On Variational Characterization of Mutual Information for Regularizing Deep Learning

Shell Xu Hu

December 9, 2018

École des Ponts ParisTech

École des Ponts
ParisTech

My research topics

My website: `http://hushell.github.io/`

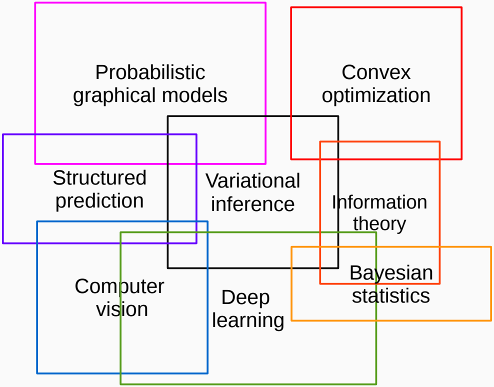


Table of contents

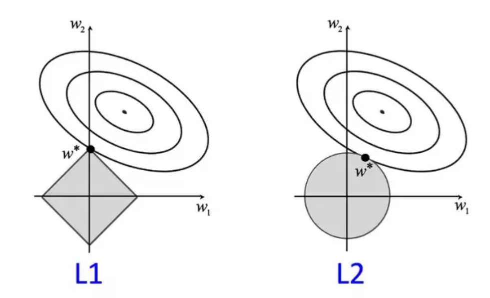
Motivation

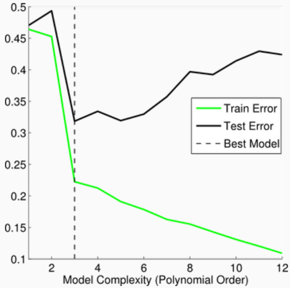
Regularization is a standard way to control model complexity

A rule of thumb of machine learning:

$$\min_{w} \ \text{loss}(w, \text{trainset}) + \beta \ \text{regularization}(w)$$

Test error $\approx$ estimator **variance** + squared estimator **bias** + noise.



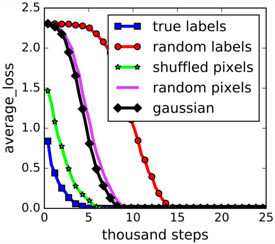


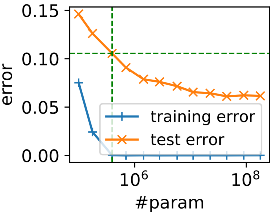
Deep learning has implicit regularization

Observations by Zhang et al. (2016); Neyshabur et al. (2018):

- **High capacity**: train error is near zero even with random labels.
- **Over-parameterization**: increasing the number of parameters does not overfit.

Complex deep models do not have high variance?

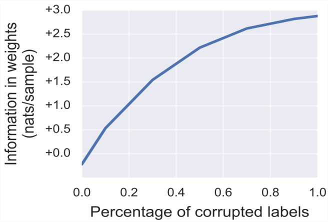




Need to understand the regularization in deep learning

Perhaps we should not link generalization with model complexity.

- Hypotheses: implicit regularization comes from either the network architecture or the stochastic gradient descent (SGD).
- Achille and Soatto (2017) look at the amount of information in the weights instead, which is inspired by the *information bottleneck* interpretation of SGD (Tishby and Zaslavsky, 2015).



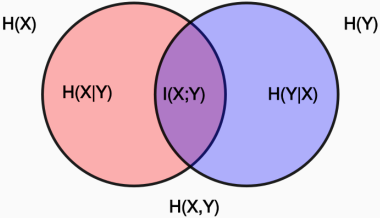
Mutual Information

Mutual information: a math concept from Shannon

Mutual information measures statistical dependency

$$
\begin{aligned}
I(X;Y) &:= \mathbb{E}_{x,y\sim p(x,y)} \log \frac{p(x,y)}{p(x)p(y)} \\
&= H(X,Y) - H(X|Y) - H(Y|X) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

$$
H(X) = I(X;X) = \text{ expected amount of information in } X
$$



Mutual information is a functional of distributions

If we decompose the joint distribution as $p(x, y) = p(x)q(y|x)$, then the mutual information can be writen as a functional of p and q:

$$I(X; Y) \equiv I(p, q) := \mathbb{E}_{x,y\sim p(x,y)} \log \frac{q(y|x)}{q(y)},$$

$$q(y) := \sum_{x} p(x)q(y|x).$$

Issue: it is computationally difficult since $q(y|x)$ and $q(y)$ are coupled.

