M2177.0043 Introduction to Deep Learning

Lecture 17: Information theory / Disentanglement

Hyun Oh Song[1]

[1]Dept. of Computer Science and Engineering, Seoul National University

May 19, 2020

Last time

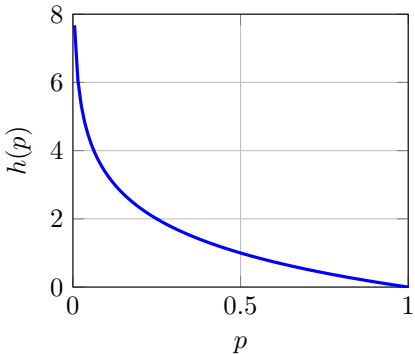
- Defense against adversarial attacks

Outline

Entropy

- Consider a discrete random variable $X \in \{1, \ldots, K\}$
- Suppose we observe event $X = k$. The information content of this event is related to its *surprise* factor

$$h(k) = \log_2 1/p(X = k) = -\log_2 p(X = k)$$



Definition 1 (Entropy)

The *entropy* of distribution p is the average information content

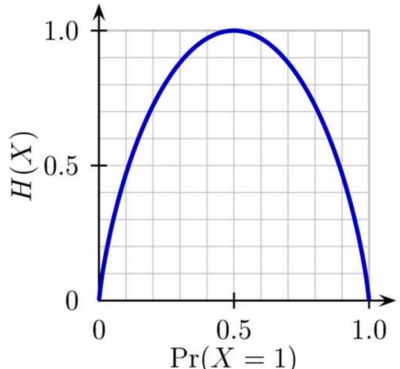
$$H(X) = -\sum_{k=1}^{K} p(X = k) \log_2 p(X = k)$$

- Maximum entropy: Discrete uniform
- Minimum entropy: Delta function

Entropy for Bernoulli R.V.

- Suppose Bernoulli r.v. $X \in \{0, 1\}, p(X = 1) = \theta$
-

$$
\begin{aligned}
H(X) &= -\left(p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)\right) \\
&= -\left(\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)\right)
\end{aligned}
$$



Properties of entropy

Lemma 2

$H(X) \geqslant 0$

Proof.

$0 \leqslant p(x) \leqslant 1$ implies $\log(1/p(x)) \geqslant 0$. Hence

$$H(X) = \mathbb{E}_p[\log(1/p(x))] \geqslant 0$$

■

Lemma 3

$H(X) \leqslant \log K$ holds with equality iff X has a uniform distribution over K elements.

Proof.

Let $u(X) = \frac{1}{K}$ be the uniform probability mass function over K elements, and let $p(X)$ be the probability mass function for X, then

$$D_{KL}(p \parallel u) = \sum_k p(X = k) \log \frac{p(X = k)}{u(X = k)} = \log K - H(X)$$

Hence by the non-negativity of KL divergence,

$$0 \leqslant D_{KL}(p \parallel u) = \log K - H(X)$$

■

Differential entropy

- Extend discrete entropy to the continuous case
- **Caution**: Differential entropy can be negative!

Definition 4

Let $f(X)$ be the probability density function with finite or infinite support $\mathbb{X}$. Differential entropy is defined as

$$h(X) = -\int_{\mathbb{X}} f(x) \ln f(x) dx$$

Joint entropy

Definition 5

The joint entropy of two r.v.s X and Y is merely the entropy of their pairings (X, Y),

$$H(X,Y) = -\mathbb{E}_{X,Y} \log p(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y)$$

Useful properties: (proof left as an exercise)

- $H(X,Y) \geqslant \max[H(X), H(Y)] \geqslant 0$
- $H(X,Y) \leqslant H(X) + H(Y)$ (holds with equality iff $X \perp\!\!\!\perp Y$)

Conditional entropy

Definition 6

The conditional entropy of X given random variable Y is the average conditional entropy over Y,

$$H(X \mid Y) = \mathbb{E}_Y H(X \mid y) = -\sum_y p(y) \sum_x p(x \mid y) \log p(x \mid y)$$
$$= -\sum_{x,y} p(x,y) \log p(x \mid y)$$

Useful properties: (proof left as an exercise)

- $H(Y \mid X) = H(X,Y) - H(X)$
- $H(Y \mid X) = 0$ iff Y is completely determined by X
- $H(Y \mid X) = H(Y)$ iff $Y \perp\!\!\!\perp X$
- $H(Y \mid X) = H(X \mid Y) - H(X) + H(Y)$

