TBSI 清华-伯克利深圳学院
Tsinghua-Berkeley Shenzhen Institute

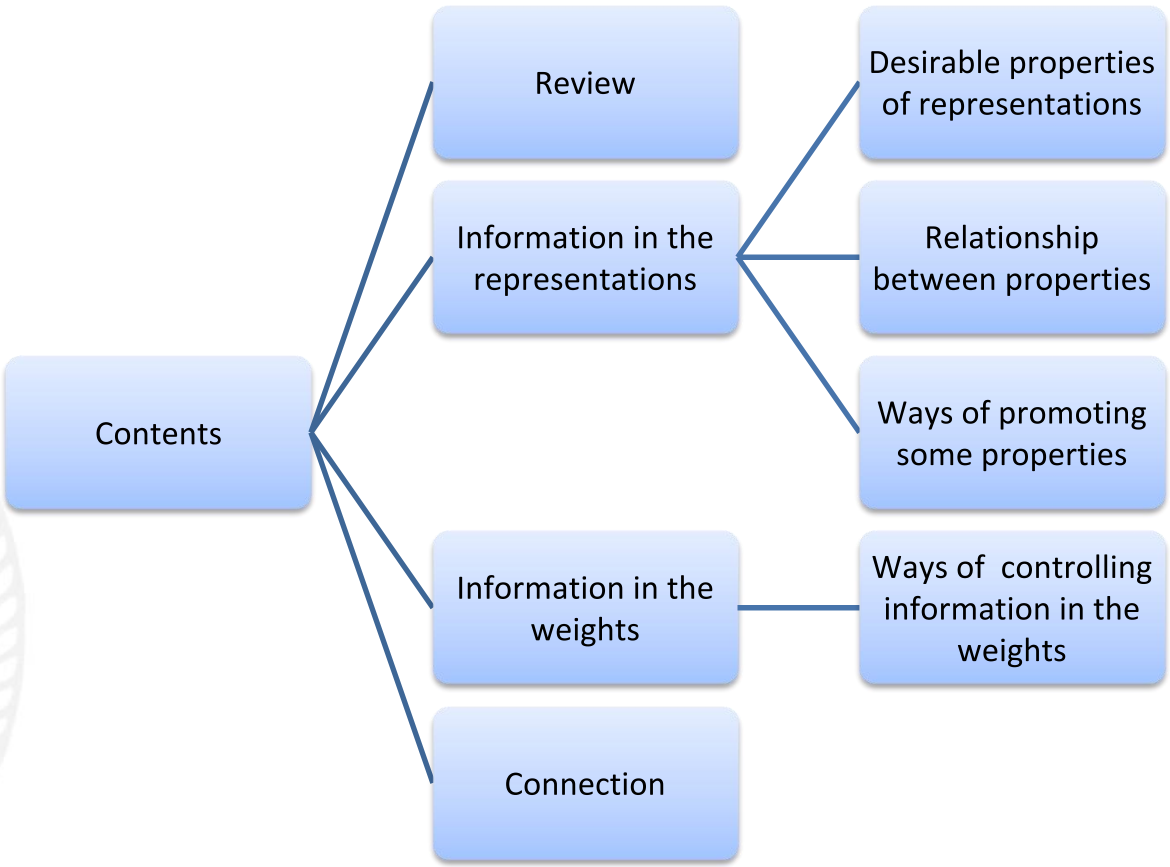
Information in the representations and information in the weights of deep learning

Ziyan Zheng
June 5, 2020

Abstract

Effectiveness of deep learning is often ascribed to the ability of deep networks to learn representations. Using established principles from Statistics and Information Theory, we introduce the desirable properties of representations and the relationships between them. It is shown that invariance to nuisance factors in a deep neural network is equivalent to information minimality of the learned representation. On the other hand, we want to find parameters (weights) that yield good generalization, so we focus on the information in the weights which can be controlled by implicit or explicit regularization. With some additional assumptions, we get a connection between information in the representations and information in the weights.