Variational characterization of mutual information

Lemma (Cover and Thomas, 2012, Theorem 10.8.1)

$$I(X;Y) = \max_{\phi(x|y)\in\Delta} \underbrace{\mathbb{E}_{x,y\sim p(x,y)} \log \frac{\phi(x|y)}{p(x)}}_{\tilde{I}(p,q,\phi)}$$

$$I(X;Y) = \min_{m(y)\in\Delta} \underbrace{\mathbb{E}_{x,y\sim p(x,y)} \log \frac{q(y|x)}{m(y)}}_{\hat{I}(p,q,m)}.$$

An Application to Bayesian Neural Networks

Bayesian inference

A brief introduction:

- Bayesians describe data Y through the latent variable model

$$p(Y, w) = p(Y|w)p(w) = p(w)\prod_i p(y_i|w),$$

  assuming the *likelihood* $p(Y|w)$ and the *prior* $p(w)$ are given.

- Bayesians make predictions according to

$$p(y_{\text{new}}|Y) = \int p(y_{\text{new}}|w)p(w|Y)dw,$$

  where $p(w|Y)$ is the *posterior*.

Bayesian neural networks

Vanilla Bayesian neural networks (BNNs) by Hinton and Van Camp (1993); Graves (2011); Blundell et al. (2015):

- Assume w is Gaussian distributed with a prior $p(w) = \mathcal{N}(0, I)$.
- Given data S, approximate the posterior $p(w|S)$ by $q(w|\theta^*)$:

$$\begin{aligned}
\theta^* &= \arg\min_\theta D_{\mathrm{KL}}\big(q(w|\theta) \| p(w|S)\big) \\
&= \arg\min_\theta \int q(w|\theta) \log \frac{q(w|\theta)}{p(w)p(S|w)} dw \\
&= \arg\min_\theta -\mathbb{E}_{q(w|\theta)}[\log p(S|w)] + D_{\mathrm{KL}}(q(w|\theta) \| p(w)).
\end{aligned}$$

Rate-distortion tradeoff: a lossy data compression framework

To induce a lossy compression of $X \to \hat{X}$, when $p(x)$ is given:

$$\min_{q(\hat{x}|x)\in\Delta} I(p,q)$$

$$\text{s.t.} \underbrace{\sum_{x,\hat{x}} p(x)q(\hat{x}|x)\ d(x,\hat{x})}_{D(p,q)} \leq \text{const.}$$

An equivalent problem by variational characterization:

$$\min_{q(\hat{x}|x)\in\Delta} \min_{m(\hat{x})\in\Delta} \hat{I}(p,q,m) + \beta\ D(p,q).$$

An algorithm for rate-distortion tradeoff

An equivalent problem by variational characterization:

$$\min_{q(\hat{x}|x)\in\Delta} \min_{m(\hat{x})\in\Delta} \hat{I}(p,q,m) + \beta\ D(p,q).$$

Alternating projection algorithm (aka Blahut-Arimoto algorithm)

Provided an initial $q_t(\hat{x}|x)$ at $t=0$. At iteration $t>0$, taking the following steps:

$$q_t(\hat{x}|x) = \frac{m_t(\hat{x})e^{-\beta d(x,\hat{x})}}{\sum_{\hat{x}'} m_t(\hat{x}')e^{-\beta d(x,\hat{x})}},$$

$$m_{t+1}(\hat{x}) = \sum_x p(x)q_t(\hat{x}|x).$$

Then, the algorithm converges to a global minimum.

