Towards Efficient Learning of Graphical Models and Neural Networks with Variational Techniques

Shell Xu Hu

École des Ponts ParisTech – Université Paris-Est

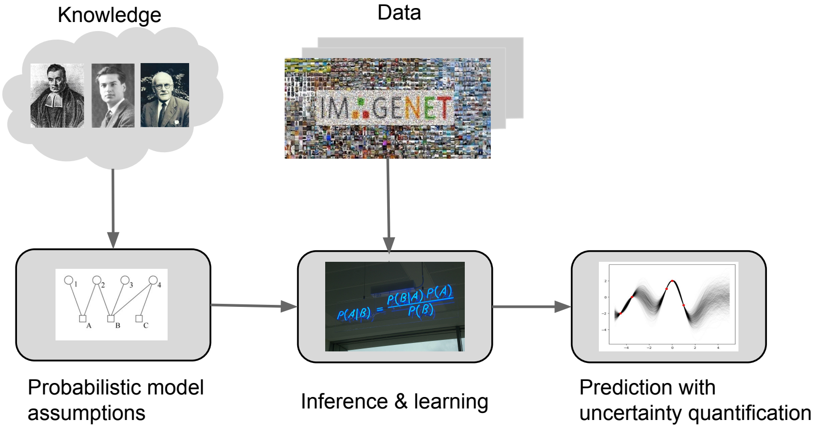
Table of contents

Related publications

- Shell X. Hu and Guillaume Obozinski. "SDCA-Powered Inexact Dual Augmented Lagrangian Method for Fast CRF Learning." International Conference on Artificial Intelligence and Statistics (AISTATS), 2018.
- Shell X. Hu, Pablo G. Moreno, Andreas Damianou, Neil D. Lawrence. "β-BNN: A Rate-Distortion Perspective on Bayesian Neural Networks." NeurIPS Workshop on Bayesian Deep Learning, 2018
- **Shell X. Hu, Pablo G. Moreno, Xi Shen, Yang Xiao, Guillaume Obozinski, Neil D. Lawrence, Andreas Damianou. "Empirical Bayes Transductive Meta-Learning with Synthetic Gradients." NeurIPS Workshop on Meta-Learning, 2019. (Long version submitted to ICLR 2020)**
- **Sungsoo Ahn, Shell X. Hu, Andreas Damianou, Neil D. Lawrence, Zhenwen Dai. "Variational Information Distillation for Knowledge Transfer." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.**
- Shell X. Hu, Sergey Zagoruyko, Nikos Komodakis. "Exploring Weight Symmetry in Deep Neural Networks." Computer Vision and Image Understanding (CVIU), 187, p.102786, 2019.

Probabilistic machine learning

The probabilistic pipeline



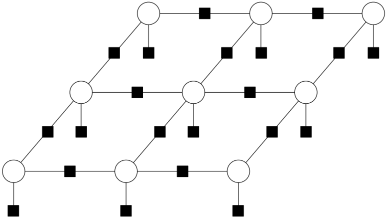
How do we model a random vector $x = (x_1, \ldots, x_n)$ when n is large?

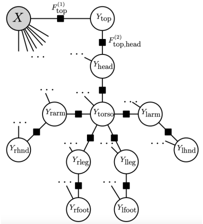
- Semantic segmentation: $x \in \{1, \ldots, K\}^n$.
- Human pose estimation: $x \in \mathbb{R}^n$.

PGMs are special distributions where **conditional independence** (CI) assumptions are made to enable a factorization according to a graph G:

$$p(x) \propto \prod_{a \in \mathcal{A}} \psi_a(x_a),$$

where $\mathcal{A}$ is a set of cliques in G.





Latent variable models (LVMs)

What if we have no idea how to make CI assumptions among x?

- A LVM introduces a latent variable z with joint distribution

$$p(x, z),$$

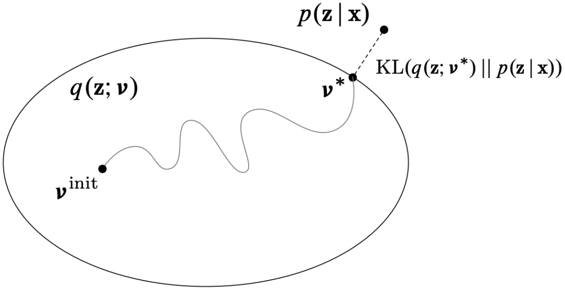
  which is the underpinning of *deep generative models* and *Bayesian neural networks*.

