (1-loop Det) to Z comes from our familiar fully connected Hamiltonian

MAGIC (in Physics: Beauty):

$$Z \to Z = \sum_{s_i \in \{\pm 1\}} e^{-\frac{\beta}{2N}\sum_{ij} W_{ij} s_i s_j}$$

- This defines free energy capturing the potential critical behaviour of the *neural network*

Criticality in deep learning nets

We need to perform the statistical sum of the system and use large N limit to derive correlation functions.

We shall use a well established Russian trick: Hubbard-Stratonovich transform in vector model (non-constant couplings W).
The non-linear terms can be integrated by introducing auxiliary parameter:

$$e^{a^2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx e^{-x^2/2+\sqrt{2}ax}$$

MORE MAGIC (IN PHYSICS: MORE BEAUTY)

Criticality in deep learning nets

- Even further we recast the problem in continous variables by introducing Dirac deltas

$$\begin{aligned}
H(s) &= -\frac{1}{2N}\sum_{ij}^{N} s_i W_{ij} s_j \\
&= -\frac{1}{2N}\prod_k \int_{-\infty}^{\infty} dV_k \delta(s_k - V_k) \sum_{ij}^{N} V_i W_{ij} V_j \\
&= \prod_k \int_{-\infty}^{\infty} dV_k \delta(s_k - V_k) H(V)
\end{aligned}$$

Criticality in deep learning nets

- Using $\delta(s) = 1/(2\pi i) \int_{\pm i\infty} \mathrm{d}y \, \exp(xy)$ we arrive at

$$\begin{aligned} Z(s) &= \prod_k \int_{-\infty}^{\infty} dV_k \delta(s_k - V_k) \sum_{s_i \in \{\pm 1\}} e^{-\beta H(V)} \\ &\sim \prod_k \int_{-\infty}^{\infty} dV_k \int_{-i\infty}^{i\infty} dU_k \sum_{s_i \in \{\pm 1\}} e^{U_k(s_k - V_k)} e^{-\beta H(V)} \\ &= \prod_k \int_{-\infty}^{\infty} dV_k \int_{-i\infty}^{i\infty} dU_k e^{-U_k V_k + \ln(\cosh U_k)} e^{-\beta H(V)} \end{aligned}$$

- We have exploited the engineered linear dependence on s_i !
- We ultimately seek $F = \lim_{N \to \infty} (-T \log Z)/N$.

Criticality in deep learning nets

- We have obtained an effective Hamiltonian H^g

$$\begin{aligned} H^g &= -\frac{\beta}{2N}\sum_{ij} W_{ij}V_iV_j + \sum_i \left[U_iV_i - \ln\left(\cosh U_i\right)\right] \\ &= -\frac{\beta}{2N}\textstyle\sum_{ij} w_{ij}V_iV_j - \beta h \sum_i V_i + \sum_i \left[U_iV_i - \ln\left(\cosh U_i\right)\right] \end{aligned}$$

  and now we need to compute

$$Z = c\prod_i \int_{-\infty}^{\infty} dV_i \int_{-i\infty}^{i\infty} dU_i \, e^{-H^g(V,U,T)}$$

- The way to proceed is to employ saddle point approximation

$$\int_{-\infty}^{\infty} dx e^{-f(x)} \approx \left(\frac{2\pi}{f''(x_0)}\right)^{1/2} e^{-f(x_0)}$$

Criticality in deep learning nets

- The second derivative matrices (Hessians $f''(x_0)$) at the stationary point read

$$
\begin{aligned}
H^g_{V_iV_j} &= -\frac{\beta}{N} w_{ij} \\
H^g_{U_iU_j} &= -\delta_{ij}(1 - \tanh^2 U_i) \\
H^g_{V_iU_j} &= \delta_{ij}
\end{aligned}
$$

while the point (x_0) itself is given by

$$
\begin{aligned}
\frac{\partial H^g}{\partial V_i} &= -\frac{\beta}{N} \sum_j W_{ij} V_j + U_i \\
&= -\beta(\textstyle\sum_j w_{ij} V_j / N + h) + U_i = 0 \\
\frac{\partial H^g}{\partial U_i} &= V_i - \tanh U_i = 0
\end{aligned}
$$

Criticality in deep learning nets

- To read-off observables we need on-shell Hamiltonian (the $f(x_0)$ piece)

$$H_0^g = \quad \tfrac{\beta}{2N} \textstyle\sum_{ij} w_{ij} V_i V_j - \sum_i \ln\cosh \beta(\sum_j w_{ij} V_j / N + h)$$