Rate-distortion perspective on supervised learning

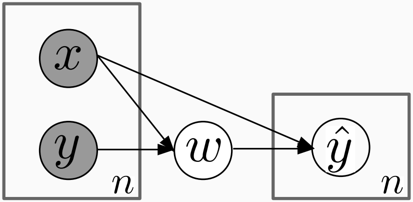
Supervised learning (with model uncertainty) can be viewed as creating a lossy compression w for the data S:

- We describe S by a latent variable model

$$p(S, w) = q(w|S)p^*(S).$$

- We make predictions according to

$$q(y \mid x, S) := \int p(y \mid x, w)q(w|S)dw.$$



Rate-distortion inspired objective for supervised learning

The compression-accuracy tradeoff:

$$\min_{q(w|S)\in\Delta} \Big[I(w;S) \equiv I(q(w|S)) \Big] \text{ s.t. } \mathbb{E}_{p^*(S)}\mathbb{E}_{q(w|S)} d(w,S) \le D$$

$$I(q(w|S)) := \mathbb{E}_{p^*(S)}\mathbb{E}_{q(w|S)}\big[\log \frac{q(w|S)}{q(w)}\big], \quad d(w,S) := -\sum_{i=1}^{n} \log p(y_i|x_i,w),$$

Applying variational characterization, we obtain

$$I(w;S) \equiv \min_{m(w)\in\Delta} I(q,m) := \mathbb{E}_{p^*(S)}\mathbb{E}_{q(w|S)}\big[\log \frac{q(w|S)}{m(w)}\big].$$

Intuition: $I(w;S)$ is a regularizer, which forces w to contain less information about a particular S. In other words, **reducing the variance**.

Approximate Blahut-Arimoto algorithm

1. We use a variational approximation $q(w|\theta)$ for $q(w|S)$ by solving

$$\begin{aligned}\theta(S) &= \arg\min_{\theta} D_{\mathrm{KL}}(q(w|\theta)\|q(w|S)) \\ &= \arg\min_{\theta} D_{\mathrm{KL}}(q(w|\theta)\|m(w)) + \beta\,\mathbb{E}_{q(w|\theta)}\big[d(w,S)\big].\end{aligned}$$

2. $m(w) \simeq \sum_S p^*(S) q(w|\theta(S)) \simeq \frac{1}{K}\sum_{k=1}^{K} q(w|\theta(B_k)) =: \tilde{m}(w)$, where B_k is a bootstrap sample of size n_b drawn from the empirical distribution $p_S(x,y) = \frac{1}{n}\sum_{i=1}^{n} \delta(x_i = x)\delta(y_i = y)$.

β-BNN

Input: S (dataset), β (coefficient), K ($\#$ mixture components), n_b (size of a bootstrap sample).
Initialize: $\Theta = \{\theta_k^{(0)} = (0, I)\}_{k=1}^{K}$; $\tilde{m}(w) = \frac{1}{K}\sum_{\theta \in \Theta} q(w|\theta)$.
for all $t = 1, \ldots, T$ **do**
  Draw K bootstrap samples $\{B_k\}_{k=1}^{K}$ of size n_b from $p_S(x, y)$.
  for all $k = 1, \ldots, K$ **do**
    $\theta_k^{(t)} \leftarrow \theta(B_k)$.
    $\Theta = \Theta \cup \{\theta_k^{(t)}\} \setminus \{\theta_k^{(t-1)}\}$.
    if do online update **or** $k = K$ **then**
      $\tilde{m}(w) = \frac{1}{K}\sum_{\theta \in \Theta} q(w|\theta)$.
Output: Θ.

