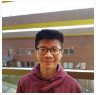
Information theoretical deep learning with variational techniques

Shell Xu Hu[1] Pablo G. Moreno[2] Andreas Damianou[2],
Sungsoo Ahn[2], Zhenwen Dai[2], Neil D. Lawrence[2,3],
Xi Shen[1] Yang Xiao[1], Guillaume Obozinski[1,4]

[1]École des Ponts ParisTech, [2]Amazon, [3]University of Cambridge, [4]Swiss Data Science Center

My research topics – focused on variational inference

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

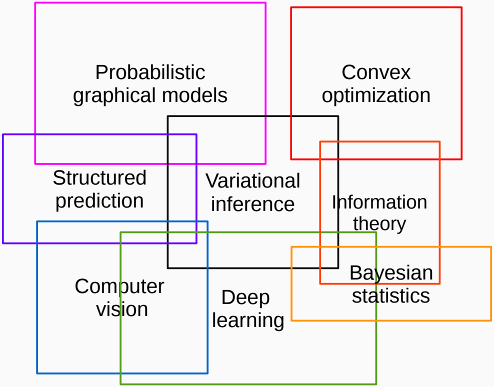


Table of contents

Introduction

The development of convolutional neural networks

CNNs have become the workhorses for computer vision.

- Many techniques, such as *residual connections*, *batch normalization*, have been developed to improve the performance on ImageNet.
- The overall trend was making CNNs deeper and wider.

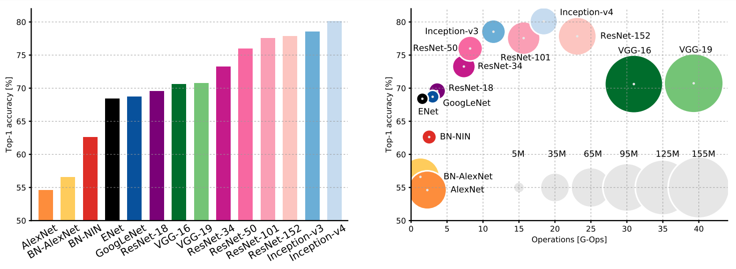
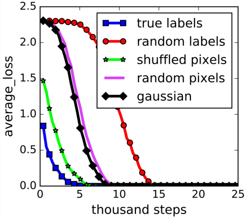


Figure 1: CNNs are over-parameterized. Source: Canziani et al. [2016].

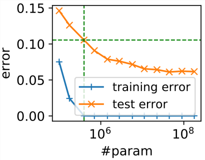
Over-parameterization and the generalization puzzle

Interesting observations on over-parameterized neural networks – the beginning of theoretical research on deep learning:

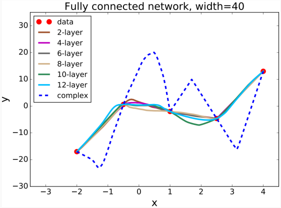
- Achieve zero training error even on damaged data.
- Generalize well on real data as #params increasing.
- Favor simple solutions.



Zhang et al. [2016]



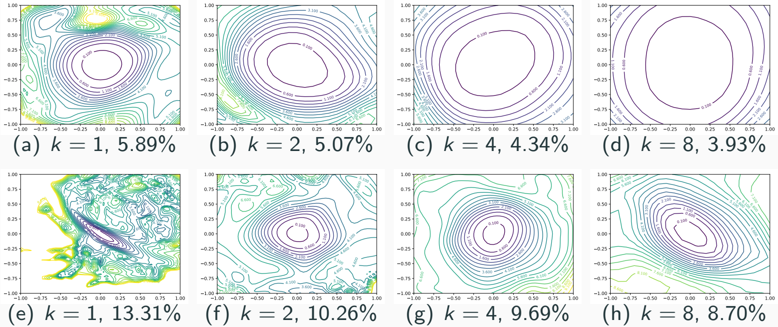
Neyshabur et al. [2018]



Wu et al. [2017]

Explanation I: Over-parameterization eliminates bad local minima and mitigates non-convexity

This explanation was suggested by Soudry and Hoffer [2017], Kawaguchi [2016], Lu and Kawaguchi [2017], Li et al. [2017], Haeffele and Vidal [2017], Wu et al. [2018].



