Talk 4: Personalization in Federated Learning – II

WEN Hao

2021-6-10

Personalization for FL

When does one need personalization?

— When data across clients are “enough” non-IID, which is more realistic.

Personalization for FL

When does one need personalization?

— When data across clients are “enough” non-IID, which is more realistic.

Means of personalization:

- Federated Multi-Task Learning (+ regularization / proximal term), e.g. [1]
- Model-Agnostic Meta Learning, e.g. [2]
- Local Fine-tuning.
- etc.

1 Model-Agnostic Meta Learning

2 Federated Multi-Task Learning

Model-Agnostic Meta Learning

"The goal of meta-learning is to train a model on a **variety of learning tasks**, *such that it can solve new learning tasks using only a small number of training samples."* *– [2]*

i.e. over a distribution of learning tasks $p(\mathcal{T})$, where

$$\mathcal{T} = \{\mathcal{L}(\{(x_t, a_t)\}), q(x_1), q(x_{t+1}|x_t, a_t), H\}$$

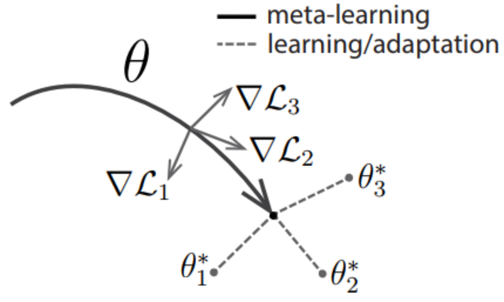
with

(x_t, a_t) :	data points
$\mathcal{L}$:	loss function
$q(x_1)$:	initial distribution
$q(x_{t+1}\|x_t, a_t)$:	transition
H :	episode length

Model-Agnostic Meta Learning – Intuition

Intuition of MAML

Some internal representations are more transferrable than others. MAML should encourage the emergence of such general-purpose representations via searching for model parameters that are sensitive to changes in the task.

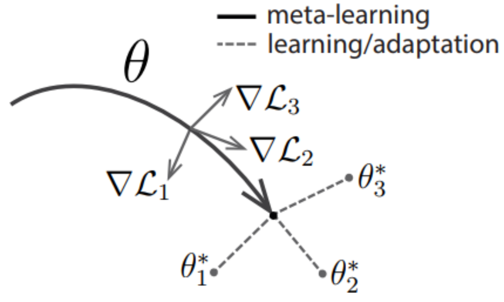


Model-Agnostic Meta Learning – Formulation

Mathematically, MAML can be formulated as a (bi-level?) optimization problem

$$\text{minimize} \quad \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$$

$$\text{where} \quad \theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$$



Model-Agnostic Meta Learning – Algorithm

Algorithm 1: MAML[2]

Require: $p(\mathcal{T})$ distribution over tasks

Require: α, β step size hyper-params

randomly initialize model params θ

while *not done* **do**

  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$

  for *all* $\mathcal{T}_i$ **do**

    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ w.r.t. K samples

    Compute adapted parameters with gradient descent $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$

  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$ (gradient through gradients, instead of "naive" averaging)

Model-Agnostic Meta Learning – Applications

In deep learning, a very commonly used architecture is as follows:

input $\rightarrow$ $\boxed{\text{CNN (encoder)}}$ $(\rightarrow$ attn$)$ $\rightarrow$ task specific module

tasks can be one or more of

- classification (global pooling + linear)
- sequence labelling (linear)
- segmentation (upsample)
- object detection
- etc.

or many sub-tasks of the above (current main concern for meta-learning).

MAML forces the feature extractor (or called encoder, etc.) to capture general-purpose internal representations (features).

