Optimization for Deep Learning

Lecture 12-2: Model-Agnostic Meta-Learning (MAML)

Kun Yuan

Peking University

Main contents in this lecture

- Meta learning formulation
- MAML
- Reptile

Meta learning

- Suppose we have a collection of M tasks $\{\mathcal{T}_i\}_{i=1}^{M}$. Each task is associated with a dataset pair $(\mathcal{D}_i^{\text{tr}}, \mathcal{D}_i^{\text{test}})$.

- Let ϕ be the hyper-parameter to learn, which is common to all tasks.

- Let θ_i be the model for task i. Given an specific algorithm $\mathcal{A}lg$, the hyper-parameter ϕ, and the training dataset $\mathcal{D}_i^{\text{tr}}$, θ_i can be learned by

$$\theta_i = \mathcal{A}lg(\phi, \mathcal{D}_i^{\text{tr}}) = \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \mathbb{E}_{\xi_i \sim \mathcal{D}_i^{\text{tr}}} [F(\theta, \phi; \xi_i)] \right\}$$

- The hyper-parameter ϕ can be learned by the mata-learning problem

$$\phi^\star = \underset{\phi \in \mathbb{R}^s}{\arg\min} \left\{ \frac{1}{M} \sum_{i=1}^{M} L(\theta_i, \mathcal{D}_i^{\text{test}}) \right\} \text{ where } \theta_i = \mathcal{A}lg(\phi, \mathcal{D}_i^{\text{tr}})$$

Model-agnostic meta learning (MAML) (Finn et al., 2017)

- Goal: 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
- In other words, we will learn a good initialization model for new tasks
- Given the learned initialization model, a small number of gradient steps with a small amount of training data will produce good generalization performance

Algorithm development

- We let ϕ be the initialization model
- We consider the **one-step SGD optimizer**

$$\theta_i = \arg\min_{\theta}\{f_i(\theta)\} \quad \text{where} \quad f_i(\theta) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i^{\text{tr}}}[F(\theta; \xi_i)]$$
$$\approx \phi - \gamma \nabla_\theta f_i(\phi)$$

- We regard the model θ_i after one-step SGD update as the the classifier associated with task i.

Algorithm development

- Substituting $\theta_i = \phi - \gamma \nabla_\theta f_i(\phi)$ to the meta learning problem

$$\min_{\phi} \quad \frac{1}{M} \sum_{i=1}^{M} L_i\Big(\phi - \gamma \nabla_\theta f_i(\phi)\Big) \quad \text{where} \quad L_i(\theta_i) := L(\theta_i, \mathcal{D}_i^{\text{test}})$$

- The gradient $\nabla L_i(\cdot)$ is calculated as

$$\begin{aligned} & \nabla_\theta L_i\Big(\phi - \gamma \nabla_\theta f_i(\phi)\Big) \\ = \; & \nabla_\theta L_i(\theta)|_{\theta = \phi - \gamma \nabla_\theta f_i(\phi)} - \gamma \frac{\partial^2 f_i(\phi)}{\partial \phi \partial \theta} \cdot \nabla_\theta L_i(\theta) \\ \approx \; & \nabla_\theta L_i(\theta)|_{\theta = \phi - \gamma \nabla_\theta f_i(\phi)} \end{aligned}$$