Contents



Contents
Review
Information in the representations
Desirable properties of representations
Relationship between properties
Ways of promoting some properties
Information in the weights
Ways of controlling information in the weights
Connection

Reference

1. Alessandro Achille, Stefano Soatto. Emergence of Invariance and Disentanglement in Deep Representations. ICML 2017.
arXiv: 1706.01350
2. Yoshua Bengio. From Deep Learning of Disentangled Representations to Higher-level Cognition.
http://t.cn.AiejL66r
3. Scott Fortmann-Roe. Understanding the Bias-Variance Tradeoff.
http://scott.fortmann-roe.com/docs/BiasVariance.html
4. Ravid Schwartz-Ziv, Naftali Tishby. Opening the black box of Deep Neural Networks via Information.
arXiv:1703.00810v3
5. Cover and Thomas. Elements of Information Theory. 2006
6. Yoshua Bengio, Aaron Courville and Pascal Vincent. Representation Learning: A Review and New Perspectives.
7. David McAllester. A pac-bayesian tutorial with a dropout bound.
arXiv:1307.2118, 2013.

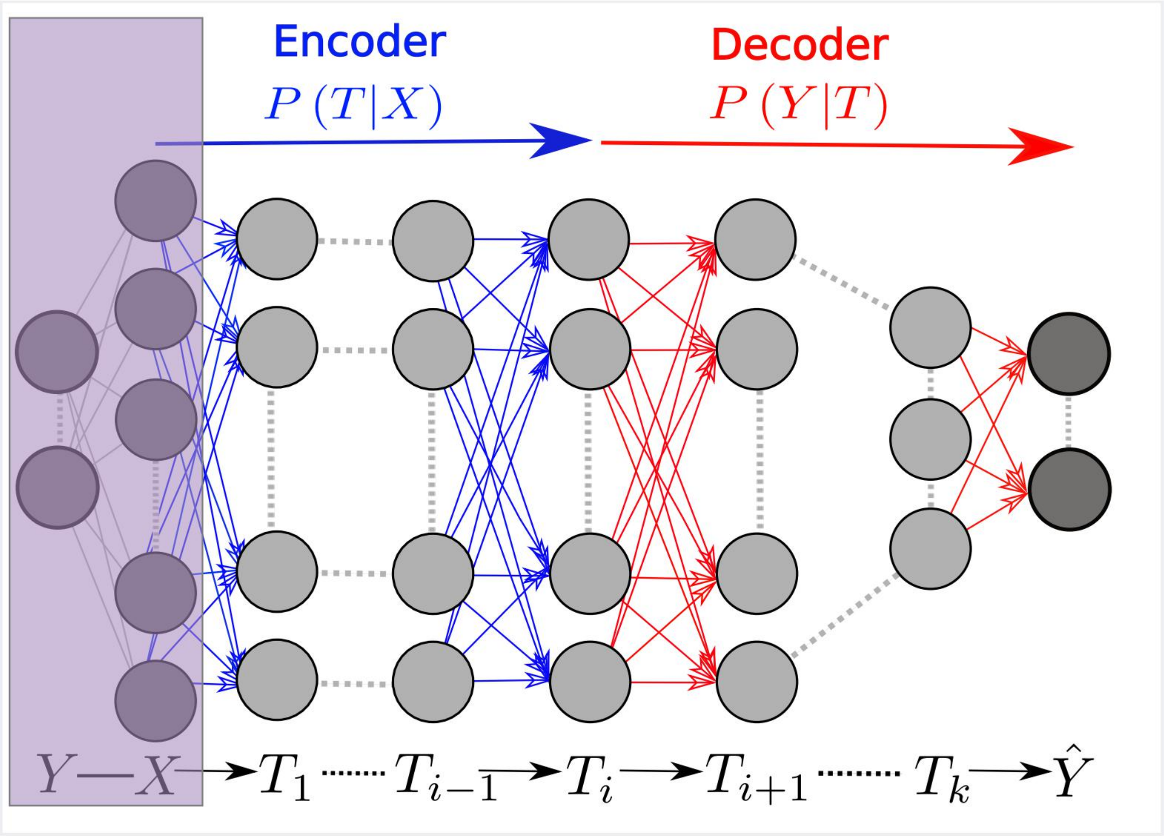
Review

Unknown (possibly complex) distribution $p(x, y)$
Training set $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x} = \left\{x^{(i)}\right\}_{i=1}^{N}$ and $\mathbf{y} = \left\{y^{(i)}\right\}_{i=1}^{N}$
Task: Given test sample r.v. x, infer r.v. y