- **Inference** about z based on the data is through **posterior**

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

- Since $p(x) = \int p(x, z)dz$ is intractable, we appeal to **approximate posterior inference**.

Variational inference for LVMs



- VI casts **inference as optimization**.
- Posit a **variational family** of distributions of the form

$$q(z; \nu).$$

- Fit the **variational posterior** $q(z; \nu)$ to be close to the true posterior $p(z|x)$ in terms of some divergence measure (e.g. Kullback-Leibler).

Variational inference for LVMs: derivation

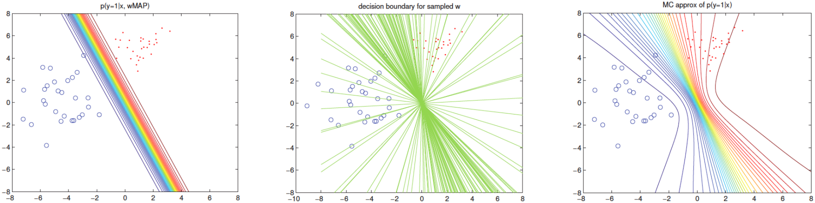
To compute $\log p(x) = \log \int p(x,z)dz$, the key idea to find an *evidence lower bound* (ELBO) by Jensen's inequality:

$$\begin{aligned}\log \int p(x,z)dz &= \log \int q(z)\frac{p(x,z)}{q(z)}dz \\ &\geq \int q(z)\log\frac{p(x,z)}{q(z)}dz =: \text{ELBO} \\ &= -D_{\text{KL}}\big(q(z)\|p(z|x)\big) + \log p(x).\end{aligned}$$

Thus, we have the classical **equivalence**

$$\max_q \text{ ELBO} \quad \Leftrightarrow \quad \min_q \ D_{\text{KL}}\big(q(z)\|p(z|x)\big).$$

Case study: Bayesian models



(a) (b) (c)

- Frequentist's parametric model: $p(y_{\text{test}}|x_{\text{test}}; w_{\text{train}})$.
- Bayesian's non-parametric model:

$$p(y_{\text{test}}|x_{\text{test}}, \mathcal{D}_{\text{train}}) = \int_{\mathcal{W}} p(y_{\text{test}}|x_{\text{test}}, w) p(w|\mathcal{D}_{\text{train}}) dw$$

.

- Compute the *posterior* via **Bayes rule**?

$$p(w|\mathcal{D}_{\text{train}}) = \frac{p(\mathcal{D}_{\text{train}}|w)p(w)}{p(\mathcal{D}_{\text{train}})},$$

given the *likelihood* $p(\mathcal{D}_{\text{train}}|w)$ and the *prior* $p(w)$.

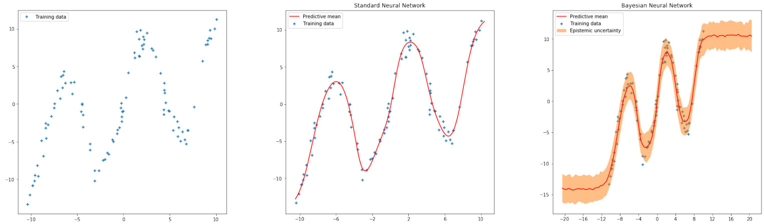
Case study: Bayesian models

- In general, unless a conjugate prior is considered for the likelihood, the posterior cannot be computed in closed form.
- Alternatively, we do **variational inference**:

$$q_{\text{train}} = \arg\min_{q \in \mathcal{Q}} D_{\text{KL}}\Big(q(w) \parallel p(w|\mathcal{D}_{\text{train}}) \Big)$$

and make **prediction** through

$$q(y_{\text{test}}|x_{\text{test}}, \mathcal{D}_{\text{train}}) = \int_{\mathcal{W}} p(y_{\text{test}}|x_{\text{test}}, w)\ q_{\text{train}}(w) dw$$

