Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

Model-Agnostic Meta-learning

慕星星　硕士研究生

2020年3月8日

内容提要

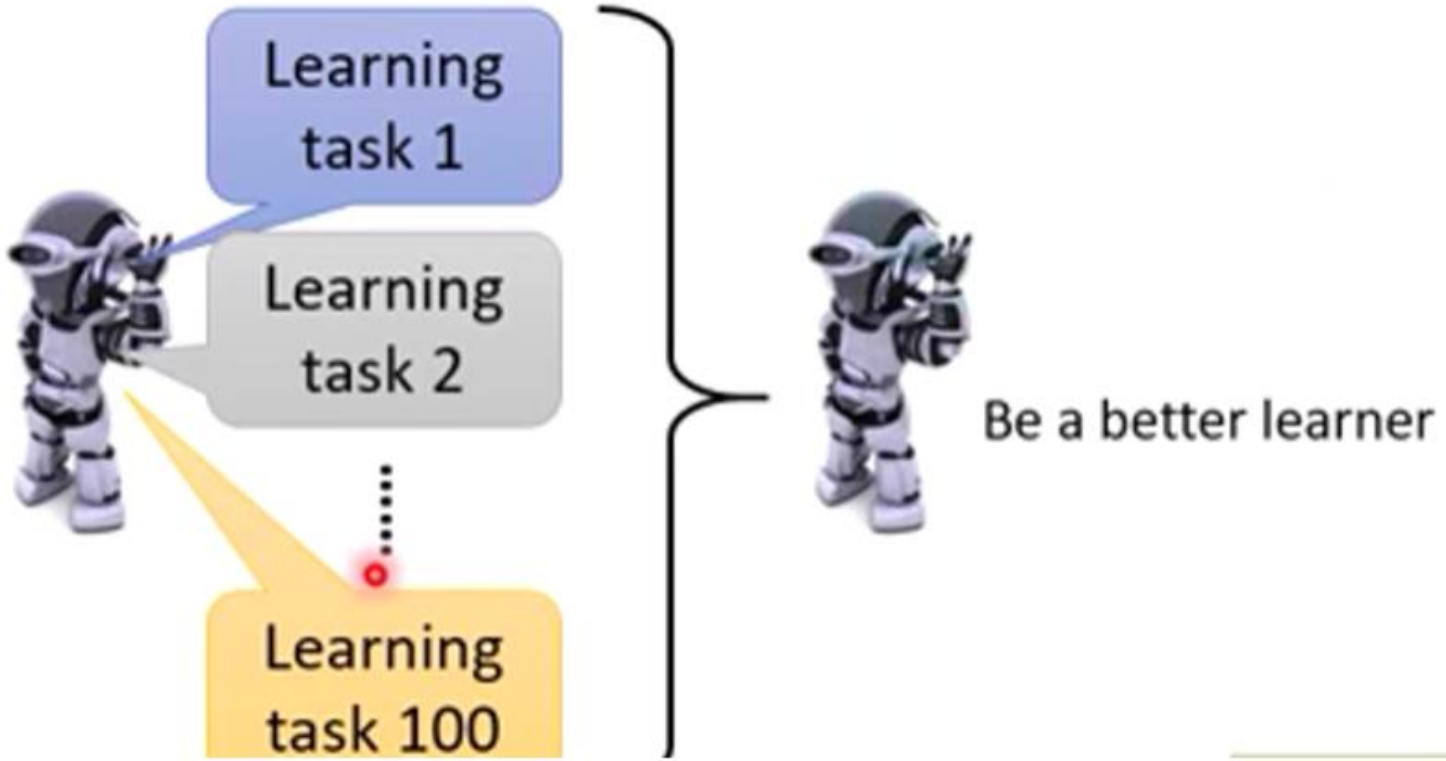
- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用介绍
- 参考文献