Frequently used quantities:
Shannon entropy: $H(x) = \mathbb{E}_p[-\log p(x)]$
Conditional entropy: $H(x|y) := \mathbb{E}_{\bar{y}}[H(x|y = \bar{y})] = H(x, y) - H(y)$
Mutual information: $I(x; y) = H(x) - H(x|y)$
Conditional mutual information: $I(x; y|z) = H(x|z) - H(x|y, z)$
KL divergence: $\mathrm{KL}(p(x)\|q(x)) = \mathbb{E}_p[\log \frac{p(x)}{q(x)}]$
Cross-entropy: $H_{p,q}(x) = \mathbb{E}_p[-\log q(x)]$
Total correlation: $\mathrm{TC}(z) = \mathrm{KL}\left(p(z)\| \prod_i p\left(z_i\right)\right)$
Often-used identity:

$$I(z; x) = \mathbb{E}_{x\sim p(x)}\mathrm{KL}(p(z|x)\|p(z))$$

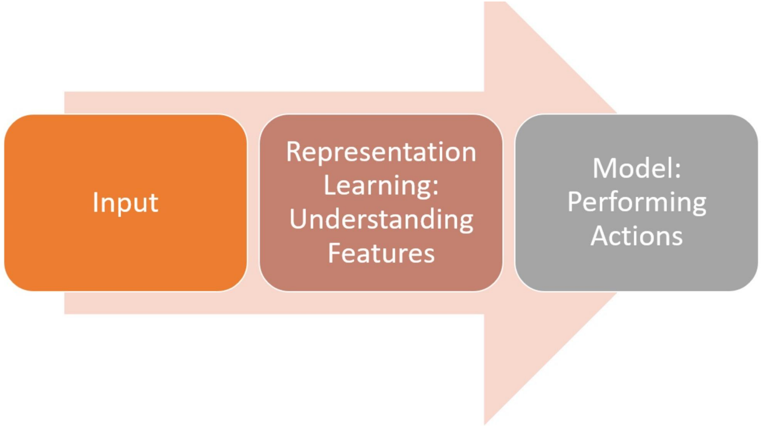
Information in the representation



Representation learning

Information in the representation

- **representation** z of x: described as $p(z|x)$



Markov chain: $y \longrightarrow x \longrightarrow z$

Data Processing Inequality (DPI): $I(z;y) \leqslant I(x;y)$

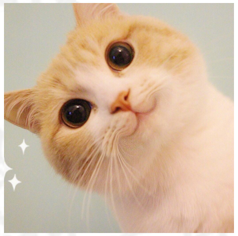
Proof.
$$\begin{aligned} I(z;y) &= I(x,z;y) - I(x;y|z) \\ &= I(x;y) + I(z;y|x) - I(x;y|z) \\ &= I(x;y) - I(x;y|z) \\ &\leqslant I(x;y) \end{aligned}$$

Information in the representation

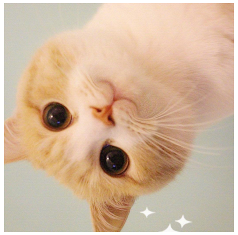
Markov chain: $(y, n) \to x \to z$

· **nuisance** n for task y: $y \perp n$, equivalently $I(y;n) = 0$

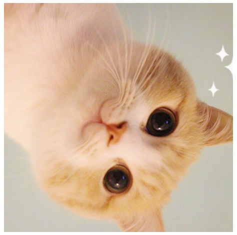
Example: Is rotation angle θ a nuisance?