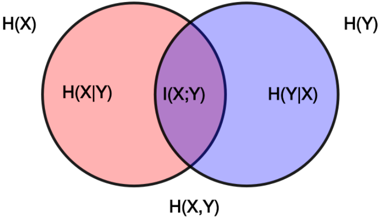


Information theoretical machine learning

Preliminary: Mutual information

Mutual information is used to measure statistical dependence

$$I(X;Y) := \mathbb{E}_{x,y\sim p(x,y)} \log \frac{p(x,y)}{p(x)p(y)}$$
$$= H(X,Y) - H(X|Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$
$$H(X) = I(X;X) = \text{ expected amount of information in } X$$



Mutual information: another variational tool

If we know the distribution of X and the joint distribution with decomposition $p(x, y) = p(x)q(y|x)$, then we can use mutual information to adjust $q(y)$ by either **minimizing** or **maximizing**

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

Note that the mutual information is a functional of p and q.

Variational characterization of mutual information

Computational issue: $I(X;Y)$ is intractable!

Solution: using variational techniques to derive bounds:

Lemma [Cover and Thomas, 2012, Theorem 10.8.1]

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

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

Rate-distortion (RD) tradeoff and information bottleneck

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

min **Rate**

s.t. **Distortion** $\leq$ const

$$\min_{q(\hat{x}|x)} \; I_q(X;\hat{X})$$

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

The **information bottleneck (IB)** [Tishby et al., 2000] is a extension for supervised learning where the distortion is defined in terms of the relevance wrt the label Y:

$$d(x,\hat{x}) = D_{\mathrm{KL}}\big(p(y|x) \| p(y|\hat{x})\big).$$

A more common form reads as (assuming $p(y|x)$ is fixed)

$$\min_{q(\hat{x}|x)} I_q(X;\hat{X}) - \beta \; I_q(Y;\hat{X}).$$

Rate-distortion based Bayesian inference

For a dataset S, consider a latent variable model defined by

$$\text{Generative process}: \quad P(S, w) = p(S \mid w)p(w).$$

The variational posterior $q(w|S)$ induces

$$\text{Inference process}: \quad q(S, w) = q(w \mid S)q^*(S).$$

The **Bayesian version of the information bottleneck (BIB)** [Achille and Soatto, 2017] can be derived from the RD tradeoff [Hu et al., 2018]:

$$\begin{aligned}
&\min_{q(w|S)} I_q(w; S) + \beta\ H_{q,p}(S|w) \\
&\text{where } H_{q,p}(S|w) := \mathbb{E}_{p^*(S)}\mathbb{E}_{q(w|S)} d(w, S) \\
&\text{and } d(w, S) := -\log p(S \mid w).
\end{aligned}$$

This is an alternative objective for variational inference.

Empirical Bayes transductive meta-learning with synthetic gradients

Meta-learning: a framework for small-data problems

Definition (meta-learning)

The problem is to solve rapidly a new task after learning several other similar tasks, where the dataset is a two-level hierarchy – dataset of datasets, one for each task. Meta-learning is sometimes called **learning to learn** [Schmidhuber, 1987, Thrun and Pratt, 1998].

Applications:

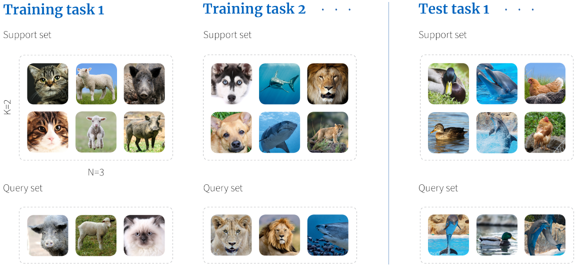
- Learning to do gradient descent [Andrychowicz et al., 2016].
- Learning to classify unseen categories [Vinyals et al., 2016].
- Learning to generalize across domains [Li et al., 2017].

