1 Information Theory basics

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

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

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

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

Definition 1.3. *Mutual information* identities:

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

Proposition 1. *Chain rule for MI:*

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

Proof.

$$\frac{p(x,y,z)}{p(x)p(y,z)} = \frac{p(x,z)p(y|x,z)}{p(x)p(z)p(y|z)} = \frac{p(x,z)}{p(x)p(z)} \frac{\frac{p(x|y,z)p(y|z)}{p(x|z)}}{p(y|z)} = \frac{p(x,z)}{p(x)p(z)} \frac{p(x|y,z)}{p(x|z)}$$

□

Definition 1.4. *Conditional MI* identities:

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

Definition 1.5. *Conditional entropy*:

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

Proposition 2. *Conditioning reduces entropy.*

Proof. $H(x|y,z) = H(x|z) - I(x:y|z)$ and $I(x:y|z) \geq 0$. □

Definition 1.6. *Conditional cross-entropy*:

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

Proposition 3 (Data-processing inequality (DPI))**.** *If we have a Markov chain $X \to Y \to Z$, then $I(X;Z) \leq I(Y;Z)$.*

Proof. By chain rule we have:

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

Since we have a Markov chain, then $I(X:Z|Y) = 0$:

$$I(X:Z|Y) = \mathop{\mathbb{E}}_{Y} \left[\mathop{\mathbb{E}}_{X|Y} [\mathrm{D}_{\mathrm{KL}}[p(z|x,y) \parallel p(z|y)]] \right] = \mathop{\mathbb{E}}_{Y} \left[\mathop{\mathbb{E}}_{X|Y} [\mathrm{D}_{\mathrm{KL}}[p(z|y) \parallel p(z|y)]] \right] = 0$$

And we have this form of KL, because

$$\frac{p(x,z|y)}{p(x|y)p(z|y)} = \frac{p(z|x,y)p(x|y)}{p(x|y)p(z|y)} = \frac{p(z|x,y)}{p(z|y)}$$

□

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

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

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 [7], 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 $T(x)$.
- By definition we have R is minimal iff $\forall T \in S_Y(X)$ there exists such function f that $R = f(T)$.

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 $S_Y(X)$ be a set of all sufficient statistics of X for Y.

Proposition 4. *If Z is a solution to*

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

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 [6].

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

Proof. Imagine that $T \in S_Y(X)$. 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:

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

Now, by chain rule for MI we have:

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

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

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

□

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

Lemma 2. *Let $Z \in S_Y(X)$, then*

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

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 $Z \in S_Y(X)$ but is not minimal. We are going to show, that $\exists T \in S_Y(X)$ such that $I(X;Z) > I(X;T)$. By Fisher-Neyman factorization theorem we have

$$Z \in S_Y(X) \Longleftrightarrow \exists\ h_Z, g_Z \text{ s.t. } \forall x, y \quad p(x|y) = h_Z(x) g_Z(Z(x), y)$$

Let's define an equivalence relation

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

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

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

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 $R \in S_Y(X)$ 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 [8]. 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 [2]).

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

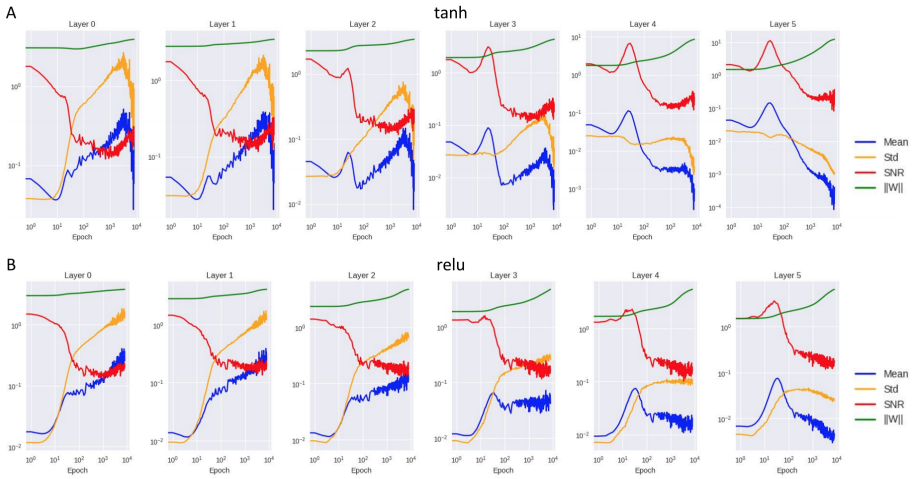


Figure 1: Figure 20 from [5]: Gradient SNR phase transition. (A) tanh networks trained in the standard setting of SZT show a phase transition in every layer. (B) ReLU networks also show a phase transition in every layer, despite exhibiting no compression.