预期收获

- 预期收获
  - 1. 熟悉元学习的基本思想
  - 2. 理解MAML的算法原理
  - 3. 了解MAML的应用

背景简介

- Meta Learning = Learn to learn
  - 让机器去学习如何进行学习：指通过学习一系列相似任务，归纳（抽象）出这些任务的本质规律（通用属性——权重/超参）。当面对全新的任务时，可以根据学得的规律，做微调，便可快速适应。



背景简介

- Meta Learning 的思路：
  - Learning good weight initializations
    - 通过使用超参数梯度下降，网络从任务的全部分布中学习到有用的表征。
  - 基于度量 metric-based
    - 学习核变换参数，更好地表示数据
  - Learning transferbale optimizers
    - 学习内部优化器的网络，使用梯度下降更新神经优化器网络参数，使得网络在整个任务中获得很好的表现

背景简介

- Meta Learning VS Lifelong Learning
  - 终身学习：着眼于用同一个模型去学习不同的任务
  - 元学习：不同任务使用不同的模型，元学习积累经验后，在新任务上训练的更快更好

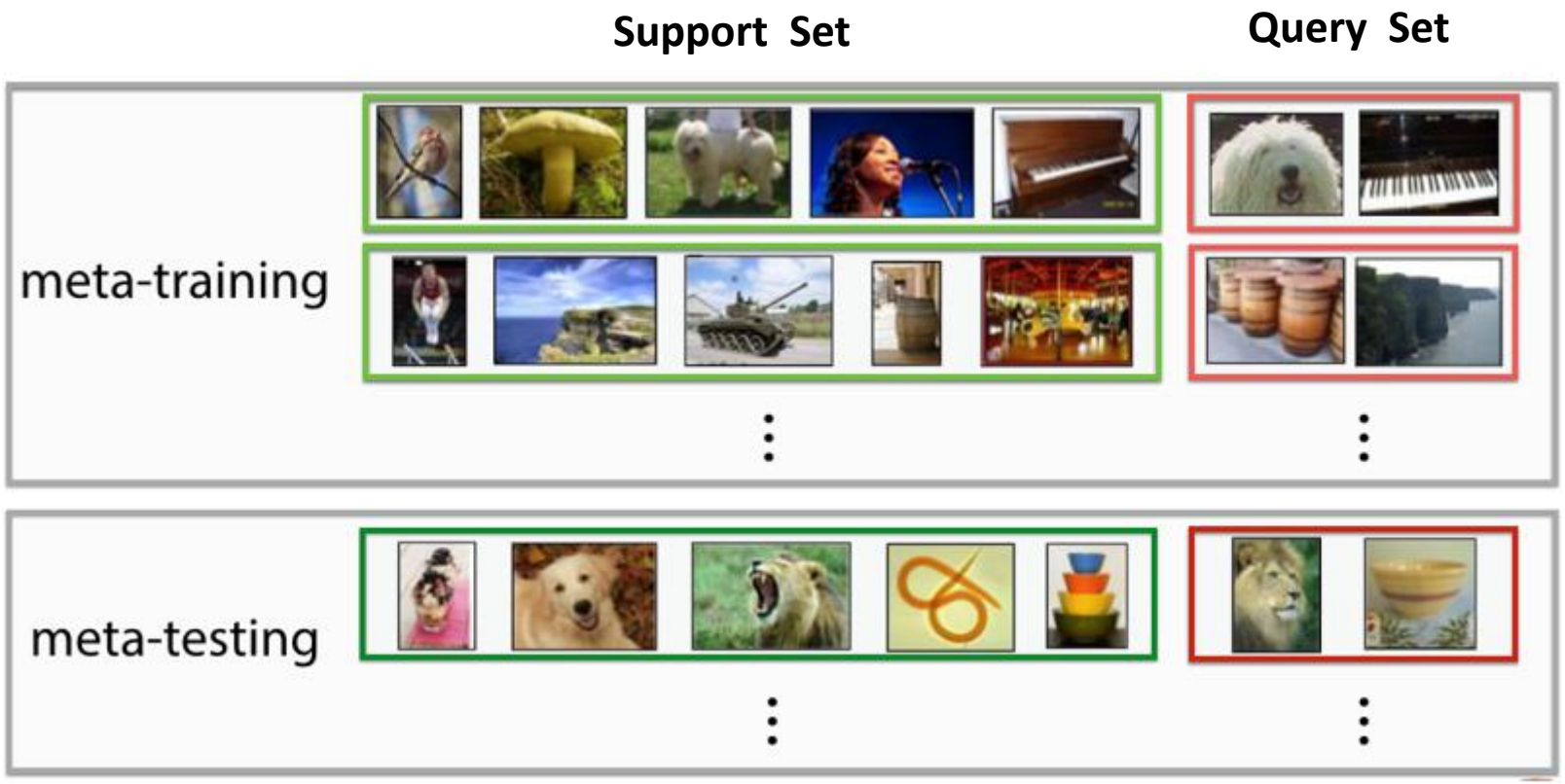
- Meta Learning VS Machine Learning
  - 机器学习：核心是通过人为设计的学习算法，利用训练数据训练得到一个函数f，这个函数可以用于新数据的预测分类。
  - 元学习：让机器自己学习找出最优的学习算法。根据提供的训练数据找到一个可以找到函数f的函数F的能力。

