Opening the black box of Deep Neural Networks via Information

Ravid Schwarz-Ziv Naftali Tishby

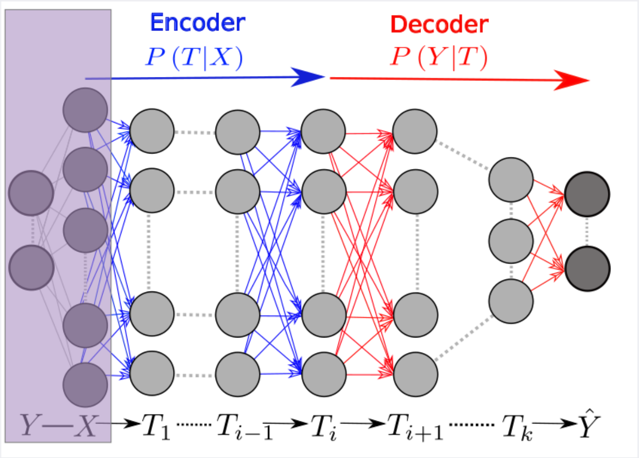
Presentation by Bogdan Kirillov for Graphical Models and Statistical Inference mini-course at Skoltech

April 29, 2017

Key points of the paper

- Most of the training time is spent on compression of the input to efficient representation rather than on fitting to labels;
- The compression begins when the training errors shrink and the SGD epochs go into random diffusion, constrained by the training error value;
- The converged layers lie very close to the theoretical bound, and the maps from the input to any hidden layer and from this hidden layer to the output satisfy the IB self-consistent equations. This mechanism is absent in one layer networks;
- Adding more hidden layers reduces training time;
- Hidden layers are expected to converge to critical points for phase transition of Informational Bottleneck curve.

Markov Chain of successive representations


Encoder
$P(T|X)$
Decoder
$P(Y|T)$
Y—$X \longrightarrow T_1 \cdots\cdots T_{i-1} \longrightarrow T_i \longrightarrow T_{i+1} \cdots\cdots T_k \longrightarrow \hat{Y}$

Information Theory of Deep Learning 1

Mutual Information:

$$\mathrm{I(Input;Label)} = \mathrm{D_{KL}[p(i,l)||p(i)p(l)]} = \sum_{\mathrm{l \in Label; i \in Input}} \mathrm{p(i,l) log} \frac{\mathrm{p(i,l)}}{\mathrm{p(i)p(l)}} \quad (1)$$

Invariance to invertible transformations:

$$\forall \phi, \gamma : \exists \phi^{-1}, \gamma^{-1} \Rightarrow \mathrm{I(A;B)} = \mathrm{I}(\phi(\mathrm{A}); \gamma(\mathrm{B})) \quad (2)$$

Data-Processing Inequality:

$$\forall \mathrm{X, Y, Z} : (\mathrm{X} \rightarrow \mathrm{Y} \rightarrow \mathrm{Z}) \Rightarrow \mathrm{I(X;Y)} \geq \mathrm{I(X;Z)} \quad (3)$$

Information Theory of Deep Learning 2

Minimal Sufficient Statistic:

$$\mathrm{T(Input)} = \underset{\mathrm{S(Input):I(S(Input),Label)=I(Input,Label)}}{\arg\min} \mathrm{I(S(Input);Input)} \tag{4}$$

Information Bottleneck tradeoff:

$$\min_{\mathrm{p(r|i),p(l|r),p(r)}} \mathrm{I(Input;Representation)} - \beta \mathrm{I(Representation;Label)} \tag{5}$$

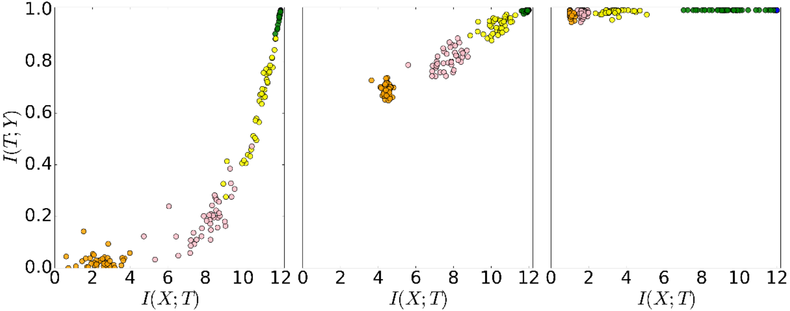
IB self-consistent equations:

$$\begin{cases} \mathrm{p(r|i)} = \frac{\mathrm{p(r)}}{\mathrm{Z(i;}\beta)} \exp(-\beta \mathrm{D_{KL}[p(l|i)||p(l|r)]}) \\ \mathrm{p(r)} = \sum_{\mathrm{i}} \mathrm{p(r|i)p(i)} \\ \mathrm{p(l|r)} = \sum_{\mathrm{i}} \mathrm{p(l|i)p(i|r)} \end{cases} \tag{6}$$