Mutual information

Definition 7

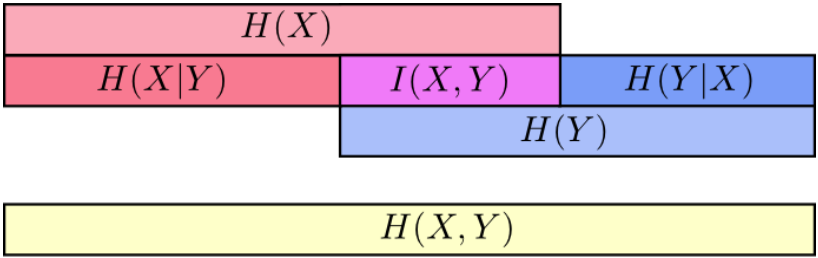
Mutual information measures the amount of information that can be obtained about one r.v. by observing another.

$$\begin{aligned} I(X;Y) &= D_{KL}(p(X,Y) \parallel p(X)p(Y)) \\ &= \mathbb{E}_{X,Y} \log \frac{p(X,Y)}{p(X)p(Y)} \end{aligned}$$

Useful properties: (proof left as an exercise)

- $I(X;Y) = I(Y;X)$ (symmetric)
- $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
- $H(X,Y) = H(X \mid Y) + H(Y \mid X) + I(X;Y)$
- $I(X;Y) = 0 \iff X \perp\!\!\!\perp Y$
- $0 \leqslant I(X;Y) \leqslant H(X) \leqslant \log_2 K$

Information diagram[1]



[1] https://colah.github.io/posts/2015-09-Visual-Information/

Total correlation

Definition 8

For a given set of n random variables $X = \{X_1, \ldots, X_n\}$, the total correlation $TC(X)$ is defined as,

$$TC(X) = D_{KL}(p(X_1, \ldots, X_n) \parallel p(X_1) \cdots p(X_n))$$

Also, this reduces to,

$$TC(X) = \sum_{i=1}^{n} H(X_i) - H(X_1, \ldots, X_n)$$

- One of the generalizations of the mutual information. Quantifies the redundancy or dependency among a set of n random variables

- $\sum_{i=1}^{n} H(X_i)$ represents the amount of information that the variables would possess if they were totally independent of one another
- $H(X_1, \ldots, X_n)$ is the actual amount of information that the variable set contains
- The difference between the two terms represent the redundancy present in the given set of variables

Information gain and VAE regularizer

Relationship between infogain and VAE regularizer

$$
\begin{aligned}
I(x;z) &= \int q(x,z) \log \frac{q(x,z)}{p(x)q(z)} dxdz \\
&= \int q(x,z) \log \frac{q(x,z)}{p(x)p(z)} dxdz + \int q(x,z) \log \frac{p(z)}{q(z)} dxdz \\
&= \int p(x)q(z \mid x) \log \frac{q(z \mid x)}{p(z)} dxdz + \int q(z) \log \frac{p(z)}{q(z)} dz \\
&= \underbrace{\mathbb{E}_{x \sim p(x)} D_{KL}(q(z \mid x) \parallel p(z))}_{\text{VAE regularizer}} - \underbrace{D_{KL}(q(z) \parallel p(z))}_{:=(*)} \qquad (1)
\end{aligned}
$$

Examine $D_{KL}(q(z) \parallel p(z))$ term

Examine $(*)$

$$
\begin{aligned}
D_{KL}(q(z) \parallel p(z)) &= \int q(z) \log \frac{q(z)}{p(z_1) \cdots p(z_d)} dz \\
&= \int q(z) \log \frac{q(z_1) \cdots q(z_d)}{p(z_1) \cdots p(z_d)} dz + \int q(z) \log \frac{q(z)}{q(z_1) \cdots q(z_d)} dz \\
&= \sum_{i=1}^{d} \int q(z_i) \log \frac{q(z_i)}{p(z_i)} dz_i + TC(z) \\
&= \sum_{i=1}^{d} D_{KL}(q(z_i) \parallel p(z_i)) + TC(z) \qquad (2)
\end{aligned}
$$

Putting Equation (1) and Equation (2) together

$$\underbrace{\mathbb{E}_{x\sim p(x)} D_{KL}(q(z \mid x) \parallel p(z))}_{\text{VAE regularizer}} = I(x;z) \; + \; TC(z) \; + \; \sum_i D_{KL}(q(z_i) \parallel p(z_i))$$

Interpretation:

