Обзор статьи
Opening the Black Box of Deep Neural Networks via Information

Глеб Пособин

9 октября 2017 г.

Взаимная информация

Взаимная информация для двух случайных величин X, Y:

$$I(X;Y) = D_{kl}(p(x,y)||p(x)p(y)) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Инвариантность к обратимым преобразованиям:

$$I(X;Y) = I(\phi(X); \psi(Y))$$

Data processing inequality: для любых трёх случайных величин X, Y, Z, таких что $X \to Y \to Z$, верно:

$$I(X;Y) \geqslant I(X;Z)$$

Достаточные статистики

Две случайные величины X, Y.

Достаточная статистика $S(X)$:

$$I(S(X); Y) = I(X; Y)$$

Минимальная достаточная статистика $T(X)$:

$$T(X) = \arg \min_{S(X): I(S(X);Y) = I(X;Y)} I(S(X); X)$$

Точные минимальные достаточные статистики очень редко существуют.

Вероятностные достаточные статистики

T — случайная величина, зависящая от X. Разрешим $I(X;Y) \neq I(T;Y)$.

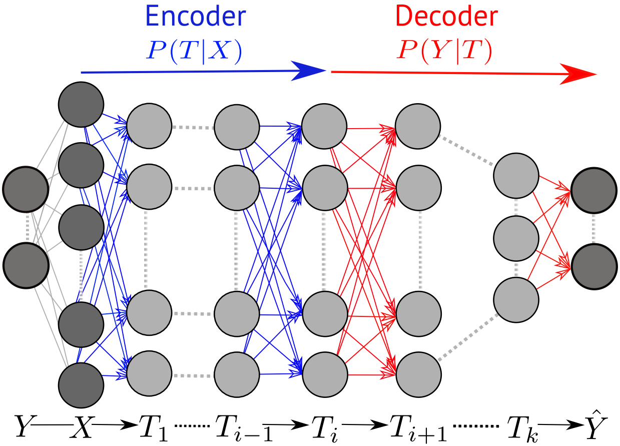
Information Bottleneck trade-off:

$$\min_{p(t|x),p(y|t),p(t)} I(X;T) - \beta I(T;Y)$$

Решения удовлетворяют системе:

$$\begin{cases} p\left(t|x\right) = \dfrac{p\left(t\right)}{Z\left(x;\beta\right)} \exp\left(-\beta D_{KL}\left[p\left(y|x\right) \parallel p\left(y|t\right)\right]\right) \\ p\left(t\right) = \sum\limits_{x} p\left(t|x\right) p\left(x\right) \\ p\left(y|t\right) = \sum\limits_{x} p\left(y|x\right) p\left(x|t\right) \end{cases}$$

Нейронные сети как марковская цепь



$$I(X;Y) \geqslant I(T_1;Y) \geqslant \ldots \geqslant I(T_k;Y) \geqslant I(\hat{Y},Y)$$

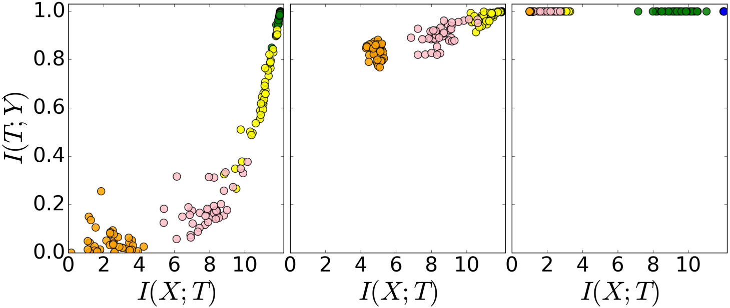
$$H(X) \geqslant I(X;T_1) \geqslant \ldots \geqslant I(X;T_k) \geqslant I(X;\hat{Y})$$

Эксперимент

Сеть: $\leqslant 7$ полносвязных скрытых слоёв с 12-10-7-5-4-3-2 нейронами, активации — $\tanh$, обучаем обычным SGD.

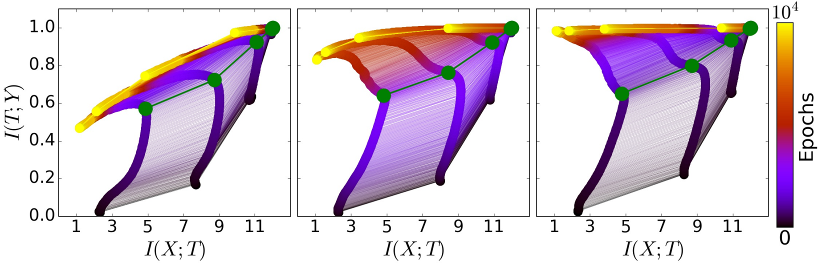
Датасет: искусственно сгенерированный, $X \in \{0,1\}^{12}$, соответствует 12 точкам на двумерной сфере, $Y \in \{0,1\}$.