Optimal β (relevance wheel):

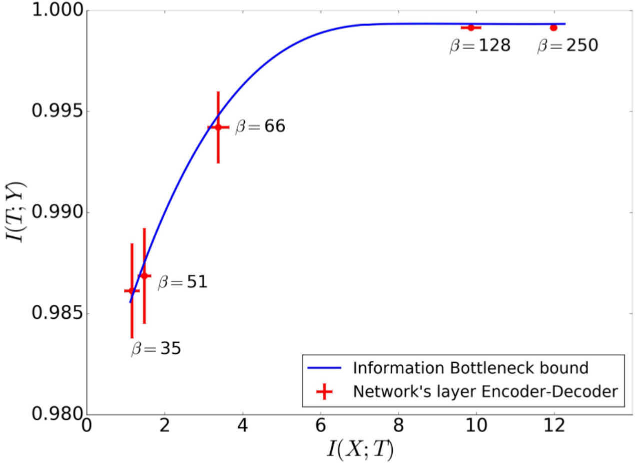
$$\beta^*_{\mathrm{j}} = \arg\min_{\beta} \mathbb{E}\, \mathrm{D_{KL}[p_k(r|i)||p^{IB}_{\beta}(r|i)]} \tag{7}$$

Network Behavior on Information Plane



- First stage - compression of representation (from left to middle);
- Second stage - shuffling compressed representation to fit the labels (right);

Consistency with theoretical boundaries

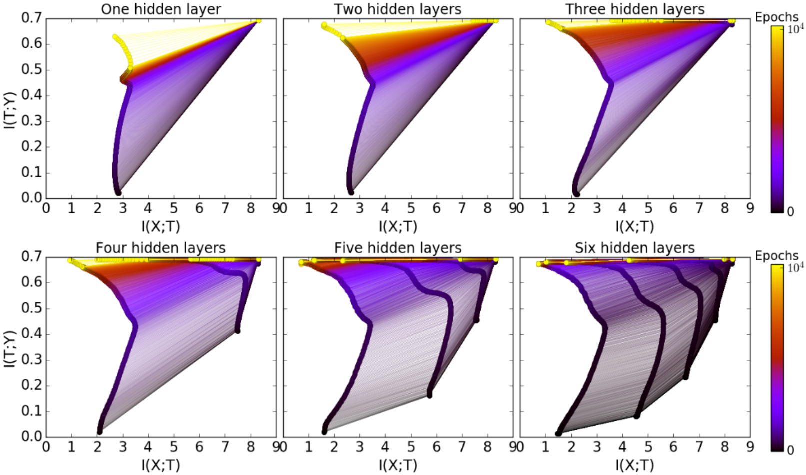


Why do the networks need noise?

- Mutual Information is invariant to any random shuffle of data;
- There is no way to distinguish between complex classes and simple classes (in terms of VC-dimension) by MI alone without tips on structure of those classes;
- The former is true only for deterministic functions;
- If the function is stochastic, the joint distributions can give a clue for distinguishing the classes.

What's the point of hidden layers?

- More hidden layers - good generalization is reached faster;
- Compression phase of a layer is shorter if it is started from previously compressed layer;
- For the deeper layers the compression is faster;



Numerical experiment setup

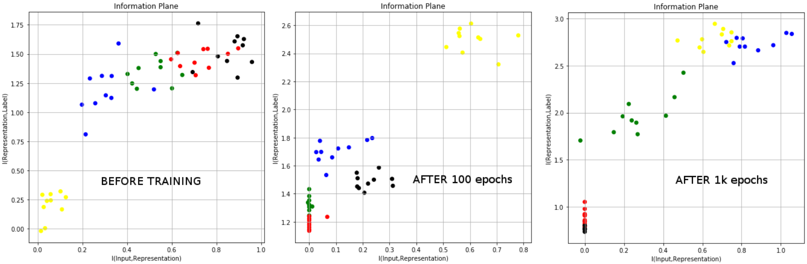
- Different way of MI estimation - I use NPEET library;
- Real data instead of generated;
- Adam instead of plain SGD.

The models are trained on a subset of MNIST:

- 1000 - training sample size;
- 100 - test sample size.

Numerical experiments on MNIST

The network used is fully-connected with 4 [256,128,64,32] hidden layers, trained to perform Regression (due to limitations of NPEET). Activations are ELU everywhere except the output layer which is Linear. The layer is marked by color in the following way: ["black", "red", "green", "blue", "yellow"].



10 randomly initialized identical models are shown. The behavior is consistent with results of Schwarz-Ziv and Tishby.

Interesting links

- The paper of Schwarz-Ziv and Tishby - https://arxiv.org/pdf/1703.00810.pdf;
- Non-Parametric Entropy Estimation Toolbox - https://github.com/gregversteeg/NPEET;
- Deep Variational Information Bottleneck - https://arxiv.org/pdf/physics/0004057.pdf;
- Code for numerical experiments in this presentation - https://github.com/bakirillov/GM_Skoltech.