(a) $k = 1$, 5.89% (b) $k = 2$, 5.07% (c) $k = 4$, 4.34% (d) $k = 8$, 3.93%

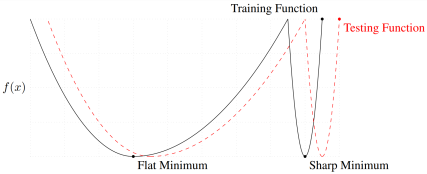
(e) $k = 1$, 13.31% (f) $k = 2$, 10.26% (g) $k = 4$, 9.69% (h) $k = 8$, 8.70%

Figure 2: With residual connections (top) and without (bottom). k = width factor, test error on CIFAR-10. Source: Li et al. [2017]

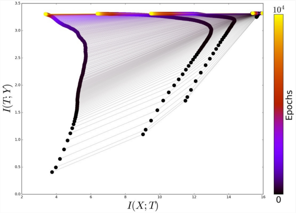
Explanation II: SGD biases towards low-complexity solutions

- Flat minima conjecture [Keskar et al., 2016, Berglund, 2011].
- Information bottleneck [Tishby et al., 2000, Tishby and Zaslavsky, 2015, Achille and Soatto, 2017]:

$$\min \qquad I(X;T) - \beta\, I(Y;T)$$



Keskar et al. [2016]



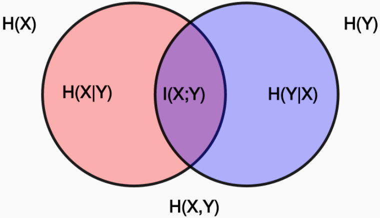
Tishby and Zaslavsky [2015]

Preliminary: Mutual information

Mutual information is used to measure 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 – a distribution manipulating tool

How do we make use of mutual information in machine learning/deep learning?

If we know the distribution of X and the joint distribution of X and Y decomposes as $p(x, y) = p(x)q(y|x)$, then we can employ mutual information to adjust the distribution of Y:

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

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

Variational characterization of mutual information

Computational issue: $I(X;Y)$ is intractable since $q(y|x)$ and $q(y) := \sum_x p(x)q(y|x)$ are coupled.

Solution: using variational techniques to derive bounds:

Lemma [Cover and Thomas, 2012, Theorem 10.8.1]

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

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

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.

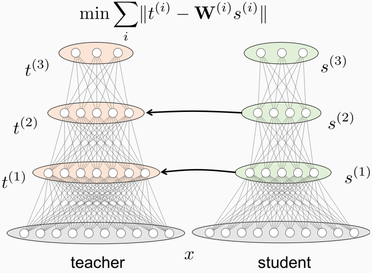


Figure 3: FitNet by Romero et al. [2014].

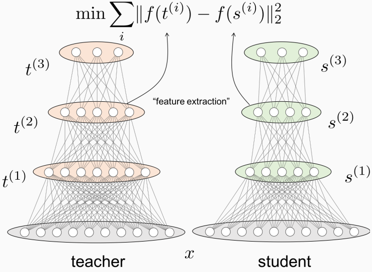
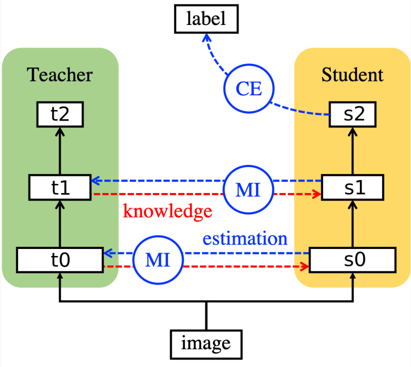


Figure 4: Attention transfer by Zagoruyko and Komodakis [2016].

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 $\boldsymbol{t} = \boldsymbol{s}$.



$\max I(\boldsymbol{t}; \boldsymbol{s})$ is inspired by *information bottleneck* [Tishby et al., 2000]:

$$\min I(\boldsymbol{x}; \boldsymbol{s}) - I(\boldsymbol{y}; \boldsymbol{s}).$$

Variational information distillation (VID)

Knowledge transfer as a regularization:

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

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.

