Dec 10, 2024

Meta-Learning Universal Priors Using Non-Injective Change of Variables

Yilang Zhang

Dept. of ECE, University of Minnesota

Aknowledgement: Alireza Sadeghi, Georgios B. Giannakis

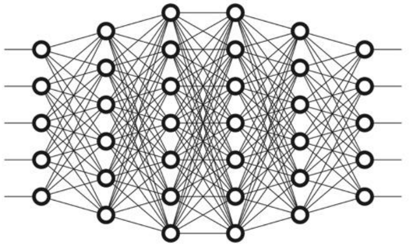
NSF grants 2220292, 2312547, 2212318, and 2126052

Motivating context of meta-learning

Challenge in deep learning: large-scale model vs. limited training data

Ex. ResNet-50 [He et al'15]

>23M parameters



VS.

HE-vs-MPM dataset [Han et al'23]

116 breast cancer images

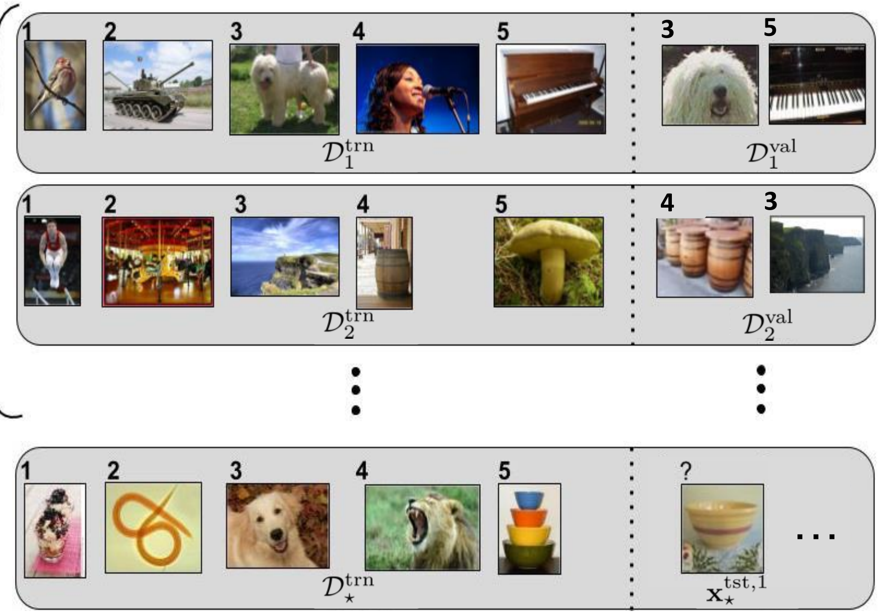
- Conventional supervised learning

$$\min_{\phi} \mathcal{L}(\phi; \mathcal{D}^{\text{trn}}) + \mathcal{R}(\phi)$$

- Model parameter $\phi \in \mathbb{R}^d$, training data $\mathcal{D}^{\text{trn}} = \{(\mathbf{x}^n, y^n)\}_{n=1}^{N^{\text{trn}}}$
- Loss $\mathcal{L}(\phi; \mathcal{D}^{\text{trn}}) = -\log p(\mathbf{y}^{\text{trn}}|\phi; \mathbf{X}^{\text{trn}})$, regularizer $\mathcal{R}(\phi) = -\log p(\phi)$ empirical prior
- Overfitting if $d \gg N^{\text{trn}}$ ➤ Rely on informative $\mathcal{R}(\phi)$

Remedy: extract and transfer task-invariant prior from related tasks learnable prior

Meta-learning in a nutshell



- Supervised meta-learning
  - Given:
    - Tasks $t = 1, \ldots, T$, each with $\mathcal{D}_t^{\text{trn}}, \mathcal{D}_t^{\text{val}}$
    - New task $\star$ with limited $\mathcal{D}_\star^{\text{trn}}$ and $\{\mathbf{x}_\star^{\text{tst},n}\}_{n=1}^{N_\star^{\text{tst}}}$
  - To-do: predict $\{y_\star^{\text{tst},n}\}_{n=1}^{N_\star^{\text{tst}}}$

✓ **Goal:** learn task-invariant prior from given tasks, with which new task can be solved