背景概念

- 元学习架构
  - 元训练
    - 用训练样本在子任务上进行训练，得到相应的损失，并对该任务的模型参数进行梯度更新，在新的数据样本上测试更新后的网络，得到错误情况。
  - 元测试
    - 根据各个子任务更新后的网络的表现求初始化参数的梯度，并对元学习模型的参数进行更新，测试其在元测试集任务上的表现，即为元学习模型的最终表现。

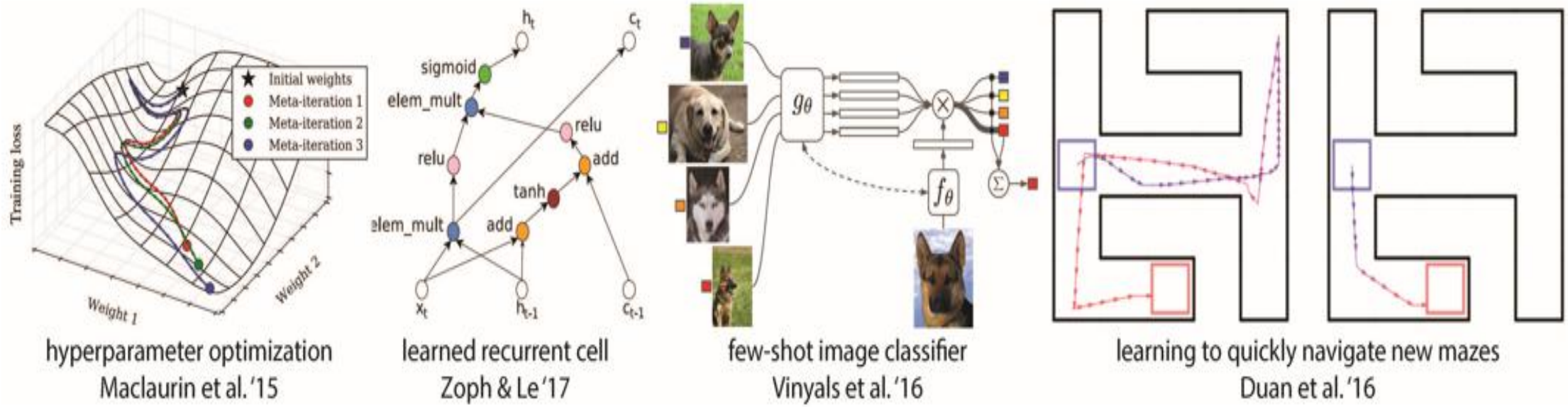
背景概念

- 元学习的训练数据
  - 元学习的训练数据是由一个个的训练任务构成的，一个训练任务对应一个传统的机器学习的应用实例。训练数据分为训练任务集和测试任务集。



基本概念

- 元学习通常被用在：优化超参数和神经网络、探索好的网络结构、小样本图像识别和快速强化学习等。



hyperparameter optimization
Maclaurin et al.'15

learned recurrent cell
Zoph & Le'17

few-shot image classifier
Vinyals et al.'16

learning to quickly navigate new mazes
Duan et al.'16

基本概念

模型无关（model-agnostic）

- model-agnostic即模型无关。
  - MAML更像是一个框架，提供一个meta-learner用于训练base-learner。Meta-learning即MAML的精髓所在，用于learning to learn；而base-learner则是在目标数据集上被训练，并实际用于预测任务的真正的数学模型。绝大多数深度学习模型都可以作为base-learner无缝嵌入MAML中，MAML甚至可以用于强化学习，这就是MAML中model-agnostic的意义。

小样本学习（Few-shot Learning）

- few-shot learning——小样本学习