An example: few-shot classification

Few-shot learning [Vinyals et al., 2016]:

	Support set	Query set	
	$d_t^l := \{(x_{t,i}^l, y_{t,i}^l)\}_{i=1}^{n^l}$	$x_t := \{x_{t,i}\}_{i=1}^{n}$	$y_t = \{y_{t,i}\}_{i=1}^{n}$
	Labeled data	Unlabeled data	
Training	✓	✓	✓
Testing	✓	✓	✗

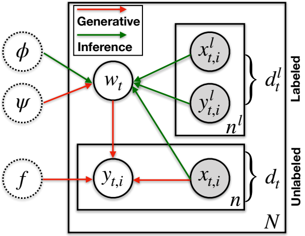
N-way-K-shot setup:



From hierarchical Bayes (HB) to empirical Bayes (EB)

Consider N training tasks with associated data $\mathcal{D} := \{d_t := (x_t, y_t)\}_{t=1}^{N}$:

$$\text{HB} \rightarrow \text{EB}: \quad p_f(\mathcal{D}) \rightarrow p_{\psi,f}(\mathcal{D}) = \int_{\psi} \left[\prod_{t=1}^{N} \int_{w_t} p_f(d_t|w_t) p(w_t|\psi) \right] p(\psi),$$



$$\log p_f(d_t|w_t)$$

$$= \sum_{i=1}^{n} \log p_f(y_{t,i}|x_{t,i}, w_t) + \log p(x_{t,i}|w_t)$$

$$= -\sum_{i=1}^{n} \ell_t(\hat{y}_{t,i}(f(x_{t,i}), w_t), y_{t,i}) + \text{const}$$

Variational inference for empirical Bayes

We derive an ELBO on the log-likelihood by introducing a variational distribution $q_{\theta_t}(w_t)$ for each task with parameter θ_t:

$$\log p_{\psi,f}(\mathcal{D}) \geq \sum_{t=1}^{N} \Big[\mathbb{E}_{w_t \sim q_{\theta_t}} \big[\log p_f(d_t | w_t) \big] - D_{\mathrm{KL}}\big(q_{\theta_t}(w_t) \| p_\psi(w_t) \big) \Big].$$

Maximizing the ELBO with respect to $\theta_1, \ldots, \theta_N$ and ψ is equivalent to

$$\min_{\theta_1,\ldots,\theta_N} \sum_{t=1}^{N} D_{\mathrm{KL}} \Big(q_{\theta_t}(\mathrm{w}_t) \, \Big\| \, p_{\psi,f}(\mathrm{w}_t | \mathrm{d}_t) \Big)$$

Amortized inference [Kingma and Welling, 2013] with transduction

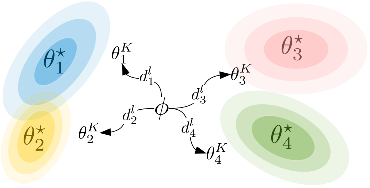
$$\text{Exact VI :} \qquad \min_{\theta_1,\ldots,\theta_N} \sum_{t=1}^{N} D_{\mathrm{KL}}\left(q_{\theta_t}(\mathrm{w}_t) \,\Big\|\, p_{\psi,f}(\mathrm{w}_t|\mathrm{d}_t) \right)$$

For scalable inference, we introduce a neural network ϕ to output θ_t.
There are **two choices to do the amortization**:

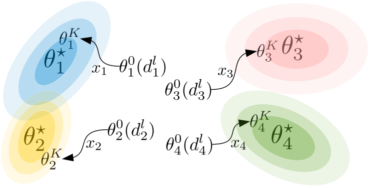
$$\text{Inductive AVI :} \qquad \min_{\phi} \sum_{t=1}^{N} D_{\mathrm{KL}}\left(q_{\phi(d_t^l)}(\mathrm{w}_t) \,\Big\|\, p_{\psi,f}(\mathrm{w}_t|\mathrm{d}_t) \right)$$

$$\text{Transductive AVI :} \qquad \min_{\phi} \sum_{t=1}^{N} D_{\mathrm{KL}}\left(q_{\phi(d_t^l, x_t)}(\mathrm{w}_t) \,\Big\|\, p_{\psi,f}(\mathrm{w}_t|\mathrm{d}_t) \right)$$

Motivation: to make use of the unlabeled data (i.e., x_t).



