Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

Chelsea Finn, Pieter Abbeel, Sergey Levine, 2017, ICML

Wonhee Cho

Vision and Learning Laboratory
School of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea
Emails : wonhee4274@cau.ac.kr

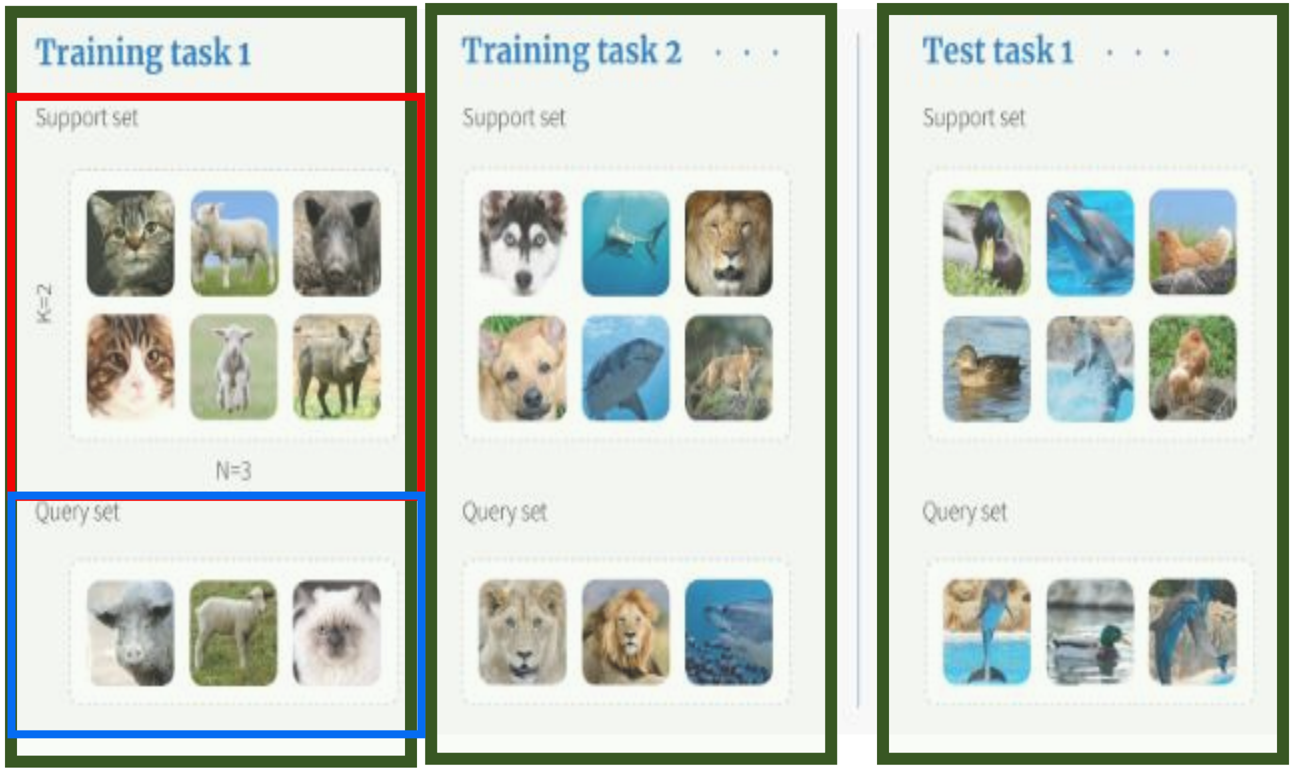
Index

Few-shot Learning

Chelsea Finn, 2019, ICML Tutorial.

Few-shot Learning(FSL)

- Few-shot learning is to classify new data having seen only a few training examples.
- Few-shot learning is useful when training examples are hard to find (e.g., cases of a rare disease), or where the cost of labelling data is high.



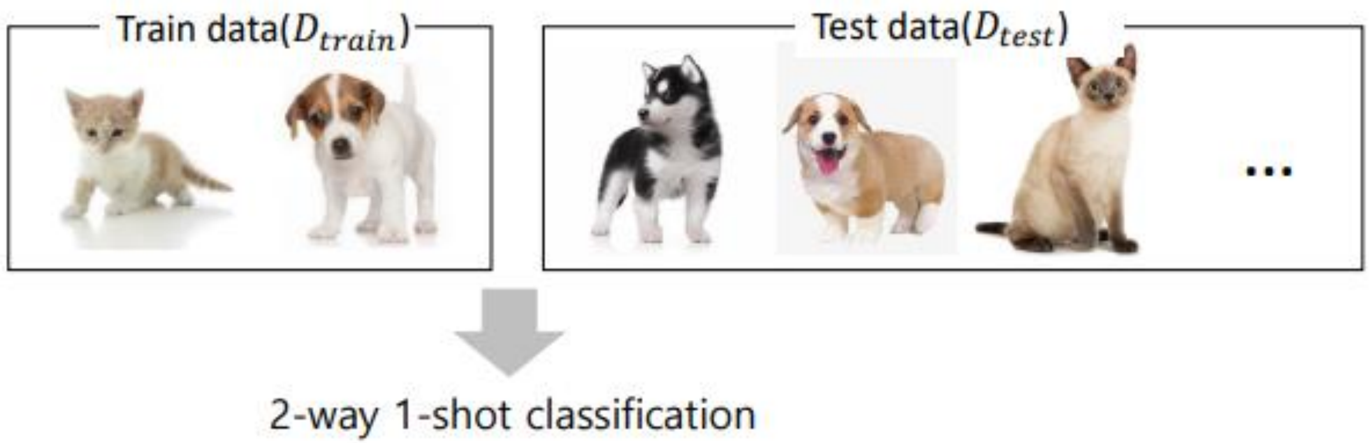
N-way K-shot classification

https://www.borealisai.com/en/blog/tutorial-2-few-shot-learning-and-meta-learning-i/

Few-shot Learning