  where the second term is omitted due to the computational complexity

First-order MAML algorithm

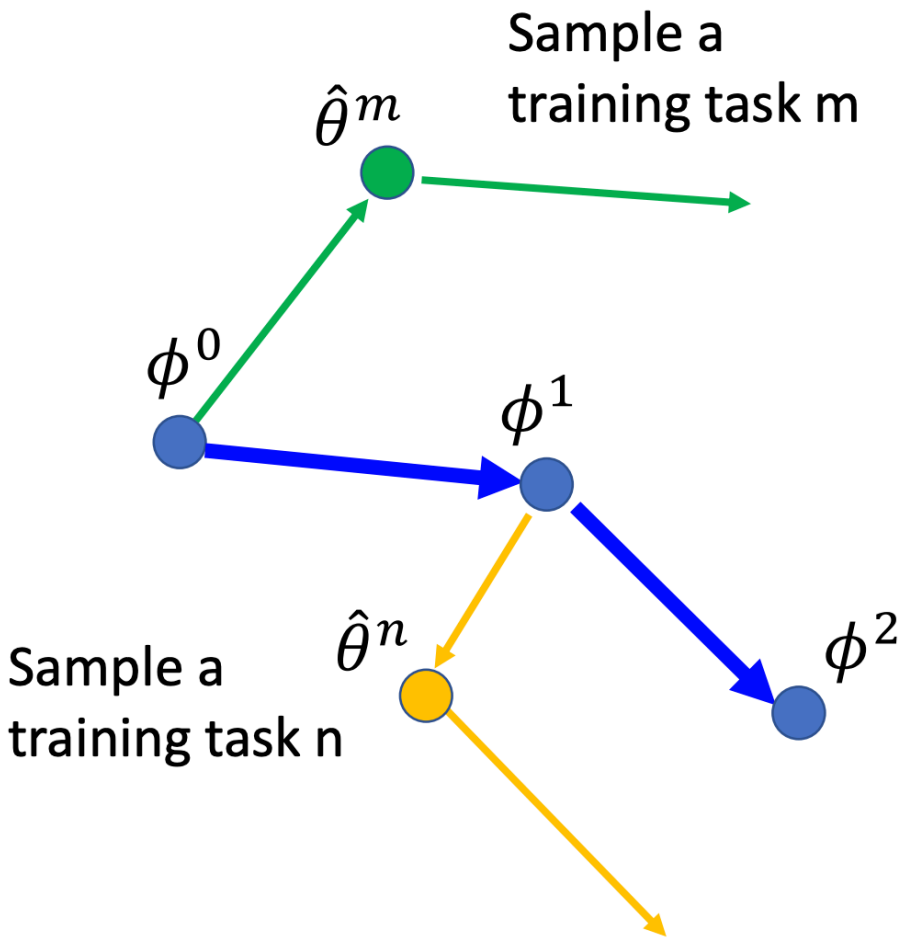
- We sample a task $\mathcal{T}_i$ with data $(\mathcal{D}_i^{\text{tr}}, \mathcal{D}_i^{\text{test}})$ at iteration k

$$\theta_{k+1} = \phi_k - \alpha \nabla F(\phi_k; \xi_i^k), \quad \text{where} \quad \xi_i^k \sim \mathcal{D}_i^{\text{tr}}$$
$$\phi_{k+1} = \phi_k - \beta \nabla L(\theta_{k+1}; \mathcal{D}_i^{\text{test}})$$

  where α and β are two different learning rates

- The one-step SGD optimizer and the removal of the higher-order term from the gradient evaluation are two keys to simplify the algorithm