(b) MAML Finn et al. [2017]



(c) Our method

- MAML is an inductive method – only use the labeled data d_t^l to construct a Dirac delta variational posterior;
- We construct a better variational posterior as a function of both labeled data d_t^l and unlabeled data x_t.

Unrolling exact inference with synthetic gradient

How do we implement the amortization network $\phi(x_t, d_t^l)$?

The best is through the exact inference (only doable in training)

$$\phi(d_t^l, x_t) = \arg\min_{\theta_t} D_{\mathrm{KL}}\Big(q_{\theta_t}(\mathrm{w}_t) \;\Big\|\; p_{\psi,f}(\mathrm{w}_t|\mathrm{d}_t)\Big)$$

However, we don't have access to y_t in testing tasks. Instead, we unroll

$$\theta_t^{k+1} = \theta_t^k - \eta \nabla_{\theta_t} D_{\mathrm{KL}}\Big(q_{\theta_t^k}(w) \,\|\, p_{\psi,f}(w|d_t)\Big).$$

up to the K-th step by parameterizing

- the **initialization** θ_t^0;
- the **gradient** $\nabla_{\theta_t} D_{\mathrm{KL}}\Big(q_{\theta_t^k}(w) \,\|\, p_{\psi,f}(w|d_t)\Big)$.

Unrolling exact inference with synthetic gradient

Key observation: y_t only appears in $\partial \ell_t$ term.

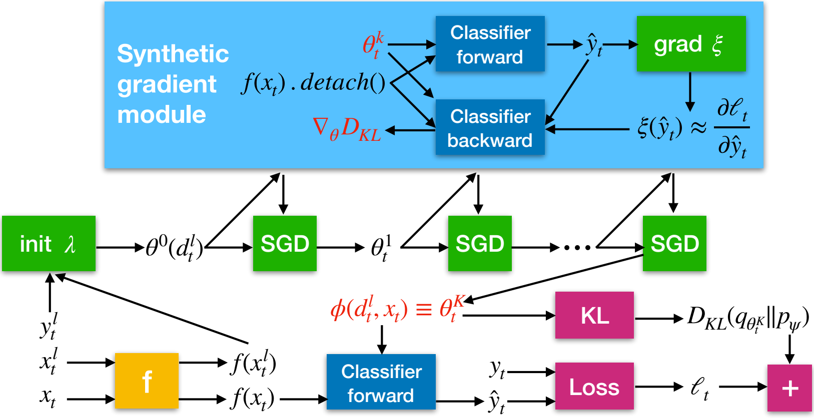
$$\nabla_{\theta_t} D_{\mathrm{KL}}\Big(q_{\theta_t} \| p_{\psi,f}\Big) = \mathbb{E}_\epsilon\Big[\sum_{i=1}^{n} \frac{\partial \ell_t(\hat{y}_{t,i}, y_{t,i})}{\partial \hat{y}_{t,i}} \frac{\partial \hat{y}_{t,i}}{\partial w_t} \frac{\partial w_t(\theta_t, \epsilon)}{\partial \theta_t}\Big] + \nabla_{\theta_t} D_{\mathrm{KL}}\Big(q_{\theta_t} \| p_\psi\Big).$$

By replacing $\frac{\partial \ell_t(\hat{y}_{t,i}, y_{t,i})}{\partial \hat{y}_{t,i}} \approx \xi(\hat{y}_{t,i})$, we can perform **synthetic gradient descent** without using y_t:

$$\theta_t^{k+1} = \theta_t^k - \eta \left[\mathbb{E}_\epsilon\Big[\sum_{i=1}^{n} \xi(\hat{y}_{t,i}) \frac{\partial \hat{y}_{t,i}}{\partial w_t} \frac{\partial w_t(\theta_t^k, \epsilon)}{\partial \theta_t}\Big] + \nabla_{\theta_t} D_{\mathrm{KL}}\Big(q_{\theta_t^k} \| p_\psi\Big)\right].$$

The idea of synthetic gradient was originally proposed by Jaderberg et al. [2017] for asynchronous forward and backward passes.

Computation graph of our method



Variational EM algorithm

while not converged **do**

Sample a task t and its data: d_t, d_t^l.

Compute the initialization $\theta_t^0 = \lambda(d_t^l)$.