  - 指通过极少的样本学习获得（监督/非监督）回归、分类模型。在现有的研究成果中，小样本学习可以基于fine-tune、metric(如孪生网络)、基于meta-learning等。在基于meta-learning的少样本学习中，已有memory-augmented neural networks (Santoro et al., 2016)、meta-learner LSTM (Ravi & Larochelle, 2017)等经典学习方法。

数据集

- Omniglot数据集
  - 组成：
    - 整个数据集由1623个符号（Characters)组成；
    - 每个符号有20个样例（Examples),每个样例由不同的人书写。
  - 使用：结合Few-shot Learning中的N-ways K-shot分类问题
    - 对于每一个训练任务和测试任务，样本数据分为N个类，每个类提供K个样本。
    - 整个字符集分为训练字符集(Training Set or Support Set)和测试字符集（Testing Set or Query Set）
    - 训练任务：从训练字符集中抽取N个类的字符，每种字符抽取K个样本，组成一个训练任务的训练数据
    - 测试任务：从测试字符集中抽取N个类的字符，每种字符抽取K个样本，组成一个测试任务的训练数据

算法原理

算法原理

T	得到仅使用少量的训练样本就可以在新任务上快速收敛的模型
I	多个训练任务task
P	两次梯度下降
O	能够快速学习解决只含有少量训练样本的新任务的模型

P	如何训练出合适的模型初始参数，使得在小规模的训练样本上迅速收敛
C	模型无关，适用于任何一种采用梯度下降算法的模型
D	二次梯度可能不稳定
L	ICML 2017

算法原理

- MAML算法伪代码

Algorithm 1 Model-Agnostic Meta-Learning

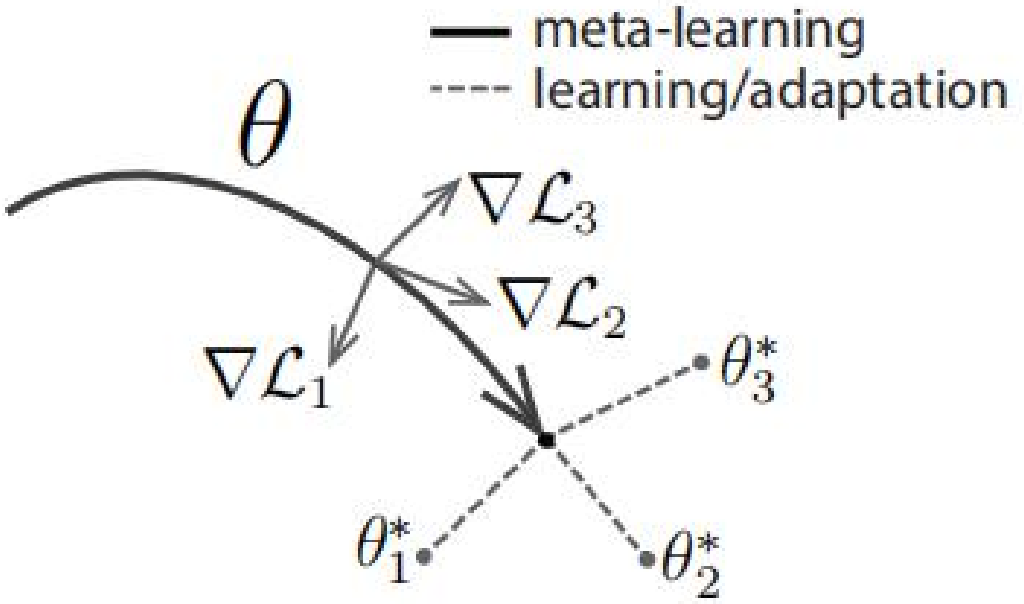
Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