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. [5]. 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

See homework.

3.4 Remarks

Remark 1. *Authors in [3] 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 Invariant 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 (main drawback is necessity of $I(z:y) = I(x:y)$).

Proposition 5. *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 (can be shown just by definition of MC), then by DPI:

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

Since $\varepsilon \geq 0$ (and $H(y|x) \geq 0$, since y is discrete):

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

□

What does it give us? Does it give us invariance and disentanglement? We see that

$$I(z;\eta) \approx \underbrace{I(z;x)}_{\text{minimality}} - \underbrace{I(z;y)}_{\text{sufficiency}}$$

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.

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. □

By the way, we can show right away, just by DPI, that:

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 (well, actually it just have a bound, which is not worse).

5 Information in the weights

Information decomposition

Consider a data distribution $(X, Y) = D \sim p(D|\theta)$. We say that our model is:

$$p(D, \theta, w) = p(\theta)p(D|\theta)p_\phi(w|D) \tag{1}$$

for some true unknown parameters ϕ. Here $p_\phi(w|D)$ is our distribution on the weights of NN — it is not deterministic due to the randomness of the optimization process. From this we can see that $p_\phi(w|D)$ is different from $p(w|D)$, which we would be computed if we'd have $p(D|w)$ to define our model (which is not). This is a very big source of confusion.

If we have some weights and an input X, then our model defines a distribution $q(Y|X, w)$. We want it to be as close as possible to $p(Y|X, w)$, which is computed from (1).

Theorem 1. *Let $D = (X, Y)$ be a fixed dataset. Then*

$$H_{p,q}(Y|X, w) = \underbrace{H(Y|X, \theta)}_{\textit{intrinsic error}} + \underbrace{I(\theta; Y|X, w)}_{\textit{sufficiency}} + \underbrace{\mathop{\mathbb{E}}_{X,w} \left[D_{KL}[p(Y|X, w) \parallel q(Y|X, w)]\right]}_{\textit{efficiency}} - \underbrace{I(Y, w|X, \theta)}_{\textit{overfitting}} \tag{2}$$

Proof. Cross-entropy can be written as

$$\begin{aligned} H_{p,q}(Y|X, w) &= \mathop{\mathbb{E}}_{p(Y,X,w)} \left[-\log q(Y|X, w)\right] \\ &= \mathop{\mathbb{E}}_{p(Y,X,w)} \left[-\log p(Y|X, w)\right] + \mathop{\mathbb{E}}_{p(Y,X,w)} \left[\log \frac{p(Y|X, w)}{q(Y|X, w}\right] \\ &= H_p(Y|X, w) + \mathop{\mathbb{E}}_{X,w} \left[\mathrm{D_{KL}}[p(Y|X, w) \parallel q(Y|X, w)]\right] \end{aligned} \tag{3}$$

Let's prove the rest. Since $I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$ (analog of $I(X; Y) = H(X) - H(X|Y)$), we can see that:

$$H_p(Y|X, w) = I(Y; \theta|X, w) + H(Y|X, w, \theta)$$

Applying the same identity to $I(Y; w|X, \theta)$ we obtain

$$H(Y|X, w, \theta) = H_p(Y|X, \theta) - I(Y; w|X, \theta)$$

hence the desired result:

$$H_p(Y|X, w) = I(Y; \theta|X, w) + H_p(Y|X, \theta) - I(Y; w|X, \theta)$$

□

Why do we call the terms so?

1. $H(Y|X,\theta)$ — intrinsic error: just some constant, which does not depend on w. This is the error we would get if we would know true data distribution.