%========== E-step ==========

for $k = 1, \dots, K$ **do**

$$\theta_t^{k+1} = \theta_t^k - \eta \text{ synthetic gradient.}$$

%========== M-step ==========

Update $\phi \leftarrow \phi - \eta \nabla_\phi D_{\mathrm{KL}}(q_{\phi(x_t, d_t^l)} \| p_f \cdot p_\psi)$.

Update $\psi \leftarrow \psi - \eta \nabla_\psi D_{\mathrm{KL}}(q_{\theta_t^K(\psi)} \| p_\psi)$.

Optionally, update $f \leftarrow f + \eta \nabla_f \log p_f(d_t | w_t)$.

Abstract form of empirical Bayes

Note that

$$\log p_{\psi,f}(w_1, \ldots, w_N, \mathcal{D}) = \sum_{t=1}^{N} \log p_f(d_t|w_t) + \log p_\psi(w_t)$$

is equal to the log-density of N iid samples drawn from

$$p(w, d, t) \equiv p_{\psi,f}(w, d, t) = p_f(d|w, t)p_\psi(w)q(t)$$

if $q(t)$ is uniform. Correspondingly, there is another decomposition

$$q(w, d, t) \equiv q_\phi(w, d, t) = q_\phi(w|d, t)q(d|t)q(t)$$

induced by the abstract variational posterior $q_\phi(w|d, t)$. When $N \to \infty$, we have an abstract form of EB:

$$\mathbb{E}_{q(t)}\mathbb{E}_{q(d|t)}\Big[\mathbb{E}_{q(w|d,t)}\big[-\log p(d|w, t)\big] + D_{\mathrm{KL}}\big(q(w|d, t)\|p(w)\big)\Big].$$

Empirical Bayes is related to information bottleneck

EB can be understood as matching the following processes:

Inference process : $q(w, d, t) = q(t)q(d|t)q(w \mid d, t)$

Generative process : $p(w, d, t) = p(d \mid w, t)p(w)q(t)$

Theorem

$$\mathbb{E}_{q(t)}\mathbb{E}_{q(d|t)}\Big[\mathbb{E}_{q(w|d,t)}\big[-\log p(d|w,t)\big] + D_{\mathrm{KL}}\big(q(w|d,t)\|p(w)\big)\Big]$$
$$\geq I_q(w; d|t) + H_{q,p}(d|w,t).$$

In light of this connection, we call our method **synthetic information bottleneck** (SIB).

Few-shot classification experiments

- **MiniImageNet** [Vinyals et al., 2016] contains 100 classes, split into 64 training classes, 16 validation classes and 20 testing classes, where each class consists of 600 image-label pairs and each image is of size 84×84.
- **CIFAR-FS** [Bertinetto et al., 2018] is created by dividing the original CIFAR-100 into 64 training classes, 16 validation classes and 20 testing classes; each image is of size 32×32.