cat



cat



cat

cat

nuisance ✓

six

nine

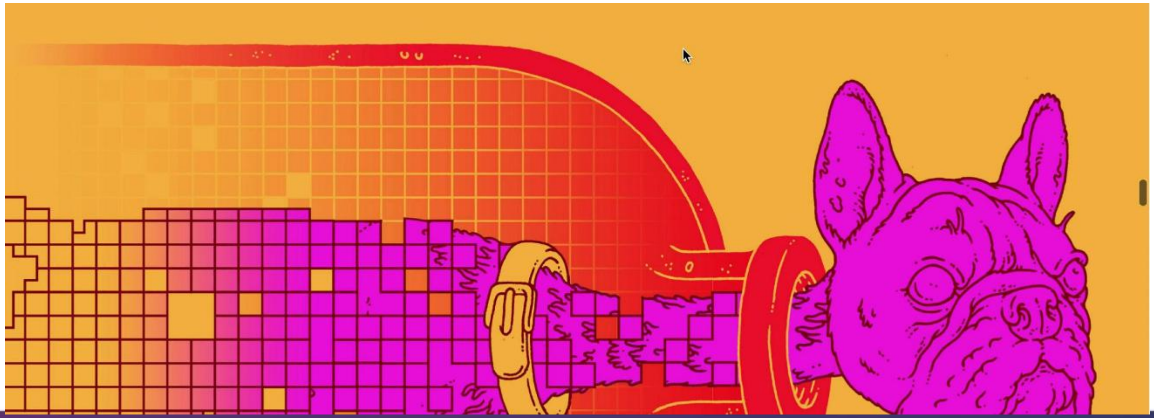
θ may not be a nuisance

Information in the representation

Markov chain: $(y, n) \rightarrow x \rightarrow z$

Intuitively, what are desirable properties of representation?

(a) **sufficient** for the task y, i.e. $I(y;z) = I(y;x)$
(b) **minimal**, i.e. $I(z;x)$ is minimized
(c) **invariant** to the effect of nuisances $I(z;n) = 0$



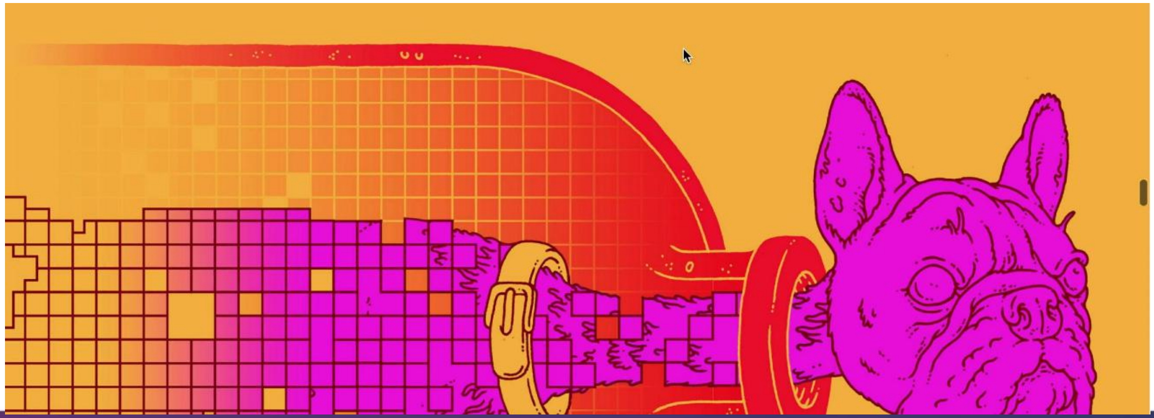
Information in the representation

Markov chain: $(y, n) \to x \to z$

Intuitively, what are desirable properties of representation?

(a) **sufficient** for the task y, i.e. $I(y;z) = I(y;x)$
(b) **minimal**, i.e. $I(z;x)$ is minimized
(c) **invariant** to the effect of nuisances $I(z;n) = 0$

enforce (a)(b) $\Rightarrow$ promote (c)



Information in the representation

Markov chain: $(y, n) \to x \to z$

Intuitively, what are desirable properties of representation?

