1 Information Theory basics

Here we will give all the required definitions and their properties.

Definition 1.1. Entropy H(X) of a discrete random variable X with probability mass function $p(X)$ is the quantity

$$H(X) = \mathop{\mathbb{E}}_{p(x)} [-\log p(x)]$$

Definition 1.2. Differential entropy of a random variable X with probability density function $p(x)$ is the quantity

$$H(X) = \mathop{\mathbb{E}}_{p(x)} [-\log p(x)]$$

Mutual information:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= \mathrm{D}_{\mathrm{KL}}[p(x,y) \parallel p(x)p(y)] \\ &= \mathrm{TC}(p(x,y)) \\ &= \mathop{\mathbb{E}}_{x} [\mathrm{D}_{\mathrm{KL}}[p(y|x) \parallel p(y)]] \end{aligned}$$

Chain rule for MI:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$$

Conditional MI identities:

$$\begin{aligned} I(X;Y|Z) &= \mathop{\mathbb{E}}_{Z} [\mathrm{D}_{\mathrm{KL}}[p(x,y|z) \parallel p(x|z)p(y|z)]] \\ &= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \\ &= H(X|Z) - H(X|Y,Z) \\ &= H(X|Z) + H(Y|Z) - H(X,Y|Z) \end{aligned}$$

Conditional entropy:

$$H_p(y|z) := \mathbb{E}_{y,z\sim p(y,z)}[-\log p(y|z)]$$

Conditional cross-entropy:

$$H_{p,q}(y|z) := \mathbb{E}_{y,z\sim p(y,z)}[-\log q(y|z)]$$

2 Information Bottleneck as such

2.1 Origins and formulation

Information Bottleneck origins from rate-distortion theory, which is solving the following task: given a random variable X and some *distortion* function $d(x,\tilde{x})$ quantize X into $\tilde{X}$ such that $\tilde{X}$ is compressed as much as possible, but not too corrupted. Strictly speaking, we are trying to solve

$$\min_{p(\tilde{x}|x)} I(\tilde{X};X) \quad \text{s.t. } D(X,\tilde{X}) \le D^*,$$

where

$$D(X,\tilde{X}) = \mathop{\mathbb{E}}_{p(x,\tilde{x})} [d(x,\tilde{x})]$$

and D^* is the maximum value of possible distortion that we permit. MSE loss or Hamming distance are common choices for $d(x,\tilde{x})$.

We can find optimal quantization by solving the variational problem:

$$\mathcal{L}(p(\tilde{x}|\tilde{x})) = I(X;\tilde{X}) + \beta D(X,\tilde{X})$$

The main problem with rate-distortion theory is that we need a distortion function $d(x,\tilde{x})$, which is difficult to specify for complex structured data, such as video or speech. And here Information Bottleneck comes to the rescue: what if after reconstruction we are only interested in some variable Y? Then we can reformulate our problem as

$$\mathcal{L}(p(\tilde{x}|\tilde{x})) = I(X;\tilde{X}) - \beta I(Y,\tilde{X})$$

For discrete case it can be shown [5], that it's a special case of rate-distortion problem with $d(x,\tilde{x}) = \mathrm{D}_{\mathrm{KL}}[p(y|x) \parallel p(y|\tilde{x})]$.

2.2 A few words about sufficient statistics

- Two datasets which give us the same inference about sufficient statistic, would give the same inference about underlying parameter θ.
- Any injective function of sufficient statistic is also a sufficient statistic.
- Factorization theorem says, that $p(x|y) = h_T(X)g_T(T(X), y)$, dependence in g_T in x is only through $T(x)$.

2.3 Sufficiency and minimality of representations

From now on we'll use symbol Z for variable $\tilde{X}$, because it's more consistent with modern DL literature.