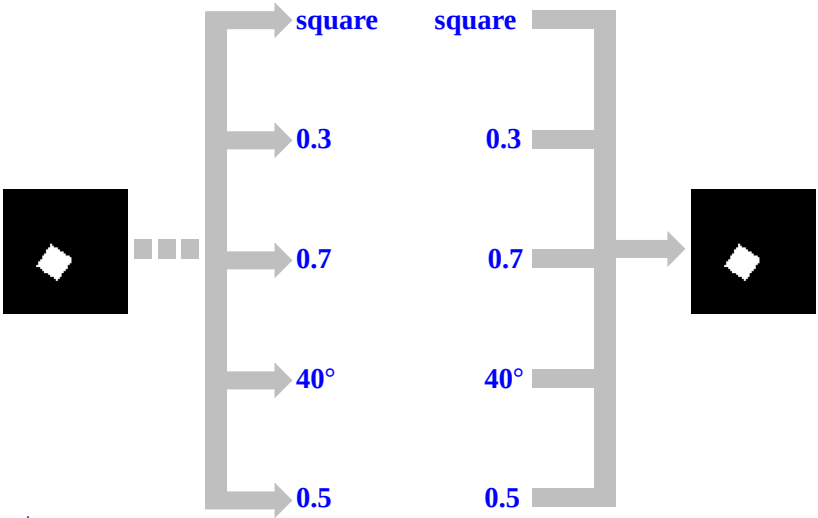
- Minimizing the VAE regularizer decreases $TC(z)$, the redundancy of the learned representation z 
- However, the information gain between the data and the representation is decreased at the same time 
- Ideally, we want to increase $I(x;z)$ and decrease $TC(z)$
- Optimizing $TC(z)$ directly is hard 

Outline

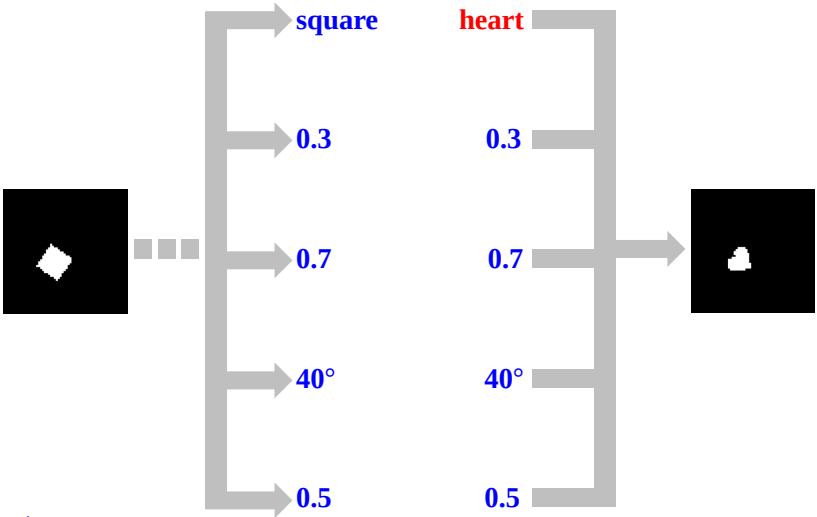
Unsupervised disentangled representation learning

- Learning to disentangle the underlying explanatory factors of data without supervision is a cruicial task for representation learning
- In a successfully disentangled representation, a single latent unit of representation should correspond to a change in a single generative factor of the data while being relatively invariant to others
- Penalizing total correlation encourages the model to learn statistically independent factors of data