2. $I(\theta; Y|X,w)$ — sufficiency: how much information does w know about θ when it sees a dataset. Why is it so? Remember the definition of sufficiency that we had: $P(A|B,\psi) = P(A|B)$ which is equivalent to $I(A:\psi|B) = H(A|B) - H(A|B,\psi) = 0$. Let's apply this to $p(Y|\theta,w,X)$. We want to measure $I(Y:\theta|w,X) = H(Y|w,X) - H(Y|\theta,w,X)$. If it is low, then we have $p(Y|\theta,w,X) \approx p(Y|w,X)$, i.e. θ is very well captured by w, i.e. w is a sufficient statistic for θ of Y (we assume, that we always know X). Besides, there is an often assumption about $p(X,\theta) = p(X)p(\theta)$, i.e. features X are generated independently.

3. $\mathop{\mathbb{E}}_{X,w}\left[\mathrm{D_{KL}}[p(Y|X,w) \parallel q(Y|X,w)]\right]$ — efficiency: how efficient is it to use neural networks of the kind f_w to approximate true data distribution. It shows how close our approximation $q(Y|X,w)$ is to a true $p(Y|X,w)$ data distribution, obtained from model (1).

4. $I(Y,w|X,\theta) = H(Y|X,\theta) - H(Y|w,X,\theta)$ — overfitting: shows, how much uninformative information does w memorizes about Y. Why it shows only uninformative information? Because in this term we assume that both θ and X are known. So, imagine, that we know θ and X — this means, that we know everything good we can about $p(Y|X,\theta)$. But the weights w are trying to memorize something more.

So we see that NN can potentially solve the task (obtain good values of loss) just by maximizing overfitting term, i.e. memorizing the dataset. For successful training variable w should only memorize θ, i.e. we need $I(w;D) = I(D;\theta) \leq H(\theta)$. For overfitting we need the term $I(\boldsymbol{y}; w|\boldsymbol{x},\theta) \leq I(w;D|\theta)$ to grow linearly with the size of the dataset. It's obvious that $I(\boldsymbol{y}; w|\boldsymbol{x},\theta) \leq I(w;D|\theta)$, because we use more information in RHS (if you want it strictly: just apply a chain rule). It's also obvious that it should grow linearly, because it factorizes into sum of MIs (because empirical distribution factorizes). So what we do is just adding a regularization term

$$L(\phi) = H_{p,q}(Y|X,w) + \beta I(w;D)$$

Note that this term is not the same thing as $I(w; y|\theta, X)$, but we cannot do better.

Unfortunately, $I(w;D)$ is intractable to compute (we need an expectation over all datasets and all training procedures), so we add an upper bound

$$\begin{aligned} I(w;D) &= \mathop{\mathbb{E}}_{D}\left[\mathrm{D_{KL}}[p_\phi(w|D) \parallel p_\phi(w)]\right] \\ &\leq \mathop{\mathbb{E}}_{D}\left[\mathrm{D_{KL}}[p_\phi(w|D) \parallel p_\phi(w)]\right] + \mathrm{D_{KL}}[p_\phi(w) \parallel p(w)] \\ &= \mathop{\mathbb{E}}_{D}\left[\mathrm{D_{KL}}[p_\phi(w|D) \parallel p(w)]\right] \end{aligned} \tag{4}$$

Here $p(w)$ is any distribution on the weights. We would like to take factorized distribution, so it's easier to compute: $p(w) = \prod p(w_i)$. Ideally we would like to take factorized marginal $\prod p_\phi(w_i)$, because it's closest factorized distribution (see homework). We could approximate this distribution by training model several times, but it's also costly and we just choose a (somewhat arbitrary) prior. So, for some factorized prior $p(w) = \prod p(w_i)$ we have a loss

$$L(\phi) = H_{p,q}(Y|X,w) + \beta \mathrm{D_{KL}}[p_\phi(w|D) \parallel p(w)]$$

Model assumptions and upper bounding the loss

We choose our $p_\phi(w|D)$ to be parametrized as

$$w|D \sim \epsilon\hat{w}, \qquad \epsilon_i \sim \log\mathcal{N}(-\frac{\alpha_i}{2}, \alpha_i) \Longrightarrow \begin{cases} \mathbb{E}[\epsilon] & = e^{\mu+\sigma^2/2} = 1 \\ \mathrm{Var}[\epsilon] & = (e^{\sigma^2}-1)e^{2\mu+\sigma^2} = e^\alpha - 1 \end{cases}$$