Few-shot classification experiments

Method	Backbone	MiniImageNet, 5-way 1-shot	MiniImageNet, 5-way 5-shot	CIFAR-FS, 5-way 1-shot	CIFAR-FS, 5-way 5-shot
Matching Net [Vinyals et al., 2016]	Conv-4-64	44.2%	57%	–	–
MAML [Finn et al., 2017]	Conv-4-64	48.7±1.8%	63.1±0.9%	58.9±1.9%	71.5±1.0%
Prototypical Net [Snell et al., 2017]	Conv-4-64	49.4±0.8%	68.2±0.7%	55.5±0.7%	72.0±0.6%
Relation Net [Sung et al., 2018]	Conv-4-64	50.4±0.8%	65.3±0.7%	55.0±1.0%	69.3±0.8%
GNN [Satorras and Bruna, 2017]	Conv-4-64	50.3%	66.4%	61.9%	75.3%
R2-D2 [Bertinetto et al., 2018]	Conv-4-64	49.5±0.2%	65.4±0.2%	62.3±0.2%	77.4±0.2%
TPN [Liu et al., 2018]	Conv-4-64	55.5%	69.9%	–	–
Gidaris et al. [2019]	Conv-4-64	54.8±0.4%	**71.9±0.3%**	63.5±0.3%	**79.8±0.2%**
SIB K=0 (*Pre-trained feature*)	Conv-4-64	50.0±0.4%	67.0±0.4%	59.2±0.5%	75.4±0.4%
SIB η=1e-3, K=3	Conv-4-64	**58.0±0.6%**	70.7±0.4%	**68.7±0.6%**	77.1±0.4%
SIB η=1e-3, K=0	Conv-4-128	53.62 ± 0.79%	71.48 ± 0.64%	–	–
SIB η=1e-3, K=1	Conv-4-128	58.74 ± 0.89%	74.12 ± 0.63%	–	–
SIB η=1e-3, K=3	Conv-4-128	62.59 ± 1.02%	75.43 ± 0.67%	–	–
SIB η=1e-3, K=5	Conv-4-128	**63.26 ± 1.07%**	**75.73 ± 0.71%**	–	–
TADAM [Oreshkin et al., 2018]	ResNet-12	58.5±0.3%	76.7±0.3%	–	–
SNAIL [Santoro et al., 2017]	ResNet-12	55.7±1.0%	68.9±0.9%	–	–
MetaOptNet-RR [Lee et al., 2019]	ResNet-12	61.4±0.6%	77.9±0.5%	72.6±0.7%	84.3±0.5%
MetaOptNet-SVM [Lee et al., 2019]	ResNet-12	62.6±0.6%	78.6±0.5%	72.0±0.7%	84.2±0.5%
CTM [Li et al., 2019]	ResNet-18	64.1±0.8%	**80.5±0.1%**	–	–
Qiao et al. [2018]	WRN-28-10	59.6±0.4%	73.7±0.2%	–	–
LEO [Rusu et al., 2019]	WRN-28-10	61.8±0.1%	77.6±0.1%	–	–
Gidaris et al. [2019]	WRN-28-10	62.9±0.5%	79.9±0.3%	73.6±0.3%	**86.1±0.2%**
SIB K=0 (*Pre-trained feature*)	WRN-28-10	60.6±0.4%	77.5±0.3%	70.0±0.5%	83.5±0.4%
SIB η=1e-3, K=1	WRN-28-10	67.3±0.5%	78.2±0.3%	76.8±0.5%	84.9±0.4%
SIB η=1e-3, K=3	WRN-28-10	69.6±0.6 %	78.9±0.4%	78.4±0.6%	85.3±0.4%
SIB η=1e-3, K=5	WRN-28-10	**70.0±0.6%**	78.9±0.4%	**80.0±0.6%**	85.3±0.4%

Variational information distillation for knowledge transfer

Deep learning is data-hungry

Issue: over-parameterized neural networks are often trained with huge data, which is infeasible for certain applications, such as

- 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].

Prior work on teacher-student knowledge transfer

It works well empirically but there is no commonly agreed theory behind this framework.

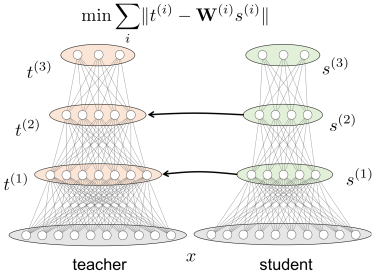


Figure 1: FitNet by Romero et al. [2014].

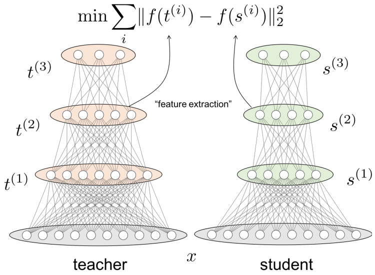
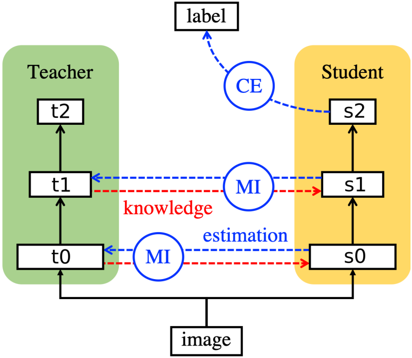


Figure 2: Attention transfer by Zagoruyko and Komodakis [2016].