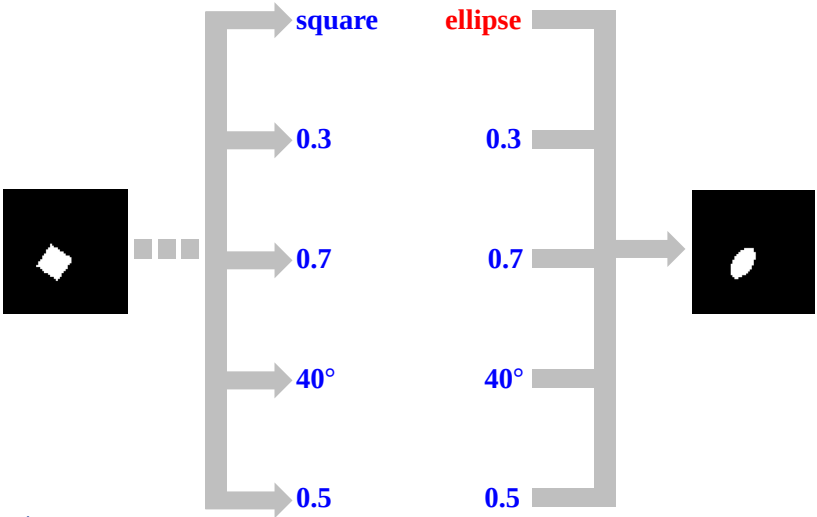
Disentanglement



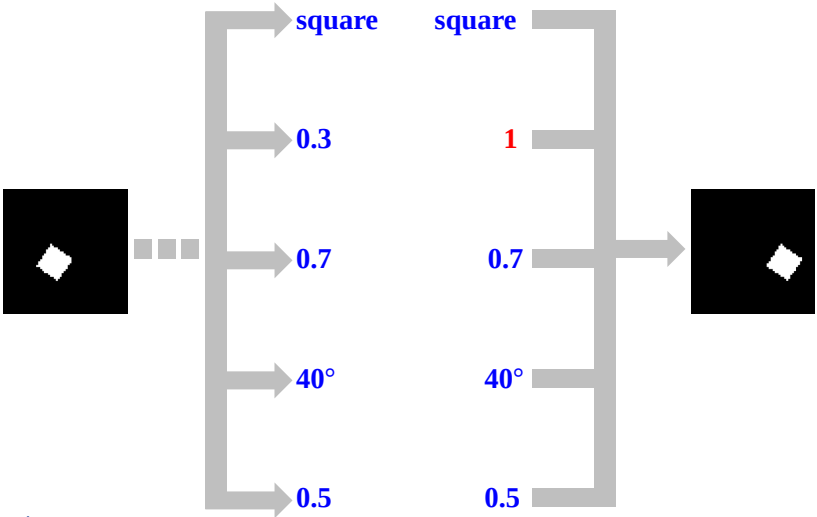
Disentanglement



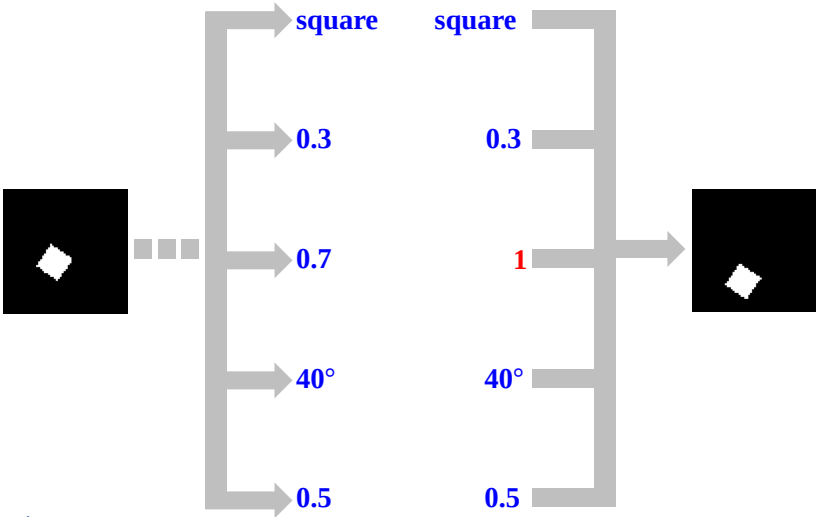
Disentanglement



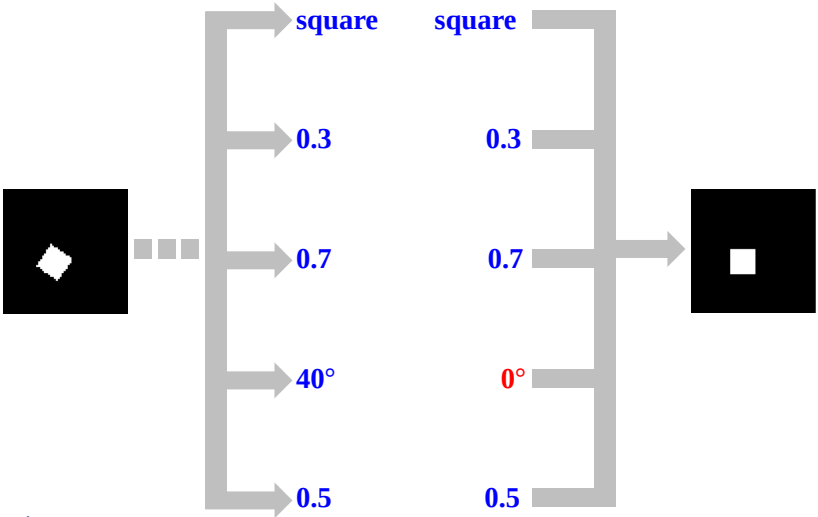
Disentanglement



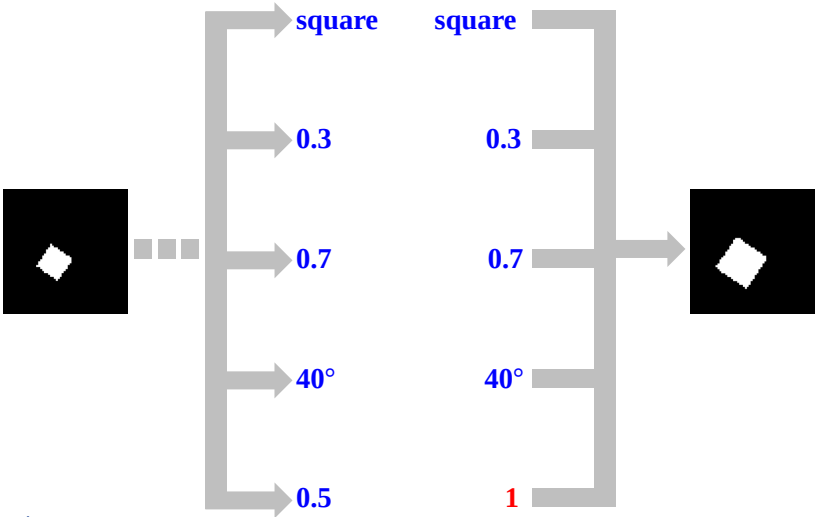
Disentanglement



Disentanglement



Disentanglement



β-VAE[2]

- If $\beta = 1$, objective same as VAE
- β-VAE sets $\beta > 1$ indirectly penalizing $TC(z)$ by increasing β coefficient to a high value (*i.e.* 10.0)

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{z \sim q_\phi(\cdot|x)} \log p_\theta(x \mid z) - \beta D_{KL}(q_\phi(z \mid x) \parallel p(z)) \right]$$