算法原理

- 整个算法分为两个循环：
  - 两个循环共享模型参数 θ
  - 两个循环的梯度更新的学习率分别由超参数α和β表示
  - 内循环计算各个子任务的损失 ζ_{t_i}和进行一至多次梯度更新后的参数 θ_i'
  - 外循环根据内循环的优化参数在新任务上重新计算损失，并计算其对初始参数的梯度，然后对初始参数进行梯度更新
  - 重复内外循环，就可以得到元学习模型对于任务分布 $p(t)$的最佳参数

算法原理

- θ是已经优化过的模型参数表示
- 当θ沿着新任务损失梯度方向变化时，会使得任务损失大幅改善，从而得到对于新任务的最佳模型参数θ^*



算法原理

- 损失函数（Loss Function）：$L(\theta)=\sum_{n=1}^{N}\varsigma_{t_i}\left(f_{\theta_i'}\right)$
  - θ_i'：第i个任务中学习到的模型参数，取决于参数 θ
  - $\varsigma_{t_i}\left(f_{\theta_i'}\right)$：第i个任务在其测试集上得到的损失

- 损失函数最小化：使用梯度下降(Gradient Descent)

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \mathrm{L}(\theta)$$

算法原理

- 只考虑一次训练之后对初始化参数的梯度更新：
  - 只取进行一次梯度更新后的参数作为当前任务的最佳参数。
  - θ 求出的是元学习模型的通用参数，$\theta_i^{'}$ 求出的是每个任务的最佳参数。
  - $\mathrm{L}(\theta)$ 和$\theta_i^{'}$用于元学习模型的参数更新
  - 既能加快模型的适应速度，在一定程度上还能减轻过拟合。

$$\theta_i^{'} = \theta - \alpha \nabla_\theta \mathrm{L}(\theta)$$

算法原理

- 梯度的计算是需要确定loss function的，MAML中loss根据不同的问题处理有不同的选择：
  - 对于可监督回归问题，采用均方差（mean-squared error）

$$\mathcal{L}_{\mathcal{T}_i}(f_\phi) = \sum_{x^{(j)}, y^{(j)} \sim \mathcal{T}_i} \| f_\phi(x^{(j)}) - y^{(j)} \|_2^2$$

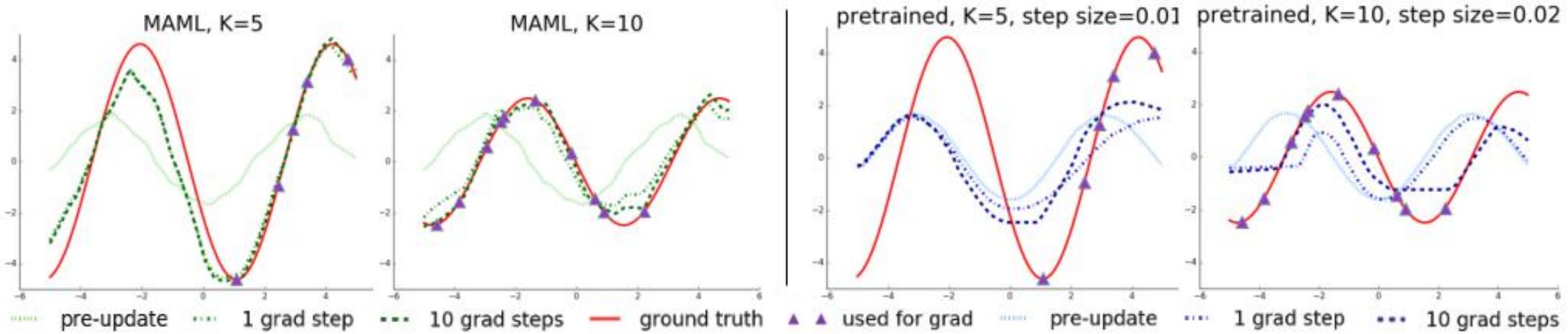
  - 对于可监督分类问题，采用交叉熵(cross entropy)