➢ Bilevel problem: task-specific parameter $\phi_t \in \mathbb{R}^d$, task-invariant meta-parameter $\boldsymbol{\theta} \in \mathbb{R}^D$

$$\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{\phi}_t^*(\boldsymbol{\theta}); \mathcal{D}_t^{\text{val}})$$

outer/meta-level

$$\text{s.t. } \boldsymbol{\phi}_t^*(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\phi}_t} \mathcal{L}(\boldsymbol{\phi}_t; \mathcal{D}_t^{\text{trn}}) + \mathcal{R}(\boldsymbol{\phi}_t; \boldsymbol{\theta}), \ \forall t$$

inner/task-level

$$\boldsymbol{\phi}_t^*(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\phi}_t} \mathcal{L}(\boldsymbol{\phi}_t; \mathcal{D}_t^{\text{trn}}, \boldsymbol{\theta}), \ \forall t$$

alternative: implicit prior

S. Ravi, and H. Larochelle, "Optimization as a model for few-shot learning," *ICLR*, 2017.

Expressiveness challenge in prior selection

Q. Which prior/regularizer to choose?

- Implicit prior via initialization
  - MAML [Finn et al'17]: Task-invariant initialization + GD

$$\phi_t^0 = \phi^{\text{init}} = \boldsymbol{\theta}, \ \forall t \qquad \phi_t^k = \phi_t^{k-1} - \alpha \nabla \mathcal{L}(\phi_t^{k-1}; \mathcal{D}_t^{\text{trn}}), \ k = 1, \ldots, K$$

Lemma [Grant et al'18]. *Under second-order approximation, MAML satisfies*

$$\phi_t^K(\boldsymbol{\theta}) \approx \phi_t^*(\boldsymbol{\theta}) = \arg\min_{\phi_t} \mathcal{L}(\phi_t; \mathcal{D}_t^{\text{trn}}) + \frac{1}{2}\|\phi_t - \boldsymbol{\theta}\|_{\boldsymbol{\Lambda}_t}^2$$

where $\boldsymbol{\Lambda}_t$ *is determined by* $\alpha, K, \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_t^{\text{trn}})$

- Implicit Gaussian prior $p(\phi_t; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_t^{-1})$

- Explicit prior via regularization
  - Isotropic Gaussian [Rajeswaran et al'19] $\mathcal{R}(\phi_t; \boldsymbol{\theta}) = \frac{\lambda}{2}\|\phi_t - \phi^{\text{init}}\|_2^2, \ \boldsymbol{\theta} := \{\phi^{\text{init}}, \lambda\}$
  - Diagonal Gaussian [Li et al'17], block-diagonal Gaussian [Park et al'19], …
  - Sparse [Tian et al'20], factorable + degenerate [Bertinetto et al'18, Lee et al'19], …

Challenge: preselected priors have limited expressiveness

C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *ICML*, 2017.
E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths, "Recasting gradient-based meta-learning as hierarchical Bayes," *ICLR,* 2018.

Data-driven priors via transform

Goal: data-driven prior $p(\phi_t; \boldsymbol{\theta})$ of sufficient expressiveness

Key idea: transform a known prior into the sought one

- ➢ Learning prior boils down to learning transform

- ❑ Conventional approaches:
  - GAN, VAE, diffusion model: tailored to nature signals
  - Normalizing flow (NF)

> Change-of-variable formula. *Let* $\mathbf{Z} \in \mathbb{R}^d$ *be a continuous random vector, and* $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ *a bijection. Then* $\mathbf{Z}' := f(\mathbf{Z})$ *has analytical pdf*
> $$p_{\mathbf{Z}'}(\mathbf{z}') = p_{\mathbf{Z}}(f^{-1}(\mathbf{z}')) \left|\det J_{f^{-1}}(\mathbf{z}')\right| = \frac{p_{\mathbf{Z}}(f^{-1}(\mathbf{z}'))}{|\det J_f(\mathbf{z}')|} \text{ (a.e.).}$$

- Probability integral transform (PIT): if d=1, the optimal $f^* = Q^{-1} \circ P_Z$
- If d>1, f^* may not exist [Kong et al'20, Sec. 4]

$P_Z, Q:$ source, target cdfs

- ➢ Limited expressiveness especially in high-dimensional spaces

Learning universal prior via non-injective change-of-variables

- **Our approach:** non-injective change-of-variable (NCoV)

> Theorem 1 (Multivariate PIT). *Let $\mathbf{Z} \in \mathbb{R}^d$ be a continuous random vector with mutually independent entries. For any differentiable a.e. cdf $Q : \mathbb{R}^d \mapsto [0,1]$, there exists $f^* : \mathbb{R}^d \mapsto \mathbb{R}^d$ for which $\mathbf{Z}' := f^*(\mathbf{Z})$ has cdf*
> $$P_{\mathbf{Z}'} = Q \text{ (a.e.)}.$$

- Q is arbitrary (even discrete), and f^* can be non-injective
- Limitation: transformed pdf may be intractable

$$p_{\mathbf{Z}'}(\mathbf{z}') = \int_{\mathbb{R}^d} p_{\mathbf{Z}}(\mathbf{z})\delta(\mathbf{z}' - f^*(\mathbf{z}))d\mathbf{z}$$