Federated Multi-Task Learning

- Mixture of global and local [3]:

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2} \sum_{i=1}^{N} \|x_i - \bar{x}\|^2$$

- pFedMe (bi-level) [4] (and similarly EASGD[5]):

$$\text{minimize} \quad \sum_{i=1}^{N} F_i(x),$$

$$\text{where} \quad F_i(x) = \min \left\{ f_i(x_i) + \frac{\lambda}{2} \|x_i - x\|^2 \right\}$$

- FedU [6]:

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \|x_i - x_j\|^2$$

pFedMe – Formulation

pFedMe (Personalized Federated Learning with Moreau Envelopes (or proximity operator)) is formulated as the following bi-level optimization problem in [4]

$$\text{minimize} \quad \sum_{i=1}^{N} F_i(x),$$

$$\text{where} \quad F_i(x) = \min \left\{ f_i(x_i) + \frac{\lambda}{2}\|x_i - x\|^2 \right\}$$

which is equivalent to (as in [5])

$$\text{minimize} \quad \sum_{i=1}^{N} \left(f_i(x_i) + \frac{\lambda}{2}\|x_i - x\|^2 \right)$$

pFedMe – Algorithm

Algorithm 2: pFedMe[4]

Input: $T, R, S, \lambda, \eta, \beta, x^0$

for $t = 0, \cdots, T-1$ **do**

Server sends x^t to all clients

for $i = 1, \cdots, N$ **do**

$x^t_{i,0} = x^t$

for $r = 0, \cdots, R-1$ **do**

Sample a fresh mini-batch $\mathcal{D}_i$, and find an approximate $x_i(x^t_{i,r})$ to the problem $\min\{\ell_i(x_i; \mathcal{D}_i) + \frac{\lambda}{2}\|x_i - x^t_{i,r}\|^2\}$

Local update $x^t_{i,r+1} = x^t_{i,r} - \eta\lambda(x^t_{i,r} - x_i(x^t_{i,r}))$

Server uniformly samples a subset of clients $\mathcal{S}^t$, each of which sends the local $x^t_{i,R}$ to the server

Sever update $x^{t+1} = (1-\beta)x^t + \beta \sum \frac{x^t_{i,R}}{\#\mathcal{S}^t}$

pFedMe – Algorithm

pFedMe observations

- global model x converges (if converges) to the average of local models, which can be inferred from

$$x^* = \min_x \{ \sum_{i=1}^{N} \left(f_i(x_i) + \frac{\lambda}{2} \|x_i - x\|^2 \right) \} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- local updates are not "totally local", i.e. the loop $r = 0, \cdots, R$ computes the "global objective" $\min\{F_i(x)\}$ locally, to reduce communication.

EASGD and ADMM

pFedMe originates from EASGD [5], which reformulate the original distributed EM problem $\min_x \mathbb{E}[\ell(x,\xi)]$ as a consensus problem as the following form

$$\text{minimize} \quad \sum_{i=1}^{N} \left(f_i(x_i) + \frac{\lambda}{2} \|x_i - x\|^2 \right)$$

More usually, it is formulated as a minimax (primal-dual) problem to be solved with ADMM

$$\max_{y_i} \min_{x_i, x} \sum_{i=1}^{N} \left(f_i(x_i) + \frac{\lambda}{2} \|x_i - x\|^2 - \langle y_i, x_i - x \rangle \right)$$

Authors of [5] claims that EASGD is more stable than ADMM, which however has to be checked!

References I

[1] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated Multi-Task Learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4427–4437, 2017.

[2] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *International Conference on Machine Learning*, pp. 1126–1135, PMLR, 2017.

[3] F. Hanzely and P. Richtárik, "Federated Learning of a Mixture of Global and Local Models," *arXiv preprint arXiv:2002.05516*, 2020.

[4] C. T. Dinh, N. H. Tran, and T. D. Nguyen, "Personalized Federated Learning with Moreau Envelopes," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, (Red Hook, NY, USA), Curran Associates Inc., 2020.

References II

[5] S. Zhang, A. Choromanska, and Y. LeCun, "Deep Learning with Elastic Averaging SGD," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 685–693, 2015.

[6] C. T. Dinh, T. T. Vu, N. H. Tran, M. N. Dao, and H. Zhang, "FedU: A Unified Framework for Federated Multi-Task Learning with Laplacian Regularization," *arXiv preprint arXiv:2102.07148*, 2021.