Графики



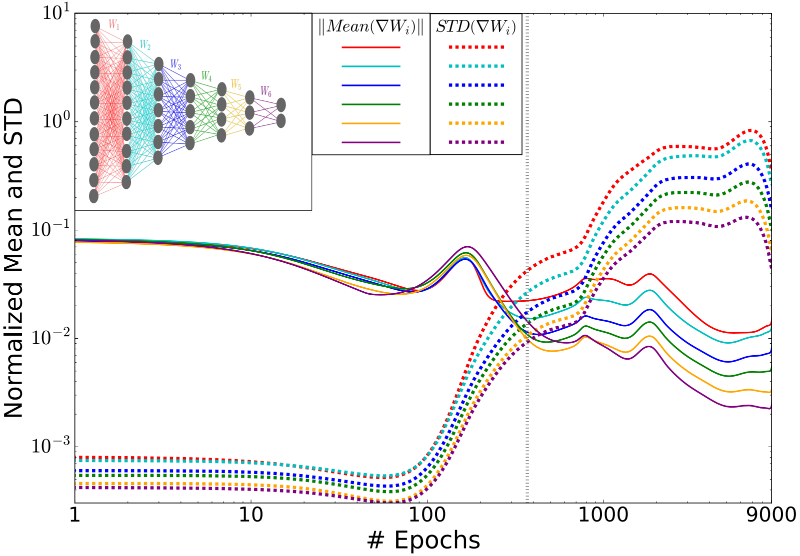
Слева — до обучения, по центру — после 400 эпох, справа — после 9000 эпох. Разные цвета — разные слои.

Графики



Слева — на 5% данных, по центру — на 45%, справа — 85% данных.

Графики



$\|Mean(\nabla W_i)\|$
$STD(\nabla W_i)$
W_1
W_2
W_3
W_4
W_5
W_6
Normalized Mean and STD
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
Epochs
1
10
100
1000
9000

Две фазы SGD

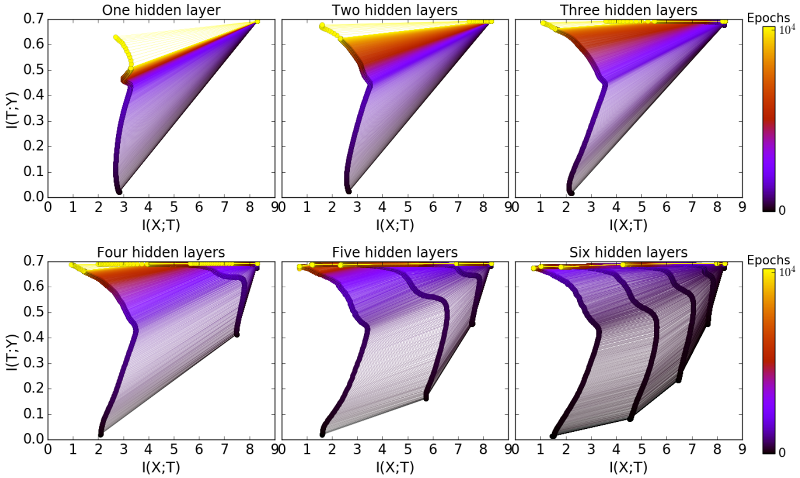
Первая фаза (дрейф): большие средние и маленькие дисперсии градиентов.

Вторая фаза (диффузия): средние значения градиентов гораздо меньше дисперсий, веса фактически совершают случайное блуждание.

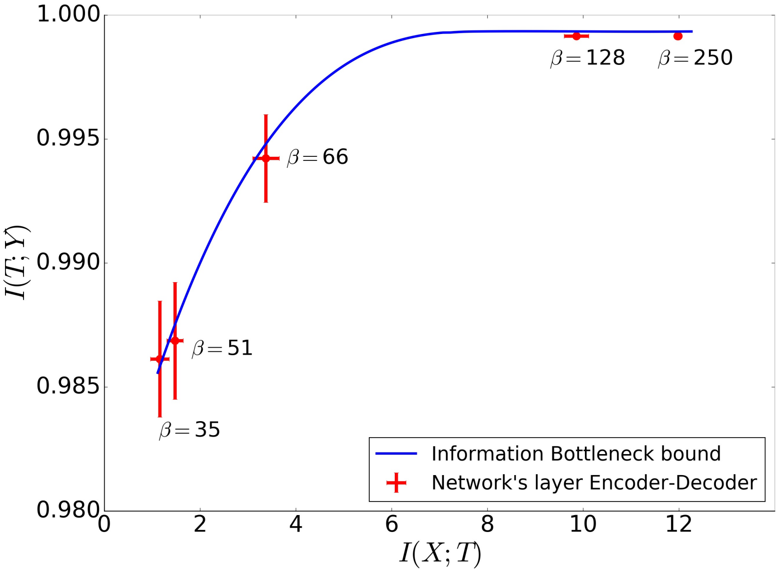
Теоретически подобные случайные блуждания максимизируют $H(X|T_i) = H(X) - I(X; T_i)$, то есть минимизируют $I(X; T_i)$.

Значит, на второй фазе градиентный спуск «сжимает» выученные представления объектов, стараясь не ухудшать $I(T; Y)$.

Влияние числа слоёв на обучение



Близость к Information Bottleneck bound



Результаты

Самое важное: возможно, SGD работает именно из-за диффузии весов сети во второй фазе, когда происходит «сжатие» представлений.

Слои сходятся к Information Bottleneck bound.

Большее число слоёв ускоряет сходимость.

Плюсы/минусы

Плюсы:

- Хорошая идея
- Интересные результаты, хоть и на искусственном примере и для маленькой сети

Минусы:

- Нигде не показаны результаты на отложенной выборке
- Маленькая сеть с только полносвязными слоями
- Странный искусственный датасет
- $\Rightarrow$ непонятно, верны ли выводы на практике, нужно дополнительно проверять

Ссылки

Ravid Shwartz-Ziv, Naftali Tishby:
Opening the Black Box of Deep Neural Networks via Information
http://arxiv.org/abs/1703.00810

Naftali Tishby, Fernando C. Pereira, William Bialek:
The information bottleneck method
http://arxiv.org/abs/physics/0004057