Лекция 10: Анализ нейронных сетей методами теории информации.

Алексей Фролов

al.frolov@skoltech.ru

Сколковский институт науки и технологий (Сколтех)

Современные методы теории информации, оптимизации и управления

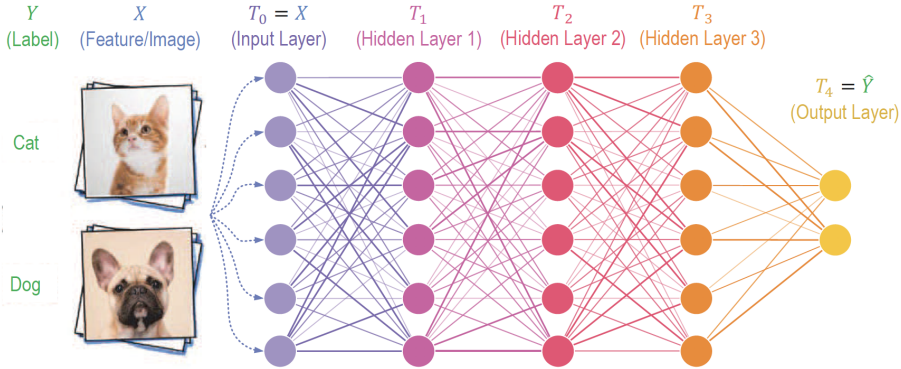
Сочи, Россия

2–23 августа, 2020

Содержание

1 Гипотеза Тишби

(Deterministic) Feedforward DNN: Each layer $T_\ell = f_\ell(T_{\ell-1})$



Гипотеза Тишби

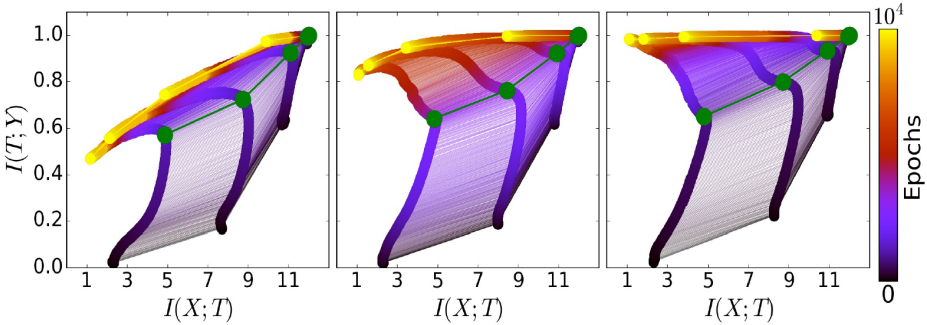


Figure 3: The evolution of the layers with the training epochs in the information plane, for different training samples. On the left - 5% of the data, middle - 45% of the data, and right - 85% of the data. The colors indicate the number of training epochs with Stochastic Gradient Descent from 0 to 10000. The network architecture was fully connected layers, with widths: input=12-10-8-6-4-2-1=output. The examples were generated by the spherical symmetric rule described in the text. The green paths correspond to the SGD drift-diffusion phase transition - grey line on Figure 4

Гипотеза Тишби

Стадии обучения нейронной сети:

1. Fitting: $I(Y; T_i)$ и $I(X; T_i)$ растут
2. Компрессия: $I(X; T_i)$ уменьшается

R. Shwartz-Ziv, N. Tishby.
Opening the Black Box of Deep Neural Networks via Information
arXiv:1703.00810, 2017.

Моменты, вызывающие вопросы

1. Квантование pdf

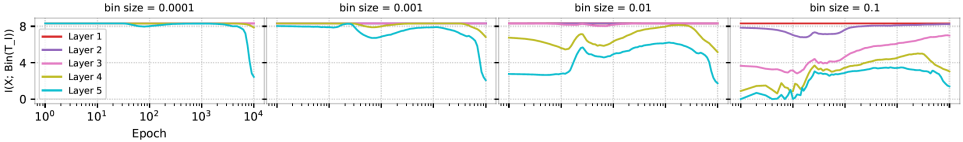
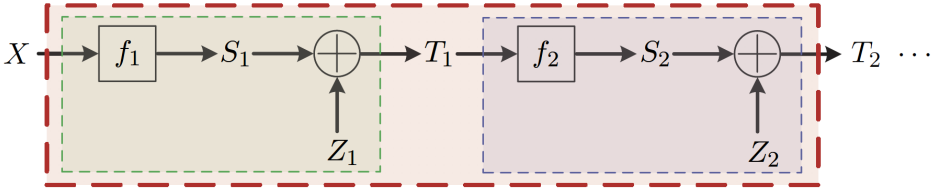


Figure 1. $I\big(X; \mathrm{Bin}(T_\ell)\big)$ vs. epochs for different bin sizes and the model in (Shwartz-Ziv & Tishby, 2017), where X is uniformly distributed over a 2^{12}-sized empirical dataset. The curves converge to $H(X) = \ln(2^{12}) \approx 8.3$ for small bins.

2. $T_i = f(X)$, поэтому $I(X; T_i) = \min\{H(X), H(T_i)\}$.

Нейронная сеть с шумом

$$T_i = f_i(T_{i-1}) + Z_i,\ Z_i \sim N(0, \sigma^2 \mathbb{I}_d).$$



Задача: оценить $h(X + Z)$, используя n сэмплов.

Методы оценки

- Козаченко–Леоненко, 1987
- взвешенный Козаченко–Леоненко, 2017
- KDE

Козаченко–Леоненко, 1987

$$\hat{H}_n = \hat{H}(X_1, \ldots, X_n) = \frac{1}{n} \sum \log \left(\frac{p^d_{(k),i} V_d(n-1)}{e^{\Psi(k)}} \right),$$

где

- V_d – объем d-мерного единичного шара
- Ψ – дигамма-функция
- $p_{(k),i}$ – расстояние до k-го ближайшего сэмпла

взвешенный Козаченко–Леоненко, 2017

$$\hat{H}_n^w = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_j \log \xi_{(j),i}\xi_{(j),i} = \frac{p_{(j),i}^d V_d(n-1)}{e^{\Psi(j)}}$$

Спасибо за внимание!