We have chosen such a distribution for ϵ, so $\mathbb{E}[w_i|D] = \hat{w}_i$, where $\hat{w}_i$ is a learned mean. The main problem here is that our $p_\phi(w|D)$ is factorized, which makes it highly limited. It's easy to see that $p(w|D)$ is log-normal, since

$$\xi \sim \log\mathcal{N}(\mu, \sigma^2), c \in \mathbb{R} \Longrightarrow c \cdot \xi \sim \log\mathcal{N}(\mu + \ln c, \sigma^2)$$

(which follows from the fact that $c \cdot \xi = c \cdot e^{\mu+\sigma\tau} = e^{\ln c+\mu+\sigma\tau}$, where $\tau \sim \mathcal{N}(0,1)$).

Our prior distribution $p(w)$ is factorized log-uniform (i.e. logarithm of this distribution is uniform):

$$p(w) = \prod_{i=1}^{n} p(w_i) \qquad p(w_i) = \frac{c}{|w_i|}$$

Let's denote $\tilde{I}(w; D) = \mathrm{D}_{\mathrm{KL}}[p_\phi(w|D) \parallel \prod p(w_i)]$. The main benefit of our parametrization is that $\tilde{I}(w; D)$ now can be written in closed form up to an arbitrary constant:

$$\begin{aligned}
\tilde{I}(w; D) &= \mathrm{D}_{\mathrm{KL}}[p_\phi(w|D) \parallel \prod p(w_i)] \\
&= \mathrm{D}_{\mathrm{KL}}[p_\phi(\log w|D) \parallel \prod p(\log w_i)] \\
&= \mathrm{D}_{\mathrm{KL}}[\mathcal{N}(-\frac{\alpha}{2}, \alpha I) \parallel c^k] \\
&= H(\mathcal{N}(-\frac{\alpha}{2}, \alpha I)||c^k) - H(\mathcal{N}(-\frac{\alpha}{2}, \alpha I)) \\
&= \mathbb{E}\left[\log c^k\right] - \sum_{i=1}^{k} \frac{1}{2} \log 2\pi e\alpha_i \\
&= -\sum \frac{1}{2} \log \alpha_i + const
\end{aligned} \tag{5}$$

Here we used the fact that $H(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{2} \log 2\pi e\sigma^2$. One can note that MI can be negative, although derivation may seem plausible. Remember that $\mathrm{D}_{\mathrm{KL}}[p \parallel q]$ is defined only for absolutely continuous distributions ($q(x) = 0 \Rightarrow p(x) = 0$), which is not the case when p is Normal and q is uniform, unless we have a uniform distribution on the whole $\mathbb{R}$, which is improper.

As our prior is improper, this constant can be arbitrary. Authors say, that although this setup leads to arbitrary constant C, we could use truncated log-uniform or just gaussian multiplicative noise instead (in the former case MI would be intractable, but can be very closely approximated).

What we immediately see from this is that

$$I(w; D) \leq \tilde{I}(w; D) = -\frac{1}{2} \sum \log \alpha_i + const$$

So information in the weights is bounded. Let's define

$$J(w; D) = -\frac{1}{2} \sum \log \alpha_i$$

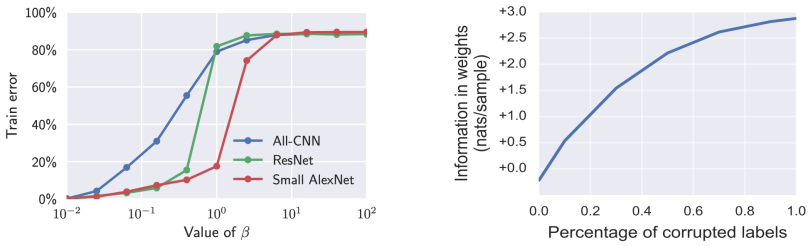
because this is what we'll use in practice and it's related to information.