(a) **sufficient** for the task y, i.e. $I(y;z) = I(y;x)$
(b) **minimal**, i.e. $I(z;x)$ is minimized
(c) **invariant** to the effect of nuisances $I(z;n) = 0$

enforce (a)(b) $\Rightarrow$ promote (c)

Proposition 1 (Invariance and minimality)
n: nuisance
z: a **sufficient** representation of the input x

$$I(z;n) \leqslant I(z;x) - I(x;y)$$

Moreover, there is a nuisance n s.t. equality holds up to a (generally small) residual ϵ

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

where $\epsilon := I(z;y|n) - I(x;y)$. In particular $0 \leqslant \epsilon \leqslant H(y|x)$, and $\epsilon = 0$ whenever y is a deterministic function of x.

Information in the representation

Markov chain: $(y, n) \to x \to z$

Intuitively, what are desirable properties of representation?

(a) **sufficient** for the task y, i.e. $I(y; z) = I(y; x)$
(b) **minimal**, i.e. $I(z; x)$ is minimized
(c) **invariant** to the effect of nuisances $I(z; n) = 0$

enforce (a)(b) $\Rightarrow$ promote (c)

Compare to minimal sufficient statistics:

$S(x)$ is **sufficient statistics** for x: $y \perp x | S(x)$

s.s. $T(x)$ is **minimal sufficient statistics**: $y \to x \to S(x) \to T(x)$ hold for any s.s. $S(x)$

$$T(x) = \arg \min_{S(x): I(S(x);y) = I(x;y)} I(S(x); x)$$

Information in the representation

Markov chain: $(y, n) \to x \to z$

Intuitively, what are desirable properties of representation?

(a) **sufficient** for the task y, i.e. $I(y; z) = I(y; x)$
(b) **minimal**, i.e. $I(z; x)$ is minimized
(c) **invariant** to the effect of nuisances $I(z; n) = 0$

enforce (a)(b) $\Rightarrow$ promote (c)

Ways of promoting invariance

1. Modify cost function

Information Bottleneck Lagrangian (Tishby et al. 1999)

$$\mathcal{L}(p(z|x)) = H(y|z) + \beta I(z; x)$$

where β trades off sufficiency and minimality.

2. Stacking

$$x \to z_1 \to z_2 \to \ldots \to z_L$$

If z_L is still sufficient for y, then it is more invariant than preceding layers.

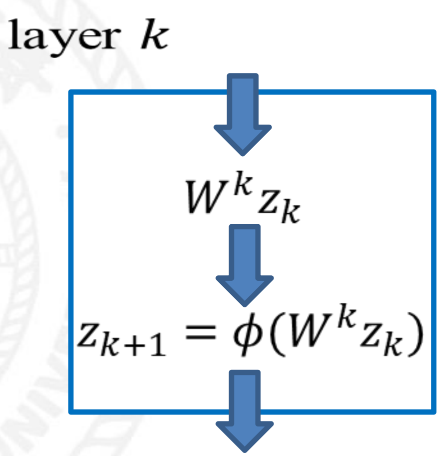
Information in the weights

Unknown (possibly complex) distribution $p(x, y)$

Training set $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x} = \{x^{(i)}\}_{i=1}^{N}$ and $\mathbf{y} = \{y^{(i)}\}_{i=1}^{N}$

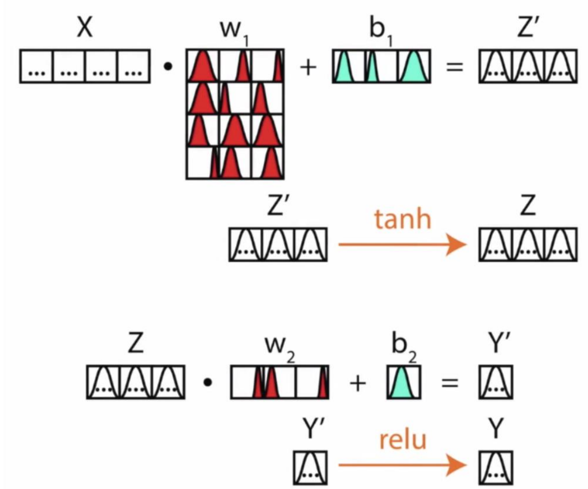
Task: Given test sample r.v. x, infer r.v. y