[2] Higgins, et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework" ICLR2017

β-VAE vs VAE

Figure: Traversal z_1, (Left) β-VAE, (Right) VAE

β-VAE vs VAE

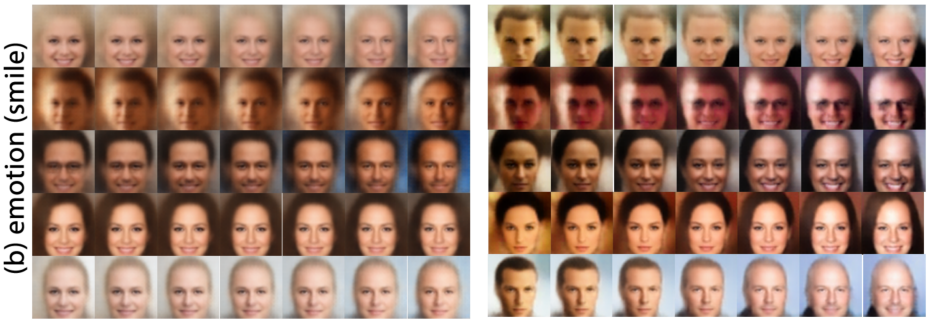


Figure: Traversal z_2, (Left) β-VAE, (Right) VAE

β-VAE vs VAE

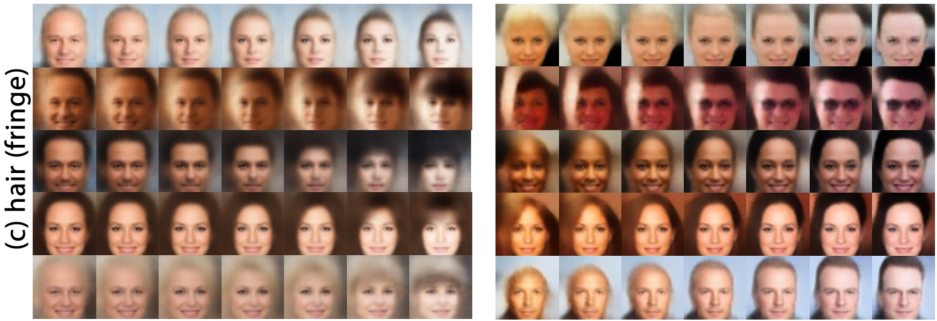


Figure: Traversal z_3, (Left) β-VAE, (Right) VAE

Next

- How we *maximize the mutual information* between the data and the latent representation $I(x; z)$ while *minimizing the total correlation* on the representation $TC(z)$? (FactorVAE, CascadeVAE, etc)
- How do we optimize for a latent representation with both discrete and continuous latent factors?