Experiments

Next, authors conduct experiments similar in nature to [9]. The claim is the following: if the model is not allowed to memorize the data (if $I(w; D)$ is low) then it does not overfit. To prove the claim they do the following experiments:

1. Take CIFAR-10, randomly permute some portion of labels
2. Choose some value for β
3. Train and watch the values of α_i, because we are going to measure MI with them.
4. If Y is completely random, then $I(w; Y|X, \theta) = I(Y; w) \leq I(w; D)$. The first equality works because $H(Y|w, X, \theta) = H(Y|w)$. Remember our loss decomposition. When $\beta = 1$ we just cancel the overfitting term. When $\beta < 1$ it's always beneficial to overfit. When $\beta > 1$ it's beneficial to learn the true data. Their empirical results suggests that this really works in practice (even when we use upper bounding and a lot of assumptions).

Authors got the following plots (second plot is for $\beta = 0.1$):



It's interesting to see that model is forced to memorize more information when there are more corrupted labels. And it's very interesting that although we have an upper bound, $\beta = 1$ is still a critical regime.

Remark 4. *Authors do not compare to just a plain Dropout. Looks like it works too. In original paper by [9] InceptionV3 with dropout overfits very well on random ImageNet-1000, but authors do not say what dropout rate they used. In our setup for a very large value of β we force our posterior to be just prior, which makes it not surprising that we can't fit random labels. For high values of β can we fit non-random labels? Yes, because amount of information needed to fit them is much less.*

Duality of the bottleneck

The following three theorems suggests that by minimizing information in the weights we obtain both sufficiency and disentanglement.

Theorem 2. *Let $z = Wx$ and our model setup. Further assume that $q(z_i)$ and $q(z_i|x)$ are normally distributed. Then*

$$\begin{aligned} I(z;x) + TC(z) &= \sum_{i=1}^{\dim z} \mathop{\mathbb{E}}_{z} \left[D_{KL}[\mathcal{N}(\mu_i, \sigma_i^2) \parallel \mathcal{N}(\nu_i, s_i^2)] \right] \\ &= -\frac{1}{2} \sum_{i=1}^{\dim z} \mathop{\mathbb{E}}_{x} \left[\log \frac{Var[\epsilon]\langle \hat{W}_i^2, x^2 \rangle}{\hat{W}_i^\top \Sigma_x \hat{W}_i + Var[\epsilon]\langle \hat{W}_i^2, \mathbb{E}[x^2] \rangle} \right] \end{aligned} \tag{6}$$

Here W_i is the i-th row of W.

This means that $I(z;x) + TC(z)$ is a monotone decreasing function of α_i, because $\mathrm{Var}[\epsilon] = \tilde{\alpha} = e^{\alpha} - 1$ (because it's more beneficial for us to decrease α, so log term becomes $-\infty$ and since it's negative we get $+\infty$.

Proof. First prove for $\dim z = 1$ (TBD), then notice that

$$I(x;z) + TC(z) = \mathop{\mathbb{E}}_{z} \left[\mathrm{D_{KL}}[\prod q(z_i|x) \parallel \prod q(z_i)] \right] = \sum_{i=1}^{\dim z} \mathop{\mathbb{E}}_{x} [\mathrm{D_{KL}}[q(z_i|x) \parallel q(z_i)]] = \sum_{i=1}^{\dim z} I(z_i;x)$$

□

The following theorem shows that $I(z;x) + \mathrm{TC}(z)$ can be uniformly bounded by $I(w;D)$.

Theorem 3 (Uniform bound for one layer)**.** *Let $z = Wx$, use our model assumptions and assumptions from theorem above. Additionally assume that $\Sigma_x = Cov[x]$ is diagonal and that it's forth moment is finite. Then there is a strictly increasing function $g(\alpha) = \frac{1}{2}\log(1 - e^{-\alpha})$ s.t.*

$$g(\alpha) \le \frac{I(x;z) + TC(z)}{\dim z} \le g(\alpha) + c \tag{7}$$

Here $c = O(1/\dim x) \le 1$ and $\alpha \approx e^{-I(w;D)/\dim z}$ (which we got from (5)).

