Using statistical mechanics to approach the optimal size of a network in image recognition

Markus Abel

Thomas Seidler

September 27 2023

Talk at the 3rd Symposium for Machine Learning and Dynamical Systems

Motivation

Image recognition and AI

Motivation

Drawing by Michel Foucault

- What do we sense?
- What do we interprete?
- What is reality?

Motivation

Drawing by Michel Foucault

- $O(10^9)$ neurons in the visual cortex
- Information stored in the brain
- There is a joint concept on "things"

Motivation

- Neural networks as a statistical system:
- many approaches analogous to statmech of the brain
- main idea: consider nodes as "neurons" and edges as "axons"
- Formulate a "Hamiltonian" as coupled units
- apply statistical mechanics to find transitions, states, phases

$$H(t) = -\sum_{i,j} J_{ij} S_i S_j - h(t) \sum_I S_I$$

with J the coupling S_i the state of a unit, and h a time-dependent forcing. Measured quantities in terms of statistics, e.g. magnetization

$$M = \sum_i s_i = N < s >$$

Deep Neural Networks

Given: input $x \in R^d$, weights $w^1, ..., w^{p-1} \in R^{dxd}$, $w^p \in R^{dxK}$, nonlinear functions σ
Output $y \in R^K$:

$$y(w,x) = \sigma(w^p \sigma(w^{p-1} \sigma(...\sigma(w^1 x))))$$

- supervised learning: given is $x^i, y^i{}_{i=1}^N$
- minimize the empirical loss

$$f(w) = 1/N \sum_{i=1}^{N} f_i(w),$$

e.g.

$$f_i(w) = 1 \; if \;\; y(w,x) \neq y^i, else \; 0$$

Deep Neural Networks

- the objective $f(w)$ is a non-convex funciton of w
- optimization problem

$$w^* = argmin_w(f(w))$$

Measured quantity, e.g., the mean classification quality

$$Q =< |f| >$$

Q is nonextensive in contrast to M

DNNs vs. Stat Mech

- weights and coupling $w \leftrightarrow J$
- classification and magnetization $y^i \leftrightarrow M(t)$
- Objective and Hamiltonian $f(w) \leftrightarrow H(J)$

Solution of finding the ground state of a system, or the optimal solution, resp. is very expensive, since the number of macrostates is huge. Approximate solutions are accepted.

Analogy

Stat Mech

- describe quantities by mean values
- mean values are sharp due to large number of variables
- parameters are equivalent to constraints
- solution (classic) by Lagrange formalism

Entropy or information

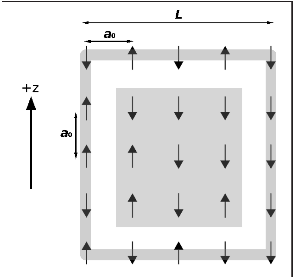
$$S = - < ln\, \omega_i > ,$$

with $\omega_i = \omega(f_i) = \frac{1}{Z} e^{-\beta f_i}$ (or E_i). f is a parameter, e.g. mean energy, or a constraint in optimization (like $f_i = f_i(w)$).

The role of noise

Side step: The Ising model once more

$$H = -J \sum_{ij} s_i s_j - h \sum_i s_i$$



mean energy or temperature determines the transition from magnetized to unmagnetic

Phase transitions in the Ising model

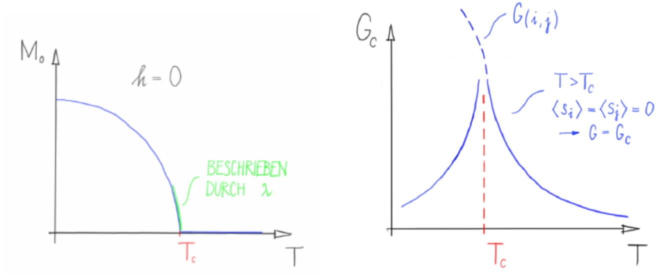


Figure: Left: Magnetization, right: Correlation function.

$$M_0 = < \Sigma_i s_i > \quad (1)$$
$$G_c(i,j) = < s_i s_j > - < s_i > < s_j > \quad (2)$$

Noise generalized

What is the temperature or the noise in DNNs?
Consider image recognition, e.g. numbers or faces.
Working definition: noise is anything that does not beolong to the object to be classified, e.g. wrong pixels, objects covering a face, insufficient resolution

with increasing noise: transition from successful classification to impossible classification

- Depends on the classification complexity
- the size of the network matters
- the evolution time, i.e. the number of epochs matters
- Asymptotics: infinite number of epochs, infinitie DNN, infinite number of images

A simple setup

classify black or white squares as 0 or 1.
Without noise: a two-neuron network is needed after the first layer.
With noise: more neurons are needed, e.g. to compute a convolution or pooling.
With noise very large: no classification possible without ensemble averaging.