Experiments: colorful MNIST

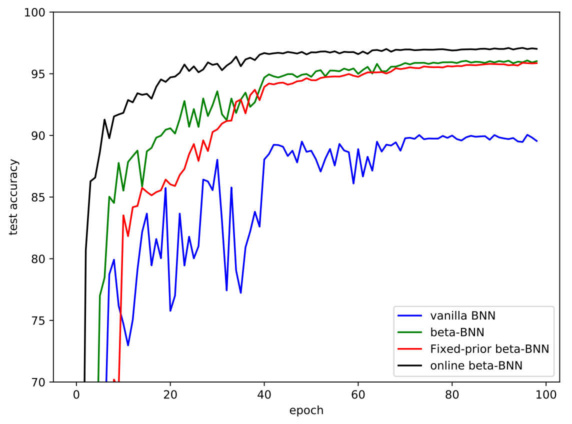
Baselines:

- Vanilla BNN: Blundell et al. (2015).
- Fixed-prior β-BNN: $\tilde{m}(w) \equiv \mathcal{N}(0, I)$.

Algorithm	β^*	Accuracy
Vanilla BNN	$\frac{1}{n}$	90.05
Fixed-prior β-BNN	10^{-10}	95.86
β-BNN	10^{-5}	96.08
Online β-BNN	10^{-3}	97.12

Experiments: colorful MNIST

Test accuracy over training epochs:



An Application to Teacher-Student Transfer

Deep learning is data-hungry

Issue: over-parameterized models are often trained with huge data.

- Medical applications is constrained by the number of patients of a particular disease.
- Semantic segmentation requires pixel-level annotation.

A potential **solution**: transfer learning.

- *Finetuning*: initialize with the weights of the source network.
- *Teacher-student knowledge transfer* by Ba and Caruana (2014); Hinton et al. (2015).

Teacher-student knowledge transfer: related work

There is no commonly agreed theory behind knowledge transfer.

$$\min \sum_i \|t^{(i)} - \mathbf{W}^{(i)} s^{(i)}\|$$

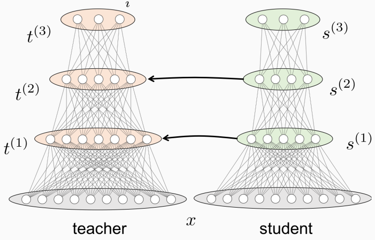


Figure 1: FitNet by Romero et al. (2014).

$$\min \sum_i \|f(t^{(i)}) - f(s^{(i)})\|_2^2$$

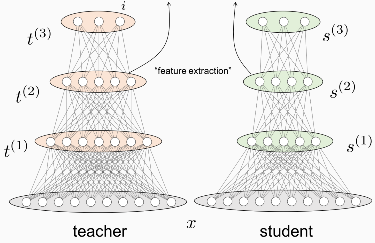
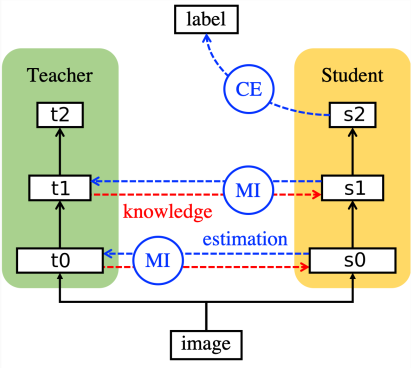


Figure 2: Attention transfer by Zagoruyko and Komodakis (2016).

Mutual information for knowledge transfer

Denote by $\mathbf{t}$ and $\mathbf{s}$ the activations of the teacher and the student respectively. Intuitively, $I(\mathbf{t};\mathbf{s})$ is maximized when $\mathbf{t} = \mathbf{s}$.



Variational information distillation (VID)

Knowledge transfer as a regularization:

$$\mathcal{L} = \mathcal{L}_{\text{task}} - \sum_{k=1}^{K} \lambda_k I(\mathbf{t}^{(k)}, \mathbf{s}^{(k)}),$$

Recall the variational characterization:

$$I(p; q) = \max_{\phi(\mathbf{t}|\mathbf{s})} \tilde{I}(p, q, \phi)$$

Instead of searching for all valid ϕ, we focus on diagonal Gaussians:

$$-\log \phi(\mathbf{t}|\mathbf{s}) = \sum_{n=1}^{N} \log \sigma_n + \frac{(t_n - \mu_n(\mathbf{s}))^2}{2\sigma_n^2} + \text{constant},$$

A related problem: channel capacity estimation

Noisy channel decoding theorem

Given a noisy channel from X to Y with transition $q(y|x)$, the channel capacity is given by

$$\begin{aligned} C &= \max_{p(x)\in\Delta} I(p,q) \\ &= \max_{p(x)\in\Delta} \max_{\phi(x|y)\in\Delta} \tilde{I}(p,q,\phi). \end{aligned}$$

Experiments: transfer from ImageNet to birds

Dataset: Caltech-UCSD Birds 200.