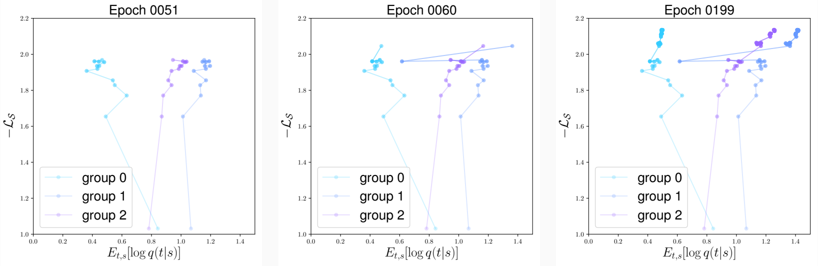
Networks: teacher (ResNet-34), student (VGG-9).

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**

Relationship between task loss and VID

Two-stage transition:

- Before epoch 51, only $\mathcal{L}_{\text{task}} \equiv \mathcal{L}_{\mathcal{S}}$ decreases significantly, $\mathbb{E}_{\boldsymbol{t},\boldsymbol{s}}[\log q(\boldsymbol{t}|\boldsymbol{s})]$ barely changes, so does $I(\boldsymbol{t};\boldsymbol{s})$;
- The first stage ends at epoch 60. At the second stage, $I(\boldsymbol{t};\boldsymbol{s})$ slowly increases, which also drives $-\mathcal{L}_{\mathcal{S}}$ increasing.



Experiments: transfer from CNNs to MLPs

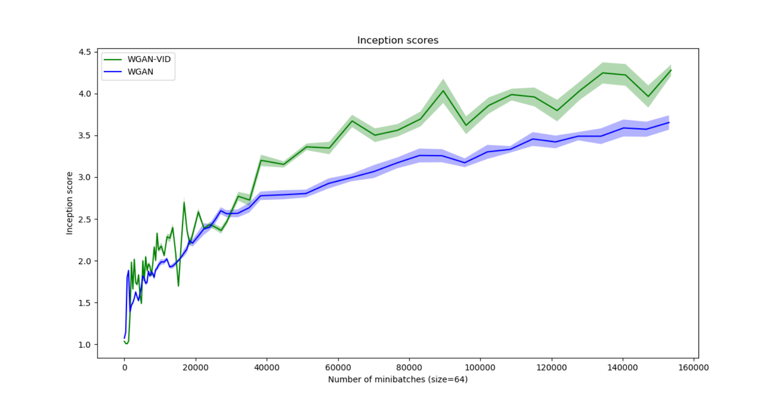
Dataset: CIFAR-10.

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	

Experiments: transfer from pretrained discriminator

Initializing from a pretrained discriminator will break the GAN balance. But a pretrained discriminator can be used to improve a poor discriminator (student).



Empirical Bayes transductive meta-Learning with synthetic gradients

Meta-learning: a framework for small-data problems

Definition: the problem of solving 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.

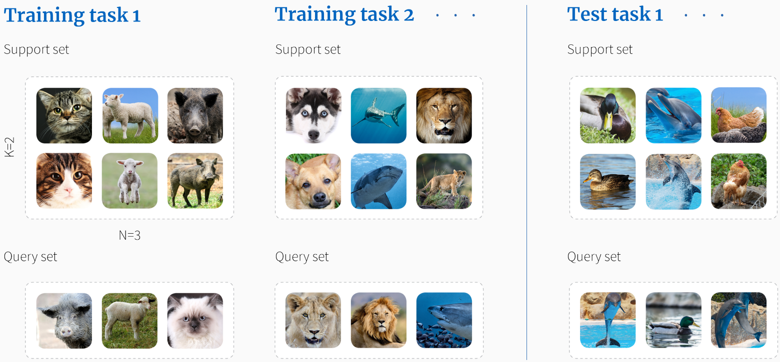
Meta-learning is sometimes called **learning to learn**.

Few-shot setting of meta-learning [Vinyals et al., 2016]

A task t, in *meta-testing*, consists of an *unlabeled set* $x_t := \{x_{t,i}\}_{i=1}^{n}$ and a *labeled set* $d_t^l := \{(x_{t,i}^l, y_{t,i}^l)\}_{i=1}^{n^l}$, and the goal is to predict $y_t = \{y_{t,i}\}_{i=1}^{n}$ corresponding to x_t. In *meta-training*, y_t is provided as ground truth.