A realistic setup

The problem is not too complex, e.g. image, or number recognition The number of weights is not too high $()O(10^12))$

Investigate noise dependence

- Expected: with diminished noise, classification is possible
- Setup: many images, high resolution, large network

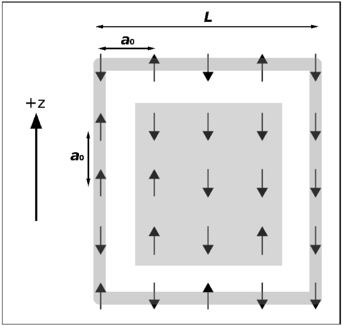
A realistic setup

The problem is not too complex, e.g. image, or number recognition The number of weights is not too high, say $O(10^12)$

Investigate noise dependence

- Expected: with diminished noise classification is possible
- Setup: many images, high resolution, large network

Finite Size Scaling - Ising again



For systems of finite size L^d and observables $Q(t) \propto (t)^y$:

$$L^{y/\nu} Q(L,t) = f(L^{1/\nu} t) \quad (3)$$

with $t = \frac{T - T_c}{T_c}$.

Application - Size optimization

What is the optimal size of a network to classify correctly in an acceptable time

- Important for saving resources
- allows estimation of needed networks and number of epochs
- answers specialized chipdesign (e.g. with autonomous drive)

statmech yields (cf. Biehl, 2000)

ambrosys

with soft-committee machines.

Signal: x

Rule (correct classification): $\tau(\xi)$

Classification result: $\sigma(\xi) = \sum \xi_{weighted}$

statmech computations

Training error: $\epsilon = \frac{1}{2n_{obs}} \sum_1^{n_{obs}} (\sigma(\xi) - \tau(\xi))^2$

Generalization error: $\epsilon_g = \frac{1}{2} < (\sigma(\xi) - \tau(\xi))^2 >$

partition function: $Z = \int d\mu exp[-\beta n_{obs} \epsilon]$

What do we want to find out?

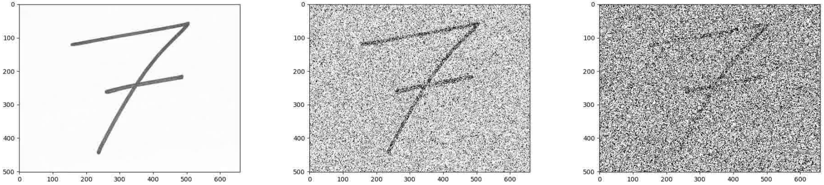
- Close to the phase transition: Optimal balance between training efficiency and model quality.

What do we want to find out?

- Close to the phase transition: Optimal balance between training efficiency and model quality.
- Finite size scaling: How many data are needed? How much can an increase in quality and amount of data improve the training?

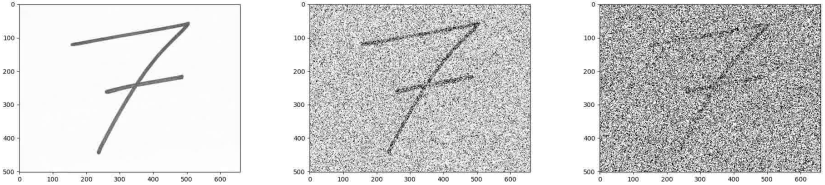
What do we want to find out?

- Close to the phase transition: Optimal balance between training efficiency and model quality.
- Finite size scaling: How many data are needed? How much can an increase in quality and amount of data improve the training?
- Model classification: Can we recommend the optimal model for a given problem ?



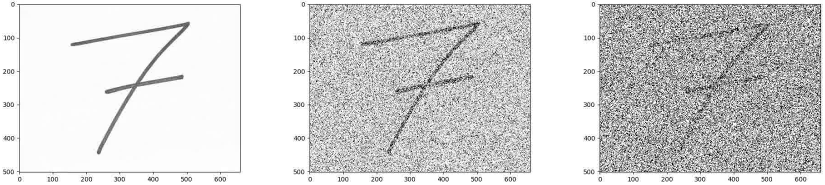
- Add Gaussian white noise to some classification data.

phase-ml.ambrosys.de

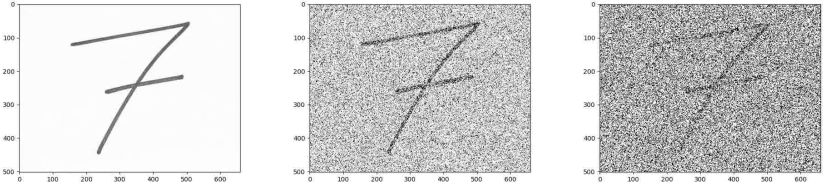


- Add Gaussian white noise to some classification data.
- Train a model and compute metrics.

phase-ml.ambrosys.de



- Add Gaussian white noise to some classification data.
- Train a model and compute metrics.
- Average over multiple noise realizations (create an ensemble of systems).