Alternative: numerical integration when d is small

- Meta-learning with NCoVs

Target pdf q is $p(\boldsymbol{\phi}_t;\boldsymbol{\theta})$; use parametric $f(\cdot;\boldsymbol{\theta})$; task-level optimizes latent variable $\mathbf{z}_t$

$$\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \mathcal{L}_t^{\text{val}}(\underbrace{f(\mathbf{z}_t^*(\boldsymbol{\theta});\boldsymbol{\theta})}_{\boldsymbol{\phi}_t^*})$$
$$\text{s.t. } \mathbf{z}_t^*(\boldsymbol{\theta}) = \arg\min_{\mathbf{z}_t} \mathcal{L}_t^{\text{trn}}(\underbrace{f(\mathbf{z}_t;\boldsymbol{\theta})}_{\boldsymbol{\phi}_t}) - \log p_{\mathbf{Z}}(\mathbf{z}_t), \ \forall t$$

Ex. $p_{\mathbf{Z}} = \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$

→ $-\log p_{\mathbf{Z}}(\mathbf{z}_t) = \frac{1}{2}\|\mathbf{z}_t\|_2^2$

$\mathbf{z}_t^0 = \mathbf{0}_d$

Side benefit: inherent initialization $\mathbf{z}_t^0 = \arg\max_{\mathbf{z}_t} p_{\mathbf{Z}}(\mathbf{z}_t)$ via maximum a priori

Numerical tests

- Few-shot classification

Method	Prior model	5-class miniImageNet 1-shot (%)	5-class miniImageNet 5-shot (%)
Meta-LSTM [41]	RNN-based	$43.44_{\pm 0.77}$	$60.60_{\pm 0.71}$
MAML [10]	implicit Gaussian	$48.70_{\pm 1.84}$	$63.11_{\pm 0.92}$
MetaSGD [29]	diagonal Gaussian	$50.47_{\pm 1.87}$	$64.03_{\pm 0.94}$
R2D2 [3]	degenerate body & Gaussian head	$51.8_{\pm 0.2}$	$68.4_{\pm 0.2}$
MC [37]	block-diagonal Gaussian	$54.08_{\pm 0.93}$	$67.99_{\pm 0.73}$
Warp-MAML [12]	Gaussian	$52.3_{\pm 0.8}$	$68.4_{\pm 0.6}$
MAML + L2F [2]	implicit Gaussian	$52.10_{\pm 0.50}$	$69.38_{\pm 0.46}$
MeTAL [1]	implicit Gaussian	$52.63_{\pm 0.37}$	$70.52_{\pm 0.29}$
Minimax-MAML [58]	inverted Gaussian & entropy	$51.70_{\pm 0.42}$	$68.41_{\pm 1.28}$
MAML + MetaNCoV	NCoV-based	$\mathbf{57.74}_{\pm 1.47}$	$70.72_{\pm 0.70}$
MetaSGD + MetaNCoV	NCoV-based	$\mathbf{59.10}_{\pm \mathbf{1.52}}$	$\mathbf{71.48}_{\pm \mathbf{0.68}}$

- Cross-domain generalization

Method	5-class TieredImageNet 1-shot (%)	5-class TieredImageNet 5-shot (%)	5-class CUB 1-shot (%)	5-class CUB 5-shot (%)	5-class Cars 1-shot (%)	5-class Cars 5-shot (%)
MAML [10]	$51.61_{\pm 0.20}$	$65.76_{\pm 0.27}$	$40.51_{\pm 0.08}$	$53.09_{\pm 0.16}$	$33.57_{\pm 0.14}$	$44.56_{\pm 0.21}$
ANIL [38]	$52.82_{\pm 0.29}$	$66.52_{\pm 0.28}$	$41.12_{\pm 0.15}$	$55.82_{\pm 0.21}$	$34.77_{\pm 0.31}$	$46.55_{\pm 0.29}$
BOIL [35]	$53.23_{\pm 0.41}$	$69.37_{\pm 0.23}$	$44.20_{\pm 0.15}$	$60.92_{\pm 0.11}$	$36.12_{\pm 0.29}$	$50.64_{\pm 0.22}$
SparseMAML+ [56]	$53.91_{\pm 0.67}$	$69.92_{\pm 0.21}$	$43.43_{\pm 1.04}$	$62.02_{\pm 0.78}$	$37.14_{\pm 0.77}$	$53.18_{\pm 0.44}$
GAP [19]	$58.56_{\pm 0.93}$	$\mathbf{72.82}_{\pm 0.77}$	$44.74_{\pm 0.75}$	$\mathbf{64.88}_{\pm 0.72}$	$38.44_{\pm 0.77}$	$55.04_{\pm 0.77}$
MetaNCoV	$\mathbf{61.50}_{\pm \mathbf{1.49}}$	$\mathbf{73.10}_{\pm \mathbf{0.74}}$	$\mathbf{47.84}_{\pm \mathbf{1.49}}$	$\mathbf{65.27}_{\pm \mathbf{0.73}}$	$\mathbf{41.66}_{\pm \mathbf{1.48}}$	$\mathbf{57.19}_{\pm \mathbf{0.75}}$

- Check our paper for additional analytical and experimental results

Thank you!