This means that $I(z;x) + \mathrm{TC}(z)$ is tightly bounded by $I(w;D)$ and increases strictly with it.

Proof. We are going to prove this theorem for $\dim z = 1$, because it will be straightforward how to generalize it. Let w is our weight row. First of all, since Σ_x is diagonal we have

$$\hat{w}^\top \Sigma_x \hat{w} = \sum w_i^2 \mathrm{Var}[x_i] = \sum w_i^2 (\mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2) \le \langle w^2, \mathbb{E}[x^2] \rangle$$

Now we can see, that

$$I(z;x) = -\frac{1}{2}\mathop{\mathbb{E}}_{x}\left[\log \frac{\tilde{\alpha}\langle \hat{w}^2, x^2\rangle}{\langle \hat{w}, \Sigma_x \hat{w}\rangle + \tilde{\alpha}\langle \hat{w}^2, \mathbb{E}\left[x^2\right]\rangle}\right]$$

(making denominator larger by using inequality above and noting that the sign is negative):

$$\begin{aligned}
&\leq -\frac{1}{2}\mathop{\mathbb{E}}_{x}\left[\log \frac{\tilde{\alpha}\langle \hat{w}^2, x^2\rangle}{(1+\tilde{\alpha})\langle \hat{w}^2, \mathbb{E}\left[x^2\right]\rangle}\right] \\
&= \frac{1}{2}\log(1+\frac{1}{\tilde{\alpha}}) - \frac{1}{2}\mathop{\mathbb{E}}_{x}\left[\log \frac{\langle \hat{w}^2, x^2\rangle}{\langle \hat{w}^2, \mathbb{E}\left[x^2\right]\rangle}\right] \\
&= \frac{1}{2}\log(1+\frac{1}{\tilde{\alpha}}) - \frac{1}{2}\mathop{\mathbb{E}}_{x}\left[\log\left(1 + \frac{\langle \hat{w}^2, (x^2 - \mathbb{E}\left[x^2\right])\rangle}{\langle \hat{w}^2, \mathbb{E}\left[x^2\right]\rangle}\right)\right]
\end{aligned}$$

Now we want to prove that the second term (expectation) is very small. For this we'll use Taylor series expansion. But first we must be sure, that the 4-th moment of it exists (we'll do Taylor series expansion up to 2nd term). Since it exists by assumption, we have for some C that:

$$\frac{\mathbb{E}\left[(x_i^2 - \mathbb{E}\left[x_i^2\right])^2\right]}{\mathbb{E}\left[x_i^2\right]^2} \leq C$$

Then:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x}\left[\log\left(1 + \frac{\langle \hat{w}^2, (x^2 - \mathbb{E}\left[x^2\right])\rangle}{\langle \hat{w}^2, \mathbb{E}\left[x^2\right]\rangle}\right)\right] &\approx \mathop{\mathbb{E}}_{x}\left[\frac{\langle \hat{w}^2, (x^2 - \mathbb{E}\left[x^2\right])\rangle}{\langle \hat{w}^2, \mathbb{E}\left[x^2\right]\rangle} - \left(\frac{\langle \hat{w}^2, (x^2 - \mathbb{E}\left[x^2\right])\rangle}{\langle \hat{w}^2, \mathbb{E}\left[x^2\right]\rangle}\right)^2\right] \\
&= -\mathop{\mathbb{E}}_{x}\left[\left(\frac{\langle \hat{w}^2, (x^2 - \mathbb{E}\left[x^2\right])\rangle}{\langle \hat{w}^2, \mathbb{E}\left[x^2\right]\rangle}\right)^2\right] \\
&= \mathrm{Var}[\frac{\hat{w}^2 \cdot \left(x^2 - \mathbb{E}\left[x^2\right]\right)}{\hat{w}^2 \cdot \mathbb{E}\left[x^2\right]}] \qquad (8) \\
&= \frac{\sum_i \hat{w}_i^4 \mathbb{E}\left(x^2 - \mathbb{E}\left[x^2\right]\right)^2}{\sum_{i,j} \hat{w}_i^2 \hat{w}_j^2 \mathbb{E}\left[x_i^2\right]\mathbb{E}\left[x_j^2\right]} \\
&\leq C\frac{\sum_i \hat{w}_i^4 \mathbb{E}\left[x_i^2\right]^2}{\sum_{i,j} \hat{w}_i^2 \hat{w}_j^2 \mathbb{E}\left[x_i^2\right]\mathbb{E}\left[x_j^2\right]} \\
&= O(1/\dim(x))
\end{aligned}$$