$$\mathcal{L}_{\mathcal{T}_i}(f_\phi) = \sum_{x^{(j)}, y^{(j)} \sim \mathcal{T}_i} y^{(j)} \log f_\phi(x^{(j)}) + (1 - y^{(j)}) \log(1 - f_\phi(x^{(j)}))$$

  - 对于强化学习问题，MAML将每一个task视为一个马尔可夫决策过程（MDP）。

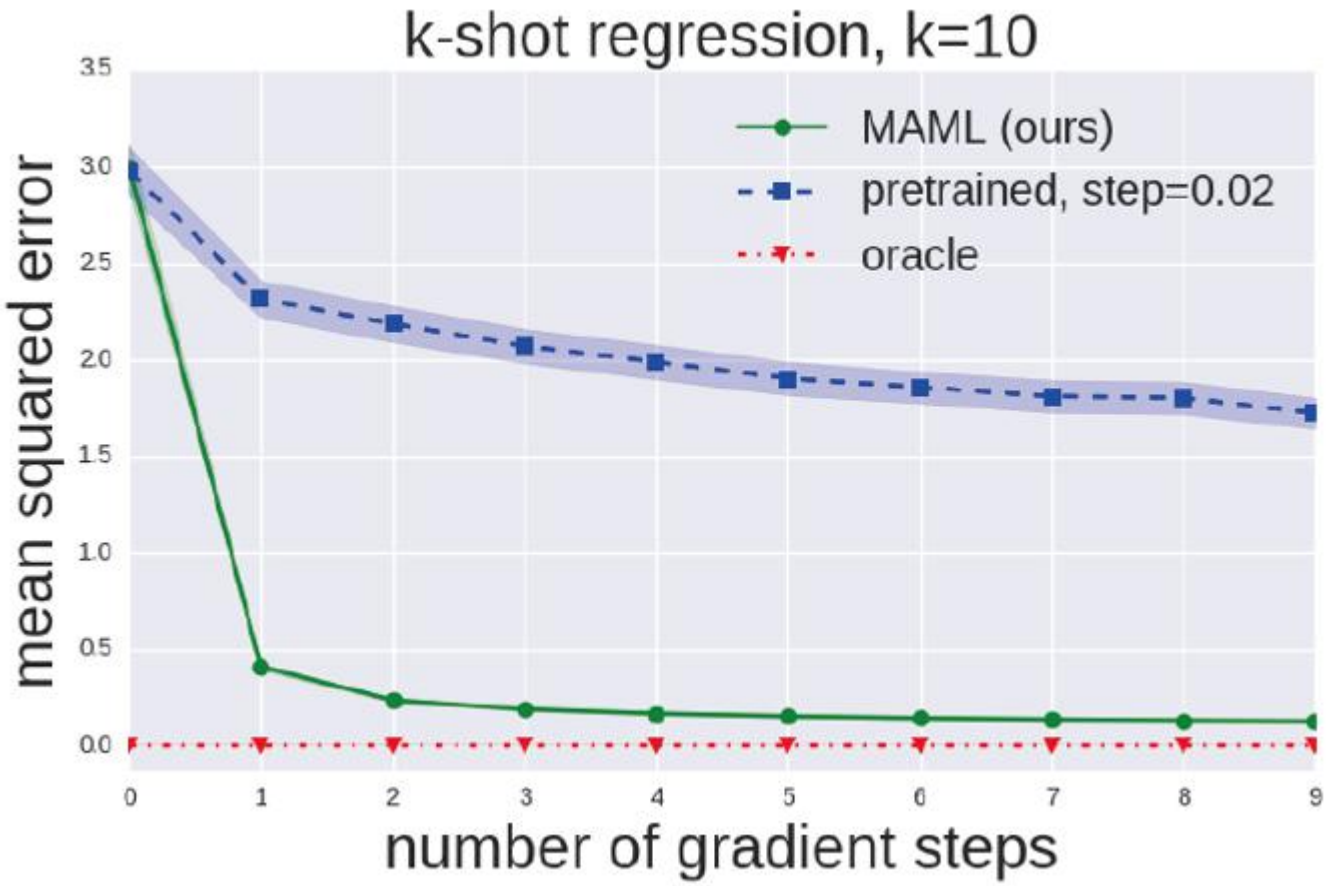
算法原理

- 算法执行结果
  - 回归任务（正弦曲线）
    - 通过将MAML算法模型与预训练模型比较，分别提供K=5和K=10个样本数据，进行回归拟合。



算法原理

- 比较MAML和预训练模型的学习曲线



算法原理

- 算法执行结果
  - 分类任务
    - MAML和简化后的FOMAML模型与用于Few-shot Learning分类的主流模型在Omniglot和MiniImage数据集上比较。

Omniglot (Lake et al., 2011)	5-way Accuracy 1-shot	5-way Accuracy 5-shot	20-way Accuracy 1-shot	20-way Accuracy 5-shot
MANN, no conv (Santoro et al., 2016)	82.8%	94.9%	–	–
MAML, no conv (ours)	**89.7 ± 1.1%**	**97.5 ± 0.6%**	–	–
Siamese nets (Koch, 2015)	97.3%	98.4%	88.2%	97.0%
matching nets (Vinyals et al., 2016)	98.1%	98.9%	93.8%	98.5%
neural statistician (Edwards & Storkey, 2017)	98.1%	99.5%	93.2%	98.1%
memory mod. (Kaiser et al., 2017)	98.4%	99.6%	95.0%	98.6%
MAML (ours)	**98.7 ± 0.4%**	**99.9 ± 0.1%**	**95.8 ± 0.3%**	**98.9 ± 0.2%**