An example: few-shot classification

N-way-K-shot setup:



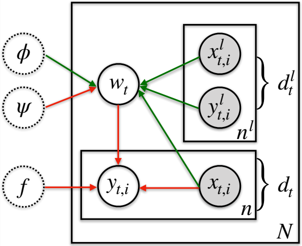
Empirical Bayes model for meta-learning

Consider a hierarchical Bayes model for the marginal likelihood

$$p_f(\mathcal{D}) = \int_\psi p_f(\mathcal{D}|\psi)p(\psi) = \int_\psi \Big[\prod_{t=1}^{N} \int_{w_t} p_f(d_t|w_t)p(w_t|\psi)\Big] p(\psi). \quad (1)$$

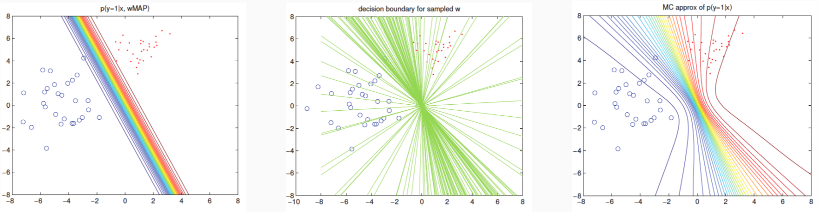
The *empirical Bayes* [Robbins, 1985, Kucukelbir and Blei, 2014], which interprets ψ in a frequentist way:

$$p_{\psi,f}(\mathcal{D}) = \prod_{t=1}^{N} p_\psi(d_t) = \prod_{t=1}^{N} \int_{w_t} p_f(d_t|w_t)p_\psi(w_t). \quad (2)$$



$$\begin{aligned}&\log p_f(d_t|w_t)\\ &= \sum_{i=1}^{n} \log p_f(y_{t,i}|x_{t,i}, w_t) + \log p(x_{t,i}|w_t)\\ &= -\frac{1}{n}\sum_{i=1}^{n} \ell_t\big(\hat{y}_{t,i}(f(x_{t,i}), w_t), y_{t,i}\big) + \text{const},\end{aligned}$$

Preliminary: why Bayesian?



(a) (b) (c)

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

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

.

- How to compute the *posterior*?

By Bayes' rule: $$p(w|\mathcal{D}_{\text{train}}) = \frac{p(\mathcal{D}_{\text{train}}|w)p(w)}{p(\mathcal{D}_{\text{train}})},$$

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

Preliminary: how practical is Bayesian?

- Unless a conjugate prior is considered for the likelihood, the posterior cannot be computed in closed form.
- Alternatively, we do *approximate Bayesian inference*:

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

  and make prediction through

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

- Most existing works, e.g., Blundell et al. [2015], opt to parameterize $q(w)$ as a Gaussian distribution – learning mean and variance.

Motivation: can we do better than classical Bayesian?

What can be improved?

- **Global vs. local**: is it necessary to condition on the entire train-set?

$$p(w|\mathcal{D}_{\text{train}}) \quad \text{vs.} \quad p(w|\mathcal{D}_{\text{context}})$$

  We will approximate $p(w|\mathcal{D}_{\text{context}})$ in the variational inference of empirical Bayes.

- **Domain shift**: to predict y_{test}, do we use

$$p(w|\mathcal{D}_{\text{train}}) \quad \text{or} \quad p(w|\mathcal{X}_{\text{test}}, \mathcal{D}_{\text{train}}) \quad \text{or} \quad p(w|\mathcal{D}_{\text{test}})?$$

Transduction: a dose to domain shift

- **Inductive learning** – $p(w|\mathcal{D}_{\text{train}})$: we first train a model on $\mathcal{D}_{\text{train}}$, and then test it on $\mathcal{D}_{\text{test}}$, one testing example at a time.
- **Transductive learning** – $p(w|\mathcal{X}_{\text{test}}, \mathcal{D}_{\text{train}})$: we are allowed to see all testing examples, i.e., $\mathcal{X}_{\text{test}}$, before making predictions.
- Cheating – $p(w|\mathcal{D}_{\text{test}})$:p