Information in the weights: $I(w; \mathcal{D}) = \mathbb{E}_{\mathcal{D}} \mathrm{KL}(q(w|\mathcal{D}) \| q(w))$



$$W^k = \varepsilon \odot \hat{W}^k$$

Layer k of deep learning



Bayesian deep learning model

Information in the weights

Information in the weights: $I(w;\mathcal{D}) = \mathbb{E}_{\mathcal{D}}\mathrm{KL}(q(w|\mathcal{D})\|q(w))$

It has been proved that we can trade off cross-entropy loss and information in the weights to obtain better generalization. (McAllester 2013)

$$\mathcal{L}(q(w|\mathcal{D})) = H_{p,q}(\mathbf{y}|\mathbf{x},w) + \beta I(w;\mathcal{D})$$

Information in the weights

Information in the weights: $I(w;\mathcal{D}) = \mathbb{E}_{\mathcal{D}}\mathrm{KL}(q(w|\mathcal{D})\|q(w))$

It has been proved that we can trade off cross-entropy loss and information in the weights to obtain better generalization. (McAllester 2013)

$$\mathcal{L}(q(w|\mathcal{D})) = H_{p,q}(\mathbf{y}|\mathbf{x}, w) + \beta I(w;\mathcal{D})$$

Same general form

$$\mathcal{L}(p(z|x)) = H(y|z) + \beta I(z;x)$$

Information in the weights

Information in the weights: $I(w;\mathcal{D}) = \mathbb{E}_{\mathcal{D}}\mathrm{KL}(q(w|\mathcal{D})\|q(w))$

Ways of controlling information in the weights

1. Modify cost function

$$\mathcal{L}(q(w|\mathcal{D})) = H_{p,q}(\mathbf{y}|\mathbf{x}, w) + \beta I(w;\mathcal{D})$$

2. SGD

Controll the learning rate and the size of mini-batches (Chaudhari and Soatto 2018).

3. Bias the optimization towards "flat minima"

Flat minima have low information. Some variants of SGD bias the optimization towards "flat minima" (Chaudhari et al., 2017).

Proposition 2 (Flat minima have low information)
Let $\hat{w}$ be a local minimum of the cross-entropy loss $H_{p,q}(\mathbf{y}|\mathbf{x}, w)$, and let $\mathcal{H}$ be the Hessian at that point. Then, for the optimal choice of the posterior $w|\mathcal{D} = \epsilon \odot \hat{w}$ centered at $\hat{w}$ that optimizes the IB Lagrangian, we have

$$I(w;\mathcal{D}) \leqslant \frac{1}{2}K\left[\log\|\hat{w}\|_2^2 + \log\|\mathcal{H}\|_* - K\log\left(K^2\beta/2\right)\right]$$

where $K = \dim(w)$ and $\|\cdot\|_*$ denotes the nuclear norm.

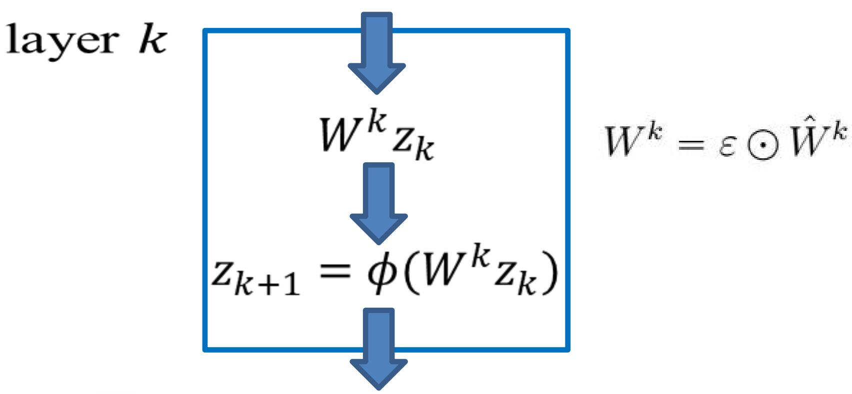
Connection

With some additional assumptions, we get a connection between information in the representations and information in the weights.