Illustration of the MAML algorithm

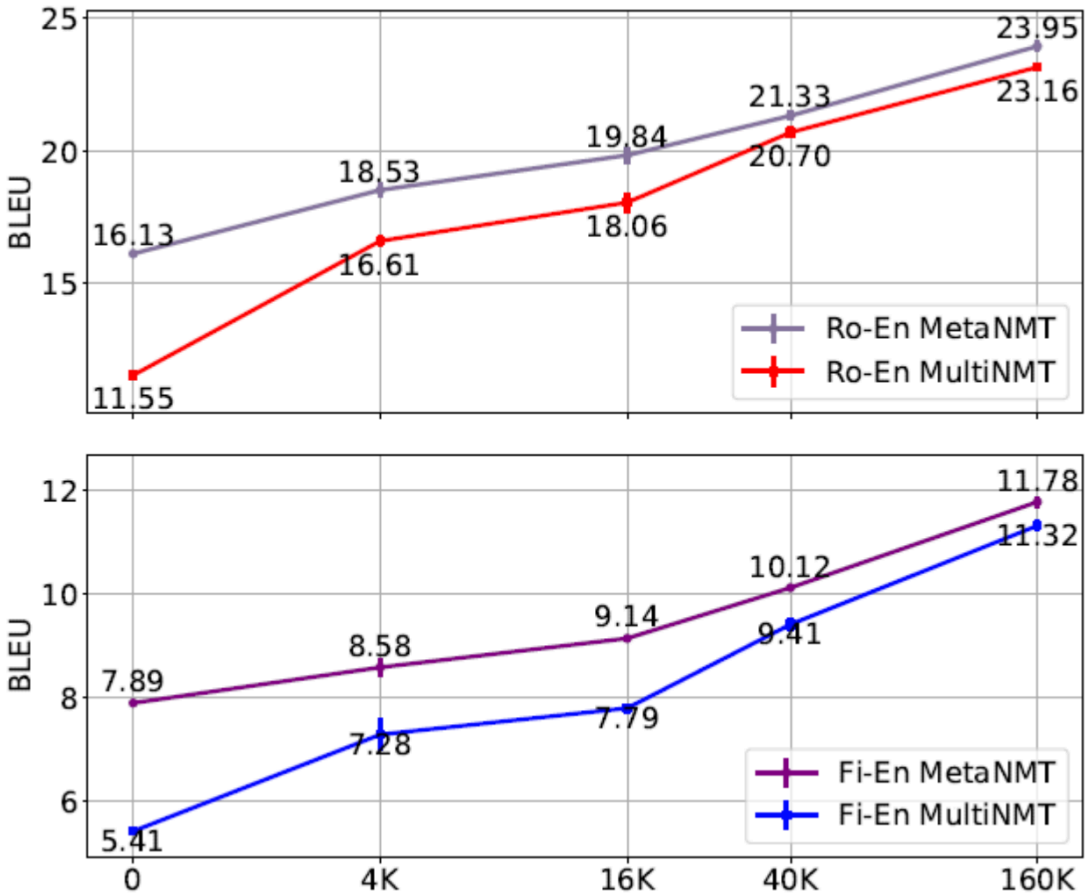


Numerical experiments

Translation

18 training tasks: 18 different languages translating to English

2 validation tasks: 2 different languages translating to English



Ro = Romanian

Fi = Finnish

https://arxiv.org/abs/1808.08437

Multi-step MAML (Antoniou et al., 2018)

- One-step MAML is not stable
- Consider a multi-step SGD optimizer

$$\theta_i^{k+1} = \theta_i^k - \gamma \nabla f_i(\theta_i^k) \quad \text{with} \quad \theta_i^0 = \phi$$

  for $k = 0, 1, \cdots, K-1$. Apparently, each θ_i^k is a function of ϕ

- Keep iterating the above recursion, we have

$$\theta_i^K = \theta_i^0 - \gamma \sum_{k=0}^{K-1} \nabla f_i(\theta_i^k) \quad \text{with} \quad \theta_i^0 = \phi$$

Multi-step MAML

- Substituting $\theta_i^K = \theta_i^0 - \gamma \sum_{k=0}^{K-1} \nabla f_i(\theta_i^k)$ to meta learning, we have

$$\min_{\phi} \quad \frac{1}{M} \sum_{i=1}^{M} L_i(\theta_i^0 - \gamma \sum_{k=0}^{K-1} \nabla f_i(\theta_i^k))$$

$$\text{where} \quad L_i(\theta_i) := L(\theta_i, \mathcal{D}_i^{\text{test}}) \quad \text{and} \quad \theta_i^0 = \phi$$

- If we ignore all higher-order term, the gradient can be approximated as

$$\nabla_\theta L_i(\theta_i^K) \approx \nabla_\theta L_i \left(\theta_i^0 - \gamma \sum_{k=0}^{K-1} \nabla f_i(\theta_i^k) \right)$$

Multi-step MAML

- We sample a task $\mathcal{T}_i$ with data $(\mathcal{D}_i^{\text{tr}}, \mathcal{D}_i^{\text{test}})$ at iteration t

$$\begin{aligned}
\theta_i^{t,0} &= \phi^t \\
\theta_i^{t,k+1} &= \theta_i^{t,k} - \alpha \nabla F(\theta_i^{t,k}; \xi_i^{t,k}) \text{ for } k = 0, \cdots, K-1 \\
\phi^{t+1} &= \phi^t - \beta \nabla L(\theta^{t,K}; \mathcal{D}_i^{\text{test}})
\end{aligned}$$

  where $\xi_i^{t,k} \sim \mathcal{D}_i^{\text{tr}}$.