We used the fact, that Taylor series of $\log(1+x)$ around $x_0 = 0$ is:

$$\log(1+x) = \log(1+0) + \frac{1}{1+0}\cdot x - x^2/2 + O(x^3) = x - x^2 + O(x^3)$$

Combining this together we get:

$$I(x;z) \leq \frac{1}{2}\log(1+\tilde{\alpha}^{-1}) + O(1/\dim(x))$$

To generalize it on higher dimensions we again just use the fact

$$I(x;z) + TC(z) = \sum_{i=1}^{\dim z} I(z_i;x)$$

□

Now we would like to generalize it on multi-layer case.

Theorem 4. *Let W^k be our weight matrices for $k = 1, ..., L$ with model assumptions and assumptions above. We can use any non-linearity ψ for $z_{k+1} = \psi(W^k z_k)$. Then:*

$$I(z_L;x) \leq \min_{k<L}\left[\dim z_k \cdot (g(\alpha^k) + 1)\right]$$

Here $\alpha^k = e^{-I(W^k;D)/\dim W^k}$.

Proof. We have Markov chain $x \to z_1 \to ... \to z_L$. By DPI we have

$$I(z_L; x) \leq \min_{k<L} I(z_{k+1}, z_k)$$

Besides, applying deterministic function can only decrease the information (since $H(y|x) = H(y|x, f(x)) \leq H(y|f(x))$), so

$$I(z_{k+1}; z_k) \leq I(\psi(W^k z_k); z_k) \leq I(W^k z_k; z_k) \leq g(\alpha^k) + c$$

□

Theorem 5 (Disentanglement for autoencoders [4])**.** *Let use have an autoencoder $h(x) = d \circ e(x)$.*

TBD.

Flat minima

Theorem 6. *Flat minima have low information.*

TBD.

6 Mutual information estimators

TBD.

7 Appendix

Why improper prior

Uninformative prior is a prior which, rouhghly speaking, does not give any preferences in the r.v. values. For finite discrete case uninformative prior is easy to define: it's just $P_i = 1/k$ where k is the number of outcomes. But what to do when our support is countable or continuous?

We can use improper priors as uninformative priors for continuous distributions. For example, log-uniform prior. Consider $p(\log x) = c$, or, which is equivalent:

$$p(x) = p(\log x) \cdot \left| \frac{d \log x}{dx} \right| = c/|x|$$

Authors say, that we can use improper prior for the weights and tell, that it is often the case in practice, referring the reader to their previous paper [1]. But in this paper they only showed that it's fine to use it for activations, not weights!

Kurtosis

What is kurtosis (excess coefficient)?

$$\gamma_2 = \text{Kurt}[X] = \frac{\mathbb{E}\left[(X-\mu)^4\right]}{\sigma^4} - 3$$

It shows, how "picked" our distribution is compared to $\mathcal{N}(0, 1)$, which has zero kurtosis. A similar notion is *skeweness* (assymmetry coefficient), which is defined as

$$\gamma_1 = \frac{\mathbb{E}\left[(X-\mu)^3\right]}{\sigma^3}$$

Symmetric distributions (with finite third moment) have skeweness of 0.

8 Bibliography

References

[1] Alessandro Achille and Stefano Soatto. *Information Dropout: Learning Optimal Representations Through Noisy Computation*. 2016. eprint: arXiv:1611.01353.

[2] 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.

[3] 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.

[4] Ori Press et al. "Emerging Disentanglement in Auto-Encoder Based Unsupervised Image Content Transfer". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=BylE1205Fm.

[5] 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-.

[6] 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.

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

[8] 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.

[9] Chiyuan Zhang et al. "Understanding deep learning requires rethinking generalization". In: 2017. URL: https://arxiv.org/abs/1611.03530.