Proposition 3 (Information in the weights bounds the information in the representations) Let W^k for $k = 1, \ldots, L$ be weight matrices, with $W^k = \epsilon^k \odot \hat{W}^k$ and $\epsilon^k_{i,j} = \log \mathcal{N}\left(-\alpha^k/2, \alpha^k\right)$, and let $z_{i+1} = \phi\left(W^k z_k\right)$,where $z_0 = x$ and ϕ is any nonlinearity. Then

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

where $\hat{W}^k$ is learned mean, $\alpha^k = \exp\left\{-I\left(W^k; \mathcal{D}\right) / \dim\left(W^k\right)\right\}$, $g(\alpha^k) = -\log\left(1 - e^{-\alpha^k}\right)/2$



Conclusion

1. Information in the representations

· Sufficiency, minimality and invariance are desirable properties of representation.

· Sufficiency and minimality of a representation enforce invariance.

· Invariance could get promoted by modifying cost function or stacking.

2. Information in the weights

· Controlling information in the weights makes better generalization.

· Information in the weights could be controlled by modifying cost function, SGD or bias towards "flat minima" in optimization.

3. Connection

· With some assumptions, information in the weights bounds the information in the representations.

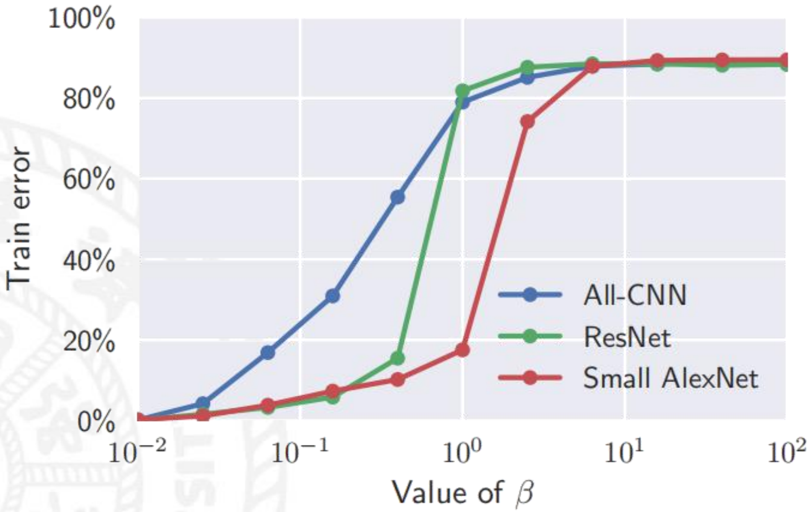
Thanks

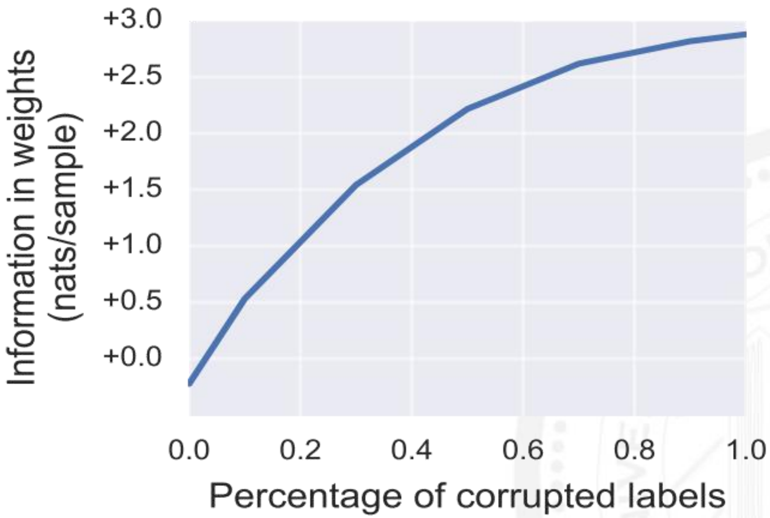
TBSI 清华-伯克利深圳学院 Tsinghua-Berkeley Shenzhen Institute

