Applications of Information Bottleneck in DL

Reza Pishkoo   Kasra Khoshjoo

Sharif University of Technology

presentation for ITSL course
February 12, 2023

Presentation Overview

Introduction

- Information Theory & Information Transfer
- Machine Learning & Prediction of Parameters
- Information Bottleneck and DNNs

Definitions

- Input Space : $\mathcal{X}$
- Output Space : $\mathcal{Y}$
- random variables $X \in \mathcal{X}$, $Y \in \mathcal{Y}$ with joint distribution $p_{X,Y}$
- x, y represent examples of X, Y respectively

Definitions

Definition

1. Conditional Entropy : $H(X|Y) = \mathbb{E}_{X,Y}[-\log p(X,Y)]$
2. Mutual Information : $I(X;Y) = \mathbb{E}_{X,Y}\left[log \frac{p(X,Y)}{p(X)p(Y)}\right]$

a property of MI:
for every bijective functions f and g: $I(X;Y) = I(f(X), g(Y))$

Relevant Information

a shared concept in Information Theory, Statistics and Machine Learning
a mathmatical formulation of "Relevant Information" is Sufficient Statistics.

Relevant Information

Definition (Sufficient Statistic)

given $S \triangleq f(X)$ for sum function f, S is a sufficient statistic of Y if :

$$\forall x \in \mathcal{X}, y \in \mathcal{Y} : \mathbb{P}\left[X = x \middle| Y = y, S = s\right] = \mathbb{P}\left[X = x \middle| S = s\right]$$

Theorem

S is a sufficient statistic for Y iff:

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

Relevant Information

Definition (Minimal Sufficient Statistic)

S is a MSS if:

$$\forall T;\ T \text{ is sufficient statistic} \Rightarrow \exists g;\ S = g(T)$$

Theorem

given $X \sim P_{Y|X}$, S *is an MSS for* Y *iff:*

$$S \in \underset{T:SS}{argmin}\ I(X;T)$$

subsequently :

Corollary

given $X \sim P_{Y|X}$, S *is an MSS for* Y *iff:*

$$S \in \underset{T:I(Y;T)=I(X;Y)}{argmin}\ I(X;T)$$

Relevant Information

	Markov Chain	Data Processing Inequality
Statistic	Y — X — S	$I(S;Y) \leq I(X;Y)$
Sufficient	Y — SS — X	$I(SS;Y) \geq I(X;Y)$
Minimal	X — SS — MSS	$\forall\, SS : I(MSS;X) \leq I(SS;X)$

Figure: intuition of MSS

The Information Bottleneck framework

let $P_{T|X}$ be a transition kernel from $\mathcal{X}$ to $\mathcal{T}$.

The kernel $P_{T|X}$ can be viewed as transforming $X \sim P_X$ into a representation of $T \sim P_T(\cdot) \triangleq \int P_{T|X}(\cdot|x)dP_X(x)$ in the $\mathcal{T}$ space.

We seek for $P_{T|X}$ s.t extracts information about Y , i.e., high $I(Y;T)$, while maximally compressing X, which is quantified as keeping $I(X;T)$ small.

the IB problem is formulated through the constrained optimization :

$$\inf_{P_{T|X}:I(Y;T)\geq\alpha} I(X;T)$$

α is a parameter which capture the trade-off between $I(X;T)$ and $I(Y;T)$.

Lagrange Dual Form

$\mathcal{L}_\beta(P_{T|X}) \triangleq I(X;T) - \beta I(Y;T)$

minimize $\mathcal{L}_\beta(P_{T|X})$ over all possible $P_{T|X}$ kernels.

Varying $\beta \in [0, \infty)$ regulates the tradeoff between informativeness and compression.

as $\beta \to \infty$, this problem is equivalent to the MSS optimization problem

β and its relation to bias-variance trade-off

Self-Consistent Equations

The Optimal assignment that minimizes $\mathcal{L}_\beta(P_{T|X})$, satisfies the equation:

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} \cdot exp\left(-\beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y|t)}\right)$$

where $Z(\cdot,\cdot)$is the normalization factor; due to the Markov chain condition $Y \to X \to T$,

$$p(y|t) = \frac{1}{p(t)} \sum_x p(y|x) \cdot p(t|x) \cdot p(x)$$

The IB Iterative Algorithm

every iteration (i) consists of three steps:

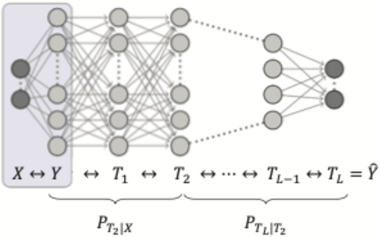
$$p_i(t|x) \leftarrow \frac{p_i(t)}{Z_i(x,\beta)} \cdot exp\left(\beta d(x,t)\right)$$