Mutual information for knowledge transfer



Denote by $\boldsymbol{t}$ and $\boldsymbol{s}$ the activations of the teacher and the student respectively. Intuitively, $I(\boldsymbol{t};\boldsymbol{s})$ is maximized when $\boldsymbol{t}=\boldsymbol{s}$. The idea is to add a term to the *information bottleneck* principle [Tishby et al., 2000]:

$$\min I(\boldsymbol{x};\boldsymbol{s})+\beta\ H(\boldsymbol{y}|\boldsymbol{s})-\lambda\ I(\boldsymbol{t};\boldsymbol{s}).$$

Variational information distillation (VID)

Knowledge transfer as a regularizer with SGD:

$$\mathcal{L} = \text{Implicit regularization} + \text{Cross-entropy} - \sum_{k=1}^{K} \lambda_k I(\boldsymbol{t}^{(k)}, \boldsymbol{s}^{(k)}),$$

Recall the variational characterization:

$$\begin{aligned} I(\boldsymbol{t}; \boldsymbol{s}) &= H(\boldsymbol{t}) - H(\boldsymbol{t}|\boldsymbol{s}) \\ &= H(\boldsymbol{t}) + \mathbb{E}_{\boldsymbol{t},\boldsymbol{s}}[\log p(\boldsymbol{t}|\boldsymbol{s})] \\ &= H(\boldsymbol{t}) + \mathbb{E}_{\boldsymbol{t},\boldsymbol{s}}[\log q(\boldsymbol{t}|\boldsymbol{s})] + \mathbb{E}_{\boldsymbol{s}}[D_{\text{KL}}(p(\boldsymbol{t}|\boldsymbol{s})||q(\boldsymbol{t}|\boldsymbol{s}))] \\ &\geq H(\boldsymbol{t}) + \mathbb{E}_{\boldsymbol{t},\boldsymbol{s}}[\log q(\boldsymbol{t}|\boldsymbol{s})], \end{aligned}$$

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

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

Experiments: transfer from ImageNet to bird data

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-scene data

Dataset: MIT-67.

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

data per class	$\approx$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**

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	

Conclusion

- **Meta-learning**
  - Formulated transductive meta-learning with empirical Bayes model.
  - Implemented transductive amortized inference using synthetic gradient descent.
  - Achieved state-of-the-art results on few-shot learning benchmarks.
  - Derived the connection to Bayesian information bottleneck.
- **Transfer learning**
  - Proposed a teacher-student knowledge transfer framework inspired by information bottleneck.
  - Achieved state-of-the-art results on transfer learning benchmarks.
  - Empirically verified knowledge transfer between CNN and MLP.

Thank you for your attention!

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*, 2017.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, pages 3981–3989, 2016.

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

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *ArXiv*, abs/1805.08136, 2018.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. *arXiv preprint arXiv:1906.05186*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Shell X. Hu, G. Pablo Moreno, Lawrence D. Neil, and Andreas Damianou. β-bnn: A rate-distortion perspective on bayesian neural networks. *NeurIPS-BDL*, 2018.

Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1627–1635. JMLR, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, 2019.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017.

Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding Task-Relevant Features for Few-Shot Learning by Category Traversal. In *CVPR*, 2019.

Zhouhan Lin, Roland Memisevic, and Kishore Konda. How far can we go without convolution: Improving fully-connected networks. *arXiv preprint arXiv:1511.02580*, 2015.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018.

Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJgklhAcK7.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy P. Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017.

Victor Garcia Satorras and Joan Bruna. Few-shot learning with graph neural networks. *ArXiv*, abs/1711.04043, 2017.

Jurgen Schmidhuber. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. Phd thesis, Technische Universitat Munchen, Germany, 1987. URL http://www.idsia.ch/~juergen/diploma.html.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Kluwer Academic Publishers, 1998.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Gregor Urban, Krzysztof J. Geras, Samira Ebrahimi Kahou, Ozlem Aslan, Shengjie Wang, Abdelrahman Mohamed, Matthai Philipose, Matt Richardson, and Rich Caruana. Do deep convolutional nets really need to be deep and convolutional? In *ICLR*, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29*, pages 3630–3638. Curran Associates, Inc., 2016.

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