Empirical validation

Transition from overfitting to underfitting

CIFAR-10 random labels



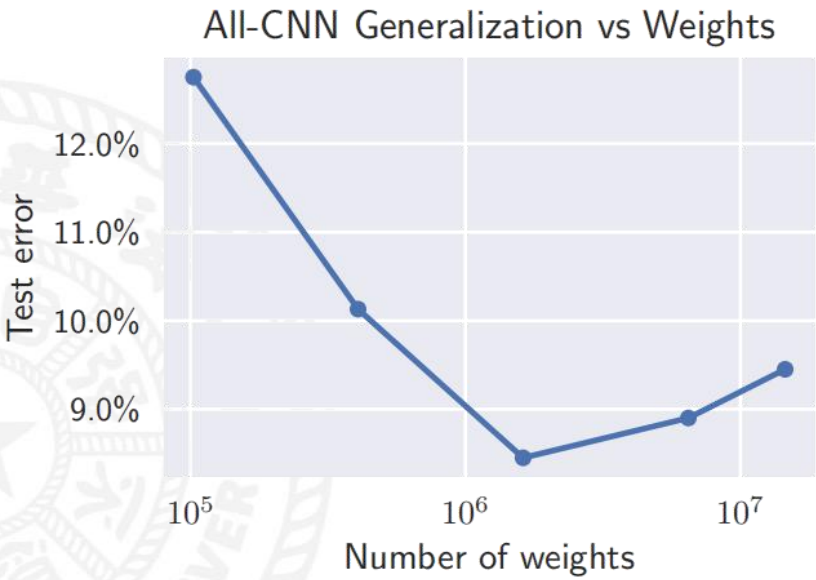


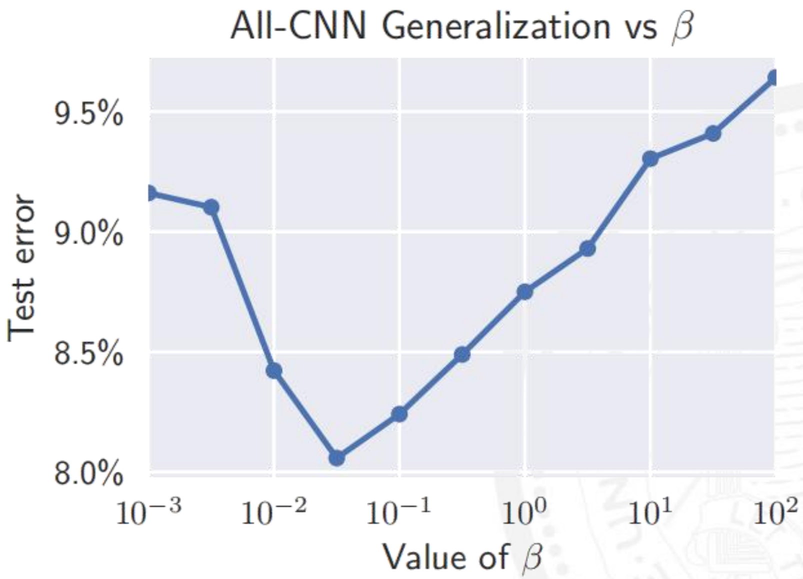
$$\mathcal{L}(q(w|\mathcal{D})) = H_{p,q}(\mathbf{y}|\mathbf{x}, w) + \beta I(w; \mathcal{D})$$

Empirical validation

Bias-variance trade-off

CIFAR-10





$$\mathcal{L}(q(w|\mathcal{D})) = H_{p,q}(\mathbf{y}|\mathbf{x}, w) + \beta I(w; \mathcal{D})$$

Empirical validation

Nuisance invariance

MNIST