MiniImagenet (Ravi & Larochelle, 2017)	5-way Accuracy 1-shot	5-way Accuracy 5-shot
fine-tuning baseline	28.86 ± 0.54%	49.79 ± 0.79%
nearest neighbor baseline	41.08 ± 0.70%	51.04 ± 0.65%
matching nets (Vinyals et al., 2016)	43.56 ± 0.84%	55.31 ± 0.73%
meta-learner LSTM (Ravi & Larochelle, 2017)	43.44 ± 0.77%	60.60 ± 0.71%
MAML, first order approx. (ours)	**48.07 ± 1.75%**	**63.15 ± 0.91%**
MAML (ours)	**48.70 ± 1.84%**	**63.11 ± 0.92%**

优劣分析

- 优点
  - MAML不会大量增加学习的参数（其实理论上只增加一个参数——元学习率）
  - MAML不会对模型进行任何限制，只限制模型训练的方法是梯度下降

- 缺点：
  - 存在元过拟合问题，因为所有task(训练及测试）需要来自同一个分布，当分布选择不合理，或者采样时多样性不足会导致过拟合问题。
  - 虽然可以解决少样本问题，但是训练所需的样本量的大小没有固定的范围，需要通过试验确定。

应用介绍

- 新神经机器翻译方法——MetaNMT
  - MetaNMT算法就是将元学习算法（MAML），用于低资源神经机器翻译（NMT）中。将翻译问题建构为元学习问题，从而解决低资源语言语料匮乏的难题。在低资源神经机器翻译（NMT）上的有着优异的性能
  - 研究人员先使用许多高资源语言（比如英语和法语），训练出了一个表现极佳的初始参数，然后构建一个所有语言的词汇表。再以初始参数/模型为基础，训练低资源语言的翻译（比如英语VS希伯来语，法语VS希伯来语）。

参考文献

- [1] Finn C,Abbeel P,Levine S.Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70.JMLR.org,2017:1126-1135

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒
胜热。清静为天下正。

谢谢！