Amortized inference with transduction

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

$$\log p_{\psi,f}(\mathcal{D}) \geq \sum_{t=1}^{N} \Big[\mathbb{E}_{w_t \sim q_{\theta_t}} \big[\log p_f(d_t|w_t)\big] - D_{\mathrm{KL}}\big(q_{\theta_t}(w_t) \| p_\psi(w_t)\big)\Big]. \tag{3}$$

Maximizing the ELBO in equation (3) with respect to $\theta_1, \ldots, \theta_N$ and ψ is equivalent to

$$\min_{\psi} \min_{\theta_1,\ldots,\theta_N} \frac{1}{N} \sum_{t=1}^{N} D_{\mathrm{KL}}\Big(q_{\theta_t}(\mathrm{w}_t) \,\|\, p_f(d_t|w_t) p_\psi(\mathrm{w}_t)\Big), \tag{4}$$

Replacing each q_{θ_t} by $q_{\phi(x_t, d_t^l))}$, equation (4) can be written as

$$\min_{\psi} \min_{\phi} \frac{1}{N} \sum_{t=1}^{N} D_{\mathrm{KL}}\Big(q_{\phi(x_t, d_t^l)}(\mathrm{w}_t) \,\|\, p_f(d_t|w_t) p_\psi(\mathrm{w}_t)\Big), \tag{5}$$

Variational inference with synthetic gradients

How do we parameterize $\phi(x_t, d_t^l)$?

If we were able to have access to the groundtruth y_t, we would perform a stochastic gradient descent on θ_t for optimizing equation (4):

$$\theta_t^{k+1} = \theta_t^k - \eta \nabla_{\theta_t} D_{\mathrm{KL}}\Big(q_{\theta_t^k}(w) \,\|\, p_f(d_t|w) \cdot p_\psi(w)\Big). \tag{6}$$

Instead, we parameterize this optimization dynamics up to the K-th step via $\phi(x_t, d_t^l)$, such that $q_{\theta_t^K}$ is a good approximation of the optimum $q_{\theta_t^\star}$. It consists of parameterizing

- the **initialization** θ_t^0
- the **gradient** $\nabla_{\theta_t} D_{\mathrm{KL}}(q_{\theta_t} \,\|\, p_f \cdot p_\psi)$.

Variational inference with synthetic gradients

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

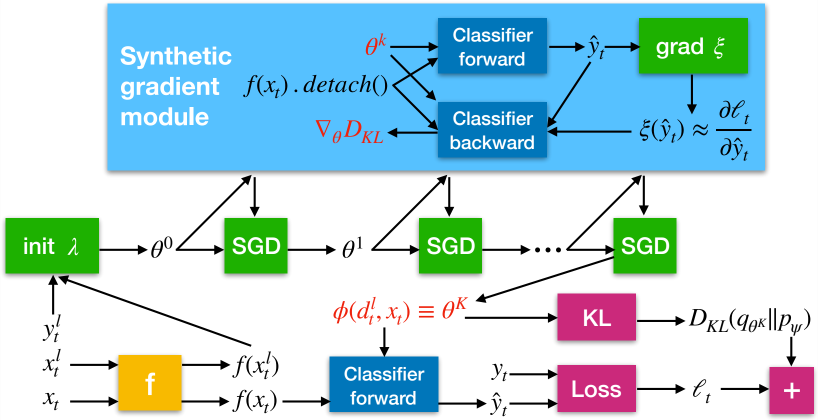
$$\nabla_{\theta_t} D_{\mathsf{KL}}\Big(q_{\theta_t} \| p_f \cdot p_\psi\Big) = \mathbb{E}_\epsilon\Big[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_t(\hat{y}_{t,i}, y_{t,i})}{\partial \hat{y}_{t,i}} \frac{\partial \hat{y}_{t,i}}{\partial w_t} \frac{\partial w_t(\theta_t, \epsilon)}{\partial \theta_t}\Big]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

under a reparameterization $w_t = w_t(\theta_t, \epsilon)$ with $\epsilon \sim p(\epsilon)$.