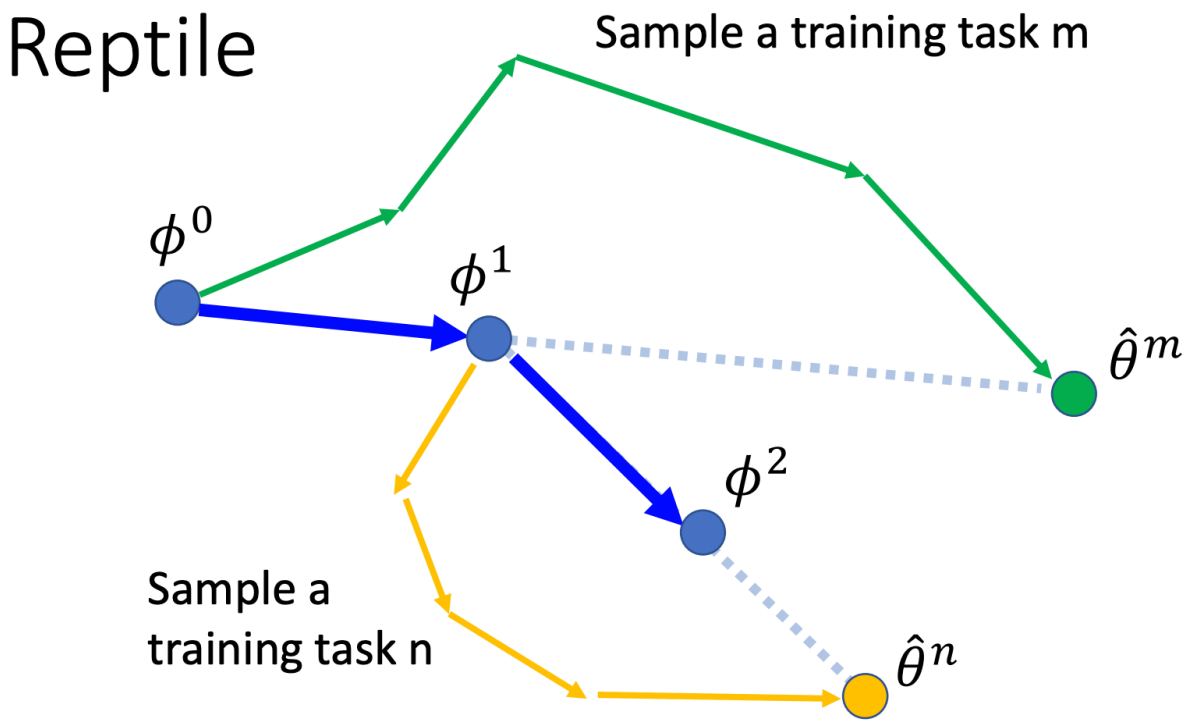
Reptile (Nichol et al., 2018)

- We sample a task $\mathcal{T}_i$ with data $(\mathcal{D}_i^{\text{tr}}, \mathcal{D}_i^{\text{test}})$ at iteration t

$$\begin{aligned}
\theta_i^{t,0} &= \phi^t \\
\theta_i^{t,k+1} &= \theta_i^{t,k} - \alpha \nabla F(\theta_i^{t,k}; \xi_i^{t,k}) \text{ for } k = 0, \cdots, K-1 \\
\phi^{t+1} &= \phi^t - \beta(\theta_i^{t,K} - \theta_i^{t,0})
\end{aligned}$$

  where $\xi_i^{t,k} \sim \mathcal{D}_i^{\text{tr}}$.

Reptile


Reptile
Sample a training task m
ϕ^0
ϕ^1
$\hat{\theta}^m$
ϕ^2
Sample a
training task n
$\hat{\theta}^n$

References I

C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," *arXiv preprint arXiv:1810.09502*, 2018.

A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.