- The large N limit leads to the Helmholtz free energy (U_j eliminated)

$$F[h] = \frac{T}{N} H_0^g = \frac{1}{2N^2} \sum_{ij} w_{ij} V_i V_j - \frac{T}{N} \sum_i \ln\cosh \beta(\textstyle\sum_j w_{ij} V_j / N + h)]$$

- We can finally read-off the magnetisation

$$m \equiv \frac{dF}{dh} = \left.\frac{\partial F}{\partial h}\right|_{V^{st}} + \frac{\partial F}{\partial V_i} (\equiv 0) \left.\frac{\partial V_i}{\partial h}\right|_{V^{st}}$$

Criticality in deep learning nets

- This leads to the familiar mean field conditions of the CW model

$$\begin{aligned} \sum_i w_{ik} V_k / N &= \frac{1}{N} \sum_i \tanh \beta(\textstyle\sum_k w_{ik} V_k / N + h) w_{ik} \\ &\Updownarrow \\ V_i &= \tanh \beta(\textstyle\sum_k w_{ik} V_k / N + h) \end{aligned}$$

- The critical point P_c is defined as the singularity of the correlation function when $h \to 0$

$$\begin{aligned} P_c &\equiv \frac{d^2 F}{dh^2} = \frac{dm}{dh} \\ &\Updownarrow \\ \frac{\partial V_i}{\partial h} &= \beta(1 - V_i^2)(1 + \sum_k w_{ik} \frac{\partial V_k}{\partial h} / N) \end{aligned}$$

FINALLY, PICTURES

Searching for criticality in Neural Networks

- Apart from the analytical results several modern architectures were tested for the presence of power laws
- A priori we would expect some features to develop behaviours of a sort

  $$P(k) \sim k^{-\gamma}$$

  The architectures included large networks of the following type:

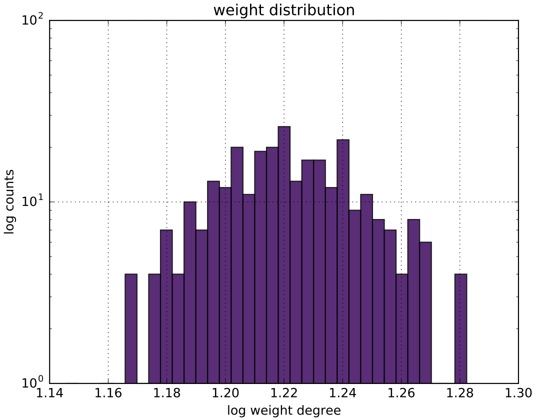
  - Deep feed-forward: 3@500-400-200-200 nodes
  - Convolutional CNN: 3xCONV-3xFULL-CONN layers
  - Autoencoder: 1x500
- Tests were run on CIFAR-10 with Adam Optimizer and ReLU activations and no gradient clipping
- They were trained for 200 epochs and inference was run for 100 epochs

The observables

The statistics gathered on the 100 inference epochs included:

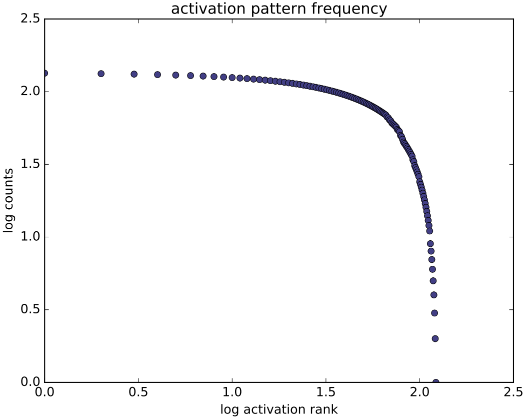
- 1. Average weights distribution across the net (histogram of weighted node order)
- 2. Activation pattern frequency for layers (what nodes patterns fired)
- 3. Node of a given weight activation frequencies
- 4. Average layer activation distribution

Class 1: Layer weight distribution



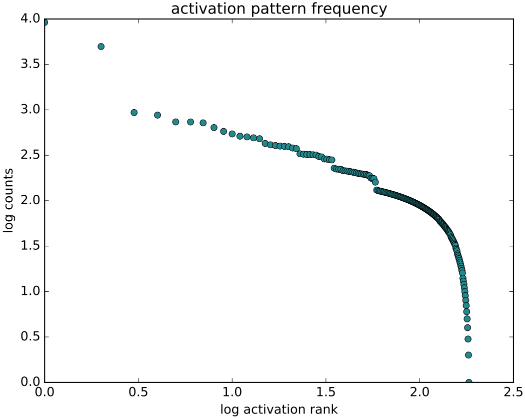
Rysunek: Feed-forward net: Layer 3 weight distribution