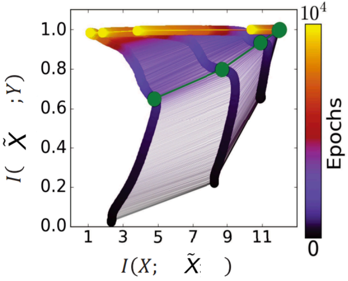
$$p_{i+1}(t) \leftarrow \sum_x p(x) p_i(t|x)$$

$$p_{i+1}(y|t) \leftarrow \sum_y p(y|x) p_i(x|t)$$

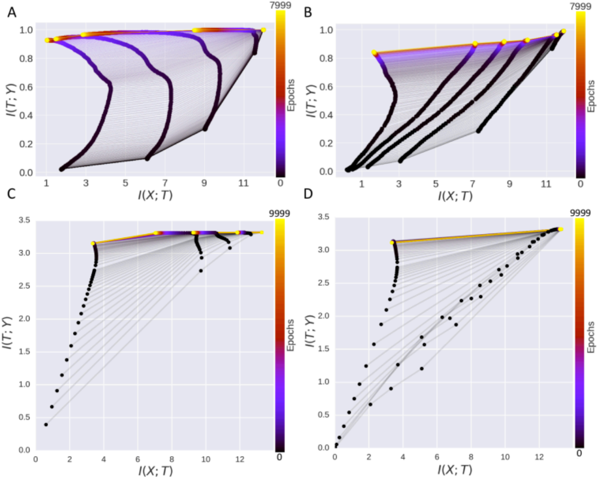
The IB method and Deep Neural Networks

Information Plane

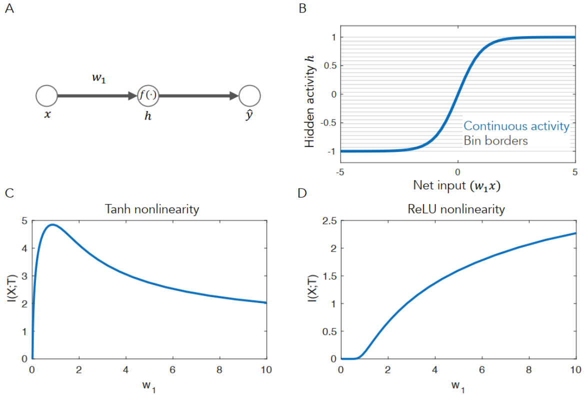




Information Plane with different Activation functions



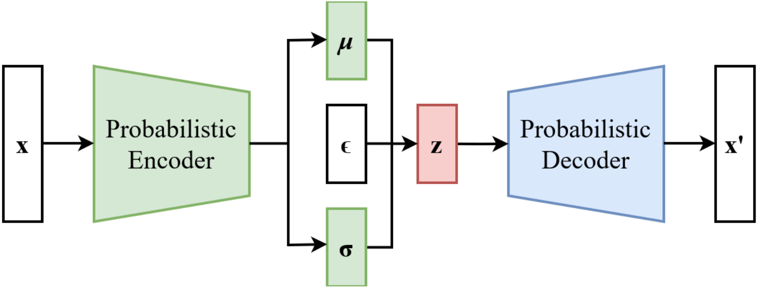
Information Plane with different Activation functions



Information Bottleneck and SGD

- continuous outputs causes infinite values of MI
- the binning method and the Data Processing Principle
- applying binning to the actual activation values
- make the NN stochastic (Amjed et al. 2018, section 5.4)

Variational Auto-Encoders



$Z \sim p_0(Z)$
objective : $\tilde{X} \sim p_\phi(\tilde{x}|x)p(X)$ close to $p_0(z)$ and $X' \sim p_\theta(x'|z)p_0(z)$ close to X

Variational Auto-Encoder

- an alternative representation :

$$\min_{p(\tilde{x}|x)} I(\tilde{X};X) + \beta H(Y|\tilde{X})$$

- approximating the loss function:

$$\mathcal{L}_{approx.} \triangleq \frac{1}{N}\sum_{i=1}^{N} \underbrace{\mathbb{E}_{p(\tilde{x}|x^{(i)})}\left[-\log \mathbb{P}\left[y^{(i)}|\tilde{x}\right]\right]}_{\text{Cross Entropy Loss}} + \beta \underbrace{D_{\mathrm{KL}}(P(\tilde{x}|x^{(i)})||P(\tilde{x}))}_{\textit{regularization}}$$

- a loss function for VAE:

$$\mathcal{L}(\phi,\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N} \underbrace{\mathbb{E}_{p(\tilde{x}|x^{(i)})}\left[-\log \mathbb{P}\left[x^{(i)}|\tilde{x}\right]\right]}_{\text{marginal likelihood of data}} + \underbrace{D_{\mathrm{KL}}(P(\tilde{x}|x^{(i)})||P_0(\tilde{x}))}_{\text{distance to the prior} P_0}$$

differences of VAE and the IB method

- $\beta = 1$
- separate optimization

Beyond the IB method

- $I(X;T)$ vs $I(S;\mathcal{A}(S))$
- a bound for generalization error :
  $\mathbb{P}\left[|err_{test} - err_{train}| > \epsilon\right] < O\left(\frac{I(S;\mathcal{A}(S))}{n\epsilon^2}\right)$

References I

Hassan Hafez-Kolahi and Shohreh Kasaei (2019)
Information Bottleneck and its Applications in Deep Learning
Journal of Information Systems and Telecommunication

Ziv Goldfeld and Yury Polyanskiy (2020)
The Information Bottleneck Problem and Its Applications in Machine Learning
CoRR

Naftali Tishby, Fernando C. Pereira and William Bialek (2000)
The information bottleneck method
arXiv

Andrew Michael Saxe and Yamini Bansal and Joel Dapello and Madhu Advani and Artemy Kolchinsky and Brendan Daniel Tracey and David Daniel Cox (2018)
On the Information Bottleneck Theory of Deep Learning
International Conference on Learning Representations

Rana Ali Amjad and Bernhard C. Geiger (2018)
How (Not) To Train Your Neural Network Using the Information Bottleneck Principle
CoRR