We have the following optimization problem:

$$\mathcal{L}(p(x; z)) = I(X; Z) - \beta I(Z; Y)$$

There is an interesting statement which connects notion of *sufficiency* and *minimality* of Z in the $\beta \to \infty$ setting. Let denote by $F(X)$ all random mappings of X (this means, that for $f \in F(X)$ we have Markov chain $Y \to X \to f(X)$). Let $S_X(Y)$ be a set of all sufficient statistics of X for Y.

Proposition 1. *If Z is a solution to*

$$\min_Z I(X; Z) \quad s.t.\ I(Z; Y) = \max_{Z'} I(Z'; Y)$$

then Z is a minimal sufficient statistics of X for Y

$$P(X|Z, Y) = P(X|Z)$$

In other words, we are getting a minimal sufficient representation by optimizing Lagrangian in the limit $\beta \to \infty$. The proof of this theorem is split into two lemmas [4].

Lemma 1. *Z is a sufficient statistic of X for Y iff $I(Z; Y) = I(X; Y)$.*

Proof. Imagine that $T \in S_X(Y)$. For any $Z \in F(X)$ we have a Markov chain $Y \to X \to Z$, so by DPI we have $I(Y; Z) \leq I(Y; X)$. But by sufficient statistic property we have $P(X|Z, Y) = P(X|Y)$, so we have a Markov chain $Y \to Z \to X$. Again by DPI we have $I(Y; Z) \leq I(Y; X)$, so $I(Y; Z) = I(Y; X)$.

Now consider $Z = f(X)$ for some $f \in F(X)$ such that $I(Y; Z) = I(Y; X)$. As $Y \to X \to Z$ is a Markov chain, then by definition of conditional MI we have:

$$\begin{aligned} I(Y : Z|X) &\triangleq \mathop{\mathbb{E}}_X [\mathrm{D_{KL}}[p(Y, Z|X) \parallel p(Y|X)p(Z|X)]] \\ &= \mathop{\mathbb{E}}_X [\mathrm{D_{KL}}[p(Z|X, Y)p(Y|X) \parallel p(Y|X)p(Z|X)]] \\ &= \mathop{\mathbb{E}}_X [\mathrm{D_{KL}}[p(Y|X)p(Z|X) \parallel p(Y|X)p(Z|X)]] = 0 \end{aligned}$$

Now, by chain rule for MI we have:

$$I(Y : X, Z) = I(Y : X) + I(Y, X|Z) = I(Y : X) + I(Y, Z|X) \Longrightarrow I(Y, X|Z) = I(Y, Z|X) = 0$$

Applying definition of conditional MI again we get $p(Y, X|Z) = p(Y|Z)p(X|Z)$. And this means that $Z \in S_X(Y)$:

$$p(X|Y, Z) = \frac{p(X, Y|Z)}{p(Y|Z)} = \frac{p(X|Z)p(Y|Z)}{p(Y|Z)} = p(X|Z)$$

□

Let's denote by $S_X^*(Y)$ a set of minimal sufficient statistics of X for Y.

Lemma 2. *Let $Z \in S_X(Y)$, then*

$$Z \in S_X^*(Y) \Longleftrightarrow I(X; Z) = \min_{T \in S_X(Y)} I(X; T)$$

Proof. First, let Z be a minimal sufficient statistic. Then for any other sufficient statistic T we have $Z = f(T)$ for some f. Then we get a Markov chain $X \to T \to Z$ and by DPI we have $I(X;Z) \leq I(X;T)$.