- Add Gaussian white noise to some classification data.
- Train a model and compute metrics.
- Average over multiple noise realizations (create an ensemble of systems).
- Repeat for multiple noise intensities.

A gentle reminder on interpretation

Drawing by Michel Foucault

- What is essential and what is not?
- What is noise and what is not?

Experiment description

Data

MNIST dataset from sklearn
Resolution: 8x8
Number of Instances: 1797
Missing Attribute Values: None
Copy of hand-written digits datasets
10 classes - 1 per digit

cf. Garris et all, NISTIR 5469, 1994, Alpaydin and Kaynak (1998) Cascading Classifiers, Kybernetika, Gentile, NIPS 2000

Experiment

Metrics : sklearn.metrics.accuracy_score

$$accuracy(y, \hat{y}) := \frac{1}{N} \sum_{i=0}^{N-1} 1(\hat{y}_i = y_i)$$

Model: sklearn.linear_model.Perceptron

tolerance $tol = 1e - 3$, random_state=0

Noise

Gaussian with mean 0 and standard deviation σ

Added pixelwise

Results

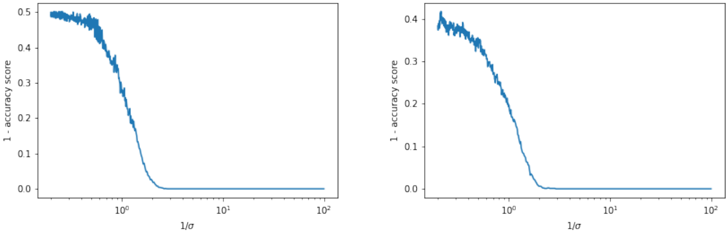
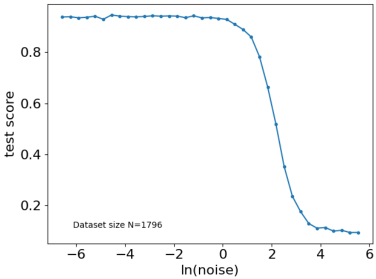
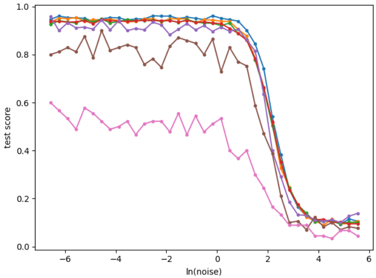


Figure: Left: Perceptron, right: RidgeClassifier.

- Classification problem: Detect black or white image
- Entropy = 2 Bit, theoretically needed: 2 neurons
- Noise: larger system needed
- Noise intensities: 1000 between 0 and 5 (with a signal intensity of 1)

Results for MNIST

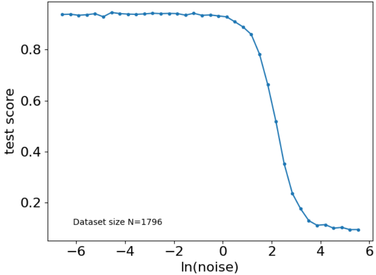


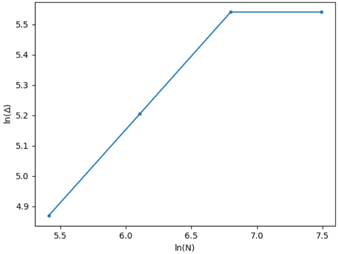


left: maximum number of data, right: different data sizes.

- Nice transition, already for $N = O(10^3)$
- Is there finite-size scaling with N?

Results for MNIST





left: maximum number of data, right: finite-size behaviour

- if the (daring) scaling is determined: exponent is $1/2$
- if true - is this a universal behaviour?

Results for MNIST - so what?

What is the practical implication? - Architectures

Knowing scaling for an architecture, we can now determine the amount of data needed to reach a certain quality of the classification

What is the practical implication? - Data sizes

Knowing the amount of data, we can determine the minimum size of a network to reach a certain quality of the classification

Universality and critical exponents

Close to critical temperature, example magnetization or correlation:

$$M_0 \propto \left(\frac{T_c - T}{T_c}\right)^{\beta} \tag{4}$$

$$G_c \propto \frac{1}{r^{d-2+\eta}} \tag{5}$$

β, η is "universal" for a whole class of systems.

Does this hold for classes of machine learning problems?

Discussion

Optimization and statistical mechanics

- Analogy can be established
- For small networks corrections are needed
- For large networks formalism may apply
- Phase transitions are observed

Application

- Classification shows a well defined transition for a model class
- Dependence on "driving", i.e. statistical properties of the data
- Dependence on size of network
- Dependence on the complexity of the optimization task

Outlook

Model and problem universality

- Problems may be classified (continuous, discrete, NP hard, nonlinear, ...)
- Models may be classified (NNs, random forests, dynamic programming,...)
- universality exponents may help to determine a good choice of method for a certain problem
- ... and the typical size of a NN needed to reach a certain quality
  given an amount of data, including if a DNN approach is **useful at all**