Now, we can perform synthetic gradient descent:

$$\theta_t^{k+1} = \theta_t^k - \eta \left[\mathbb{E}_\epsilon\Big[\frac{1}{n}\sum_{i=1}^{n} \xi(\hat{y}_{t,i}) \frac{\partial \hat{y}_{t,i}}{\partial w_t} \frac{\partial w_t(\theta_t^k, \epsilon)}{\partial \theta_t}\Big] + \nabla_{\theta_t} D_{\mathsf{KL}}\Big(q_{\theta_t^k} \| p_\psi\Big)\right]. \tag{7}$$

Variational inference with synthetic gradients: computation graph



Variational inference with synthetic gradients: algorithm

Input: the dataset $\mathcal{D}$; the step size η; the number of inner iterations K; pretrained f.
Initialize the meta-models ψ, and $\phi = (\lambda, \xi)$.
while not converged **do**
    Sample a task t and the associated dataset d_t (plus optionally the support set d_t^l).
    Compute the initialization $\theta_t^0 = \lambda$ or $\theta_t^0 = \lambda(d_t^l)$.
    for $k = 1, \ldots, K$ **do**

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

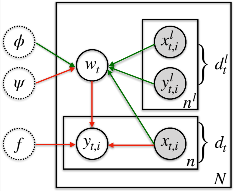
    Compute $w_t = w_t(\theta_t^K, \epsilon)$ with $\epsilon \sim p(\epsilon)$.
    Update $\psi \leftarrow \psi - \eta \nabla_\psi D_{\mathsf{KL}}(q_{\theta_t^K(\psi)} \| p_\psi)$.
    Update $\phi \leftarrow \phi - \eta \nabla_\phi D_{\mathsf{KL}}(q_{\phi(x_t, d_t^l)} \| p_f \cdot p_\psi)$.
    Optionally, update $f \leftarrow f + \eta \nabla_f \log p_f(d_t | w_t)$.

Empirical Bayes is equivalent to information bottleneck

An abstract view of the model (controllable pieces are marked in red):

$$\text{Inference}: \quad q(w,d,t) = q(t)q(d|t)q(w \mid d,t)$$

$$\text{Generative}: \quad p(w,d,t) = p(d \mid w,t)p(w)q(t)$$



Theorem

$$\min_{p(w)} \mathbb{E}_{q(t)}\mathbb{E}_{q(d|t)}\Big[D_{\mathrm{KL}}\big(q(w \mid d,t) \,\|\, p(w \mid d,t)\big)\Big]$$
$$= I_q(w; d \mid t) - \beta\, I_{q,p}(w; d \mid t) \text{ with } \beta = 1,$$

where I_q and $I_{q,p}$ are mutual information and cross mutual information respectively.

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

Mini-ImageNet experiments

		MiniImageNet, 5-way		CIFAR-FS, 5-way	
Method	**Backbone**	1-shot	5-shot	1-shot	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%

Questions?

References

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

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

Nils Berglund. Kramers' law: Validity, derivations and generalisations. *arXiv preprint arXiv:1106.5799*, 2011.

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

References ii

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

Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

O Chapelle, B Schölkopf, and A Zien. A discussion of semi-supervised learning and transduction. *Semi-Supervised Learning*, pages 457–462, 2006.

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

Misha Denil, Babak Shakibi, Laurent Dinh, Nando De Freitas, et al. Predicting parameters in deep learning. In *Advances in neural information processing systems*, pages 2148–2156, 2013.

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.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

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.

Benjamin D Haeffele and René Vidal. Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339, 2017.

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

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Alp Kucukelbir and David M Blei. Population empirical bayes. *arXiv preprint arXiv:1411.0292*, 2014.

References v

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

Hao Li, Zheng Xu, Gavin Taylor, and Tom Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 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.

Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 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.

Herbert Robbins. An empirical bayes approach to statistics. In *Herbert Robbins Selected Papers*, pages 41–47. Springer, 1985.

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.

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

Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 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.

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

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.

Chenwei Wu, Jiajun Luo, and Jason D Lee. No spurious local minima in a two hidden unit relu network. In *International Conference on Learning Representation Workshop*, 2018.

Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

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.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.