Now, let's prove reverse direction of the claim. Imagine, that $Z \in S_X(Y)$ but is not minimal. We are going to show, that $\exists T \in S_X(Y)$ such that $I(X;Z) > I(X;T)$. By Fisher-Neyman factorization theorem we have

$$Z \in S_X(Y) \iff \exists\ h_Z, g_Z \text{ s.t. } \forall x, y \quad p(x|y) = h_Z(x) g_Z(Z(x), y)$$

Let's define an equivalence relation

$$a \sim b \iff \forall y \exists \lambda s.t. \frac{g_Z(a,y)}{g_Z(b,y)} = \lambda(a,b)$$

Now we define a deterministic function $T : \mathcal{X} \to \mathcal{Z}$ such that $\forall x : T(x) = \bar{z}$ — a representative of $[Z(x)]$ (TODO: exists by axiom of choice?). Let's prove, that it is a sufficient statistic. For this let's define

$$h_T(x) \triangleq h_Z(x) \frac{g_Z(Z(x), y)}{g_Z(T(x), y)}$$
$$g_T(T(x), y) \triangleq g_Z(T(x), y)$$

Then we have

$$p(x|y) = h_Z(x) g_Z(Z(x), y) = h_Z(x) \frac{g_Z(Z(x), y)}{g_T(T(x), y)} g_T(T(x), y) = h_T(x) g_T(T(x), y)$$

Hence T is a sufficient statistic.

Now let's show that $I(X;Z) > I(X;T)$. Since Z is not minimal, then there is such $R \in S_X(Y)$ that Z is not a function of R. Let's show, that T is a function of R (btw, this will show, that T is minimal). For this we are going to show that if $R(x_1) = R(x_2)$ then $T(x_1) = T(x_2)$: this would allow us to build a function $\phi : \mathcal{R} \to \mathcal{T}$ which just take value r, find it's preimage $R^{-1}(r)$, take any sample $x \in R^{-1}(r)$ and compute $T(x)$. For any x_1, x_2 such that $R(x_1) = R(x_2)$ we have:

$$\begin{aligned}
\frac{g_Z(Z(x_1), y)}{g_Z(Z(x_2), y)} &= \frac{p(x_1|y) h_Z(x_2)}{p(x_2|y) h_Z(x_1)} \\
&= \frac{h_R(x_1) g_R(R(x_1), y) h_Z(x_2)}{h_R(x_2) g_R(R(x_2), y) h_Z(x_1)} \\
&= \frac{h_R(x_1) g_R(R(x_1), y) h_Z(x_2)}{h_R(x_2) g_R(R(x_1), y) h_Z(x_1)} \\
&= \frac{h_R(x_1) h_Z(x_2)}{h_R(x_2) h_Z(x_1)} \\
&= \lambda(Z(x_1), Z(x_2))
\end{aligned}$$

This means that $Z(x_1) \sim Z(x_2)$, which in turn means that $T(x_1) = T(x_2)$, so T is minimal sufficient statistic and a function of R. □

3 First steps: SZT experiments, critics

3.1 Main ideas

First attempts to apply IB to deep learning are to Ravid Shwartz-Ziv and Naftali Tishby [6]. They proposed to view at a NN as a Markov chain

$$Y \to X \to Z_1, ..., \to Z_l$$

where X is an input, and Z_i is our hidden representations.

SZT theory claims that NNs implicitly minimizes IB Lagrangian for each layer and does the following claims:

- There are two phases of training: fitting (when MI with input grows) and compression (when MI with input decreases).

- Compression results in good generalization.
- Compression occurs due to diffusion-like behaviour of SGD (looks like it was recently proved by [1]).

They give nice pictures for this. TODO: plot nice pictures of $I(X;Z_i)$ growing, then decreasing and $I(Z_i;Y)$ just growing. Do not forget, that by DPI further layers should be lower for both $I(Z_i;Y)$ plot and $I(X;Y)$ plot as our Markov chain implies $y \to x \to z_1 \to ... \to z_l$.