Networks: teacher (ResNet-34), student (ResNet-18).

data per class	≈29.95	20	10	5
Student	37.22	24.33	12.00	7.09
Finetuned	76.69	71.00	59.25	44.07
LwF	55.18	42.13	26.23	14.27
FitNet	66.63	56.63	46.68	31.04
AT	54.62	41.44	28.90	16.55
NST	55.01	41.87	23.76	15.63
VID	**73.25**	**67.20**	**56.86**	**46.21**

Experiments: transfer from ImageNet to indoor scenes

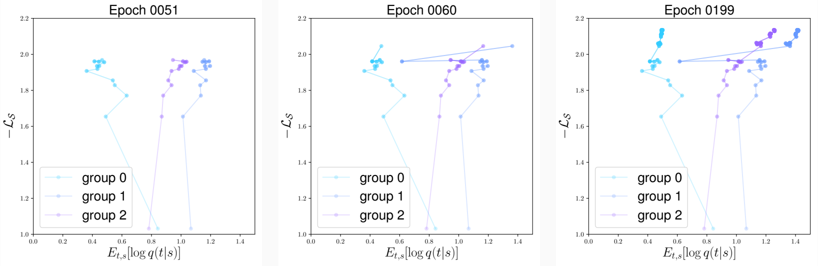
Dataset: MIT-67.

Networks: teacher (ResNet-34), student (VGG-9).

data per class	≈80	50	25	10
Student	53.58	43.96	29.70	15.97
Finetuned	65.97	58.51	51.72	39.63
LwF	60.90	52.01	41.57	27.76
FitNet	70.90	64.70	54.48	40.82
AT	60.90	52.16	42.76	25.60
NST	55.60	46.04	35.22	21.64
VID	**72.01**	**67.01**	**59.33**	**45.90**

Relationship between task loss and VID

Two-stage transition: before epoch 51, only $-\mathcal{L}_{\mathcal{S}}$ increases significantly, $\mathbb{E}_{\mathbf{t},\mathbf{s}}[\log \phi(\mathbf{t}|\mathbf{s})]$ barely changes, so does $I(\mathbf{t};\mathbf{s})$; the first stage ends at epoch 60; at the second stage, $I(\mathbf{t};\mathbf{s})$ slowly increases, which also drives $-\mathcal{L}_{\mathcal{S}}$ increasing.



Experiments: transfer from CNNs to MLPs

Dataset: CIFAR-10.

Networks: teacher (WRN-40-2), student (MLP).

Network	MLP-4096	MLP-2048	MLP-1024
Student	70.60	70.78	70.90
KD	70.42	70.53	70.79
FitNet	76.02	74.08	72.91
VID	**85.18**	**83.47**	**78.57**
Urban et al. (2017)		74.32	
Lin et al. (2015)		78.62	

Questions?

References

Achille, A. and Soatto, S. (2017). Emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*.

Ba, J. and Caruana, R. (2014). Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Graves, A. (2011). Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356.

References ii

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. ACM.

Lin, Z., Memisevic, R., and Konda, K. (2015). How far can we go without convolution: Improving fully-connected networks. *arXiv preprint arXiv:1511.02580*.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. (2018). Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*.

References iii

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2014). Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*, pages 1–5. IEEE.

Urban, G., Geras, K. J., Kahou, S. E., Aslan, O., Wang, S., Mohamed, A., Philipose, M., Richardson, M., and Caruana, R. (2017). Do deep convolutional nets really need to be deep and convolutional? In *ICLR*.

Zagoruyko, S. and Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.