Class 2: Layer 2 activation pattern frequencies



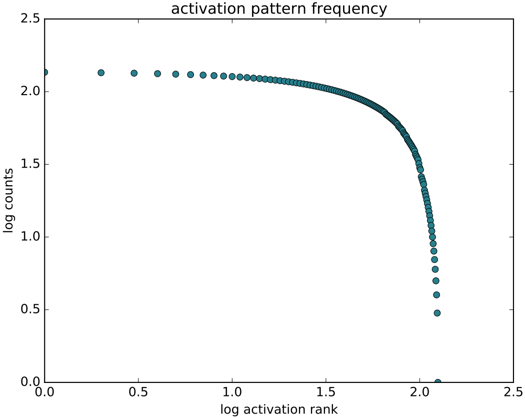
Rysunek: Feed-forward net: Log-log plot of layer activation pattern frequencies by rank

Class 2: Layer 1 activation pattern frequencies



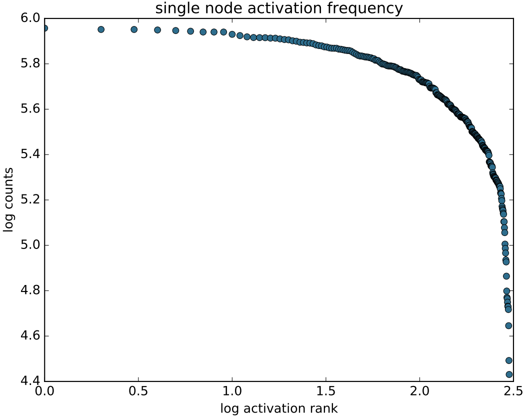
Rysunek: Autoencoder: Log-log plot of layer activation pattern frequencies by rank

Class 2: Layer 3 activation pattern frequencies



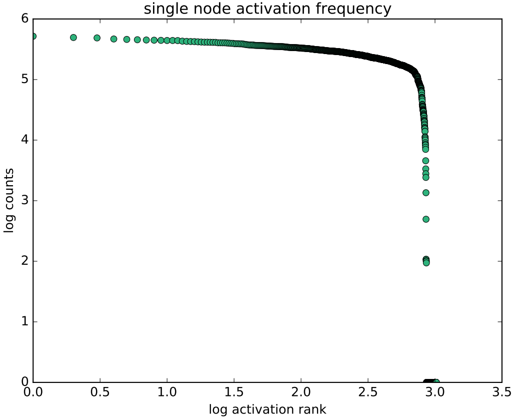
Rysunek: CNN: Log-log plot of layer activation pattern frequencies by rank

Class 3: Single neuron activations per rank, layer 2



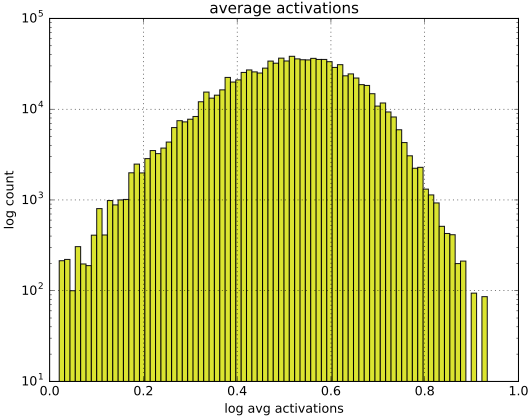
Rysunek: Feed-forward net: Log-log plot of single node activation frequencies by rank

Class 3: Single neuron activations per rank, layer 2



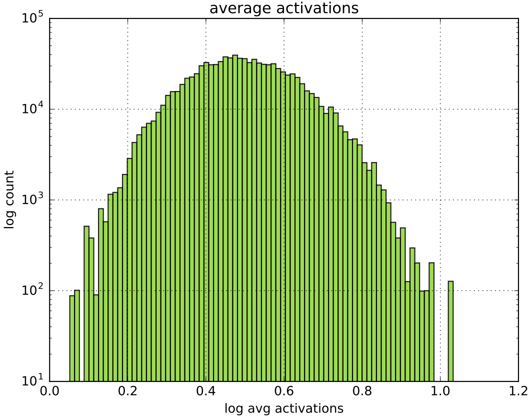
Rysunek: CNN: Log-log plot of single node activation frequencies by rank

Class 4: Average layer activation distribution, layer 4



Rysunek: Feed-forward net: Average layer activation distribution

Class 4: Average layer activation distribution, layer 4



Rysunek: CNN: Average layer activation distribution

Conclusions

- Analytical tools were developed to study statistical mechanics of systems dual to neural network layer
- Almost no trace of criticality was detected during the numerical search
- A set of new training criteria based on the developed formalism will be proposed by authors
- A follow-up paper just appeared on ArXiv..
- Intuitively, criticality should be present for the efficiency of information flow in the neural network