We define SNR for SGD as

$$\text{SNR} = \frac{m_l}{s_l} \quad m_l = \left\| \text{Mean}\left(\frac{\partial E}{\partial W_l}\right) \right\|_F \quad s_l = \left\| \text{Std}\left(\frac{\partial E}{\partial W_l}\right) \right\|_F$$

SZT claim that it this SGD behaviour was connected to fitting and compression phases. Pictures on SNR for SGD looks like

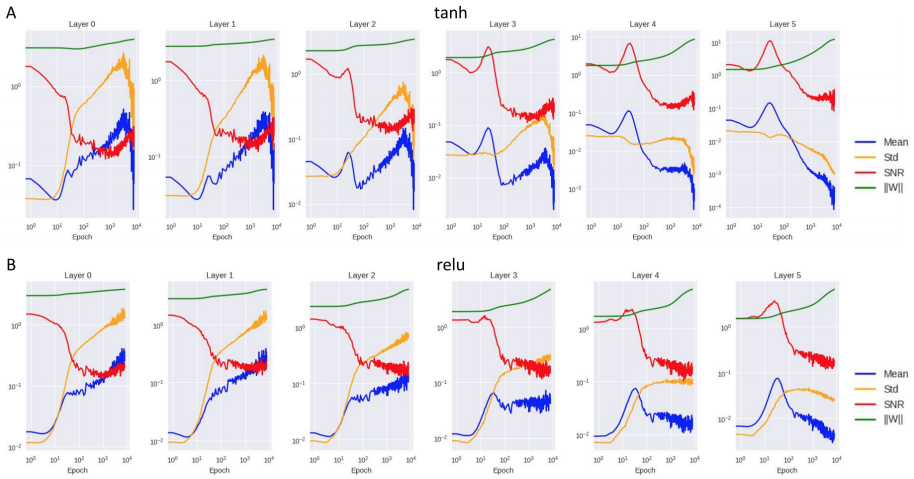

Figure 1: Figure 20 from [3]: Gradient SNR phase transition. (A) tanh networks trained in the standard setting of SZT show a phase transition in every layer. (B) ReLU networks also show a phase transition in every layer, despite exhibiting no compression.

3.2 Critics

But there are two problems with their experiments:

- Values of $I(X;Z_i)$ and $I(Y;Z_i)$ are either constant or infinite. That's why it's meaningless to measure them.
- Only toy experiments are performed. Authors claimed (in 2017) that they have CIFAR-10 experiments running, but no paper about this was published since.
- It does not work for ReLU, i.e. [3]. N.Tishby answer, that bad binning procedure was used, that's why authors got such results. But more sophisticated methods (Kraskov k-means and KDE) of estimating MI didn't lead to something better. Besides, they built ReLU network which exhibit right SNR behaviour, but didn't compress.

3.3 Why MI is either constant or infinite

Proposition 2. *Consider a discrete distribution $x \sim p(x)$ (possible with countable support) and a NN*

$$f(x) = \sigma(...\sigma(W_2\sigma(W_1x + b_1) + b_2)...)$$

with some injective non-linearity σ (sigmoid, tanh, LeakyReLU, SELU, etc). Then there is only measure-zero set of weights $W_1, ..., W_L, b_1, ..., b_L$ s.t. for some $x_i \neq x_j$ from $p(x)$ we get $f(x_i) = f(x_j)$. This means, that NN is injective for $p(x)$.

Proposition 3. *If NN is deterministic, then $I(z_l;x)$ is either constant (for descrete x) or infinite (for continuous x), regardless of training.*

Proof. Consider a discrete setting. Then

$$I(z_l; x) = H(x) - H(x|z_l)$$

We are going to prove that $H(x|z_l) \equiv 0$:

$$H(x|z_l) = H(x|x) = 0$$

We used the fact that since z_l is an injective and deterministic mapping of x, then we can invert it (it is clearly surjective by design: we use only support of z_l).

TODO: prove for continuous distribution. Looks like we'll need local injectivity (which should work for ReLU too since they are piecewise-linear?). And show that conditional entropy is $-\infty$ for some regions. □

3.4 Remarks

Remark 1. *Authors in [2] claim that we measure not MI, but some other quantity, connected to clusterization. So although we have just proved that it's not possible for MI to change actually, another term is changing and it's changing for ReLU too. Unfortunately, proofs about clusterization are only empirical (yet).*

Remark 2. *Actually, their claims about SGD are hold true and are observed in other works.*

Remark 3. *Actually, N.Tishby claims in his talk, that they have rigorous proofs about SGD behaviour and compression generalization bounds. It was pusblished on ICLR 2019, but got somewhat "retracted".*

4 Disentangled representations

The following theorem gives us a hope to build invariant and disentangled representations. Informally, it says that if we can build sufficient and minimal representations (which we are building by optimizing IB Lagrangian, for example), then we get invariant and disentangled ones.

Proposition 4. *If η is a nuisance for the task y and z is a sufficient representation of x and we have a Markov chain $\eta \to x \to z$, then*

$$I(z; \eta) \leq I(z; x) - I(x; y)$$

Moreover, if y is discrete, then we can use task-decomposition lemma and prove something more strict:

$$I(z; \eta) = I(z; x) - I(x; y) - \epsilon,$$

where $\epsilon \triangleq I(z; y|\eta) - I(x; y)$.

Proof. As we have a Markov chain $(y, \eta) \to x \to z$ then by DPI $I(z; y, \eta) \leq I(z; x)$. By chain rule we have

$$I(z; \eta) = I(z; y, \eta) - I(z; y|\eta) \leq I(z; x) - I(z; y|\eta)$$

By definition of nuisance $y \perp \eta$ so $I(z; y|\eta) \geq I(z; y)$, because (by one of identities for conditional MI):

$$\begin{aligned} I(z; y|\eta) &= H(y|\eta) - H(y|z, \eta) \\ &= H(y) - H(y|z, \eta) \\ &\geq H(y) - H(y|z) \\ &= I(z; y) \end{aligned}$$

As z is sufficient, i.e. $I(x; y) = I(z; y)$ we obtain

$$I(z; \eta) \leq I(z; x) - I(z; y|\eta) \leq I(z; x) - I(z; y) = I(z; x) - I(x; y)$$

Now consider, that we have $p(x; y)$ and y is discrete. Then by task-nuisance decomposition lemma we can introduce a nuisance η s.t. $x = f(y; \eta)$ and f is deterministic. That's why we have

$$I(z; x) = I(z; y, \eta) = I(z; \eta) + I(z; y|\eta)$$

Rearranging terms we get

$$I(z;\eta) = I(z;x) - I(x;y|\eta) = I(z;x) - I(x;y) - \underbrace{(I(x;y|\eta) - I(x;y))}_{\varepsilon}$$

Let's prove that $0 \leq \varepsilon \leq H(y|x)$. By definition of ε:

$$\varepsilon = I(z;y|\eta) - I(x;y)$$

Since $y|\eta \to x|\eta \to z|\eta$ is a Markov chain, then by DPI:

$$\begin{aligned} &\leq I(x;y|\eta) - I(x;y) \\ &= H(y|\eta) - H(y|x,\eta) - H(y) + H(y|x) \\ &= H(y) - H(y|x,\eta) - H(y) + H(y|x) \\ &= H(y|x) - H(y|x,\eta) \end{aligned}$$

The last inequality is subtle, actually. We use the fact $H(y|x) - H(y|x,\eta) = I(y;\eta,x) \geq 0 \Rightarrow H(y|x) \geq H(y|x,\eta)$. And $\epsilon \geq 0$ follows from equation proved above: $I(z;y|\eta) \geq I(z;y) = I(x;y)$. That's why $H(y|x) \geq 0$, because it's a valid upper bound.

$$\leq H(y|x)$$

□

What does it give us? Does it give us invariance and disentanglement? First of all, we see, that as more minimal our z the more invariant it is, because minimal sufficient statistic implies that $I(x;z)$ is the most minimal possible! And the first term is the only term we can influence with our z.

Unfortunately, it does not give us disentanglement yet. To ensure disentanglement we need additional assumptions. But before diving into it, let's think about how can we achieve invariance with what we already now?

Corollary 1. *Minimizing IB Lagrangian with $\beta \to 0$*

$$H(y|z) + \beta I(x;z)$$

Proof. This is true by theorem proved above. □

Introducing bottlenecks by adding noise or reducing dimensionality. TODO: shouldn't we prove that bottlenecks reduces MI?

Corollary 2. *Imagine we have a Markov chain $(y,\eta) \to x \to z_1 \to z_2$ and $I(z_1;z_2) < I(x;z_1)$, i.e. there is some bottleneck on the road between $z_1 \to z_2$. Then if z_2 is sufficient, it is more invariant*

Bottlenecks promote invaraince.

$$I(z_2;\eta) \leq I(z_2;z_1) - I(z_1,y) = I(z_2;z_1) - I(x,y) < I(z_1;x) - I(x,y)$$

□

Corollary 3. *Imagine setup above. If y is discrete and is a deterministic function of x, then inequality is strict.*

Proof. We have

$$I(z_2;\eta) \leq I(z_2;z_1) - I(x;y) < I(z_1;x) - I(x;y) = I(z_1;\eta) + \varepsilon$$

Since y is a deterministic function of x then $\varepsilon = 0$, hence the desired result. □

Corollary 4 (Stacking increases invariance)**.** *Consider we a have a Markov chain of layers*

$$(y,\eta) \to x \to z_1 \to ... \to z_l$$

If z_L is still sufficient, then it's more invariant.

TODO: Well, this is because we have a Markov chain $\eta \to ... \to z_l$, it's not a corollary: $I(\eta;z_l) \leq I(\eta;z_i)$.

5 Information in the weights

5.1 Properties

5.2 Minimum description length

5.3 Connection to representations

6 Mutual information estimators

6.1 Binning

6.2 MINE

6.3 MI estimator under gaussian convolutions

7 Bibliography

References

[1] Pratik Chaudhari and Stefano Soatto. "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=HyWrIgW0W.

[2] Ziv Goldfeld et al. "Estimating Information Flow in Neural Networks". In: *CoRR* abs/1810.05728 (2018). arXiv: 1810.05728. URL: http://arxiv.org/abs/1810.05728.

[3] Andrew Michael Saxe et al. "On the Information Bottleneck Theory of Deep Learning". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=ry_WPG-A-.

[4] Ohad Shamir, Sivan Sabato, and Naftali Tishby. "Learning and generalization with the information bottleneck". In: *International Conference on Algorithmic Learning Theory*. Springer. 2008, pp. 92–107.

[5] Naftali Tishby, Fernando C. Pereira, and William Bialek. "The Information Bottleneck Method". In: 1999, pp. 368–377.

[6] Naftali Tishby and Noga Zaslavsky. "Deep Learning and the Information Bottleneck Principle". In: *CoRR* abs/1503.02406 (2015). arXiv: 1503.02406. URL: http://arxiv.org/abs/1503.02406.

Exercises

1. Prove that

$$\arg \min_{q_1(x_1),...,q_n(x_n)} \mathrm{D}_{\mathrm{KL}}[p(x_1,...,x_n) \parallel q_1(x_1)\cdot...\cdot q_n(x_n)] = p(x_1)\cdot...\cdot p(x_n) \iff q_1(x_1),...,q_n(x_n) = p(x_1),...,p(x_n)$$

2. Prove that if X, Y are normally distributed, then Z is normally distributed too.

3. Maybe something on finding differential entropy of r.v. with infinite differential entropy. Or constructing such a variable.

4. Prove that $I(X;T)$ is either infinity or constant (except measure zero of weights).

5. Prove DPI

6. IB solution is a minimal sufficient statistic (based on http://www.cs.huji.ac.il/labs/learning/Papers/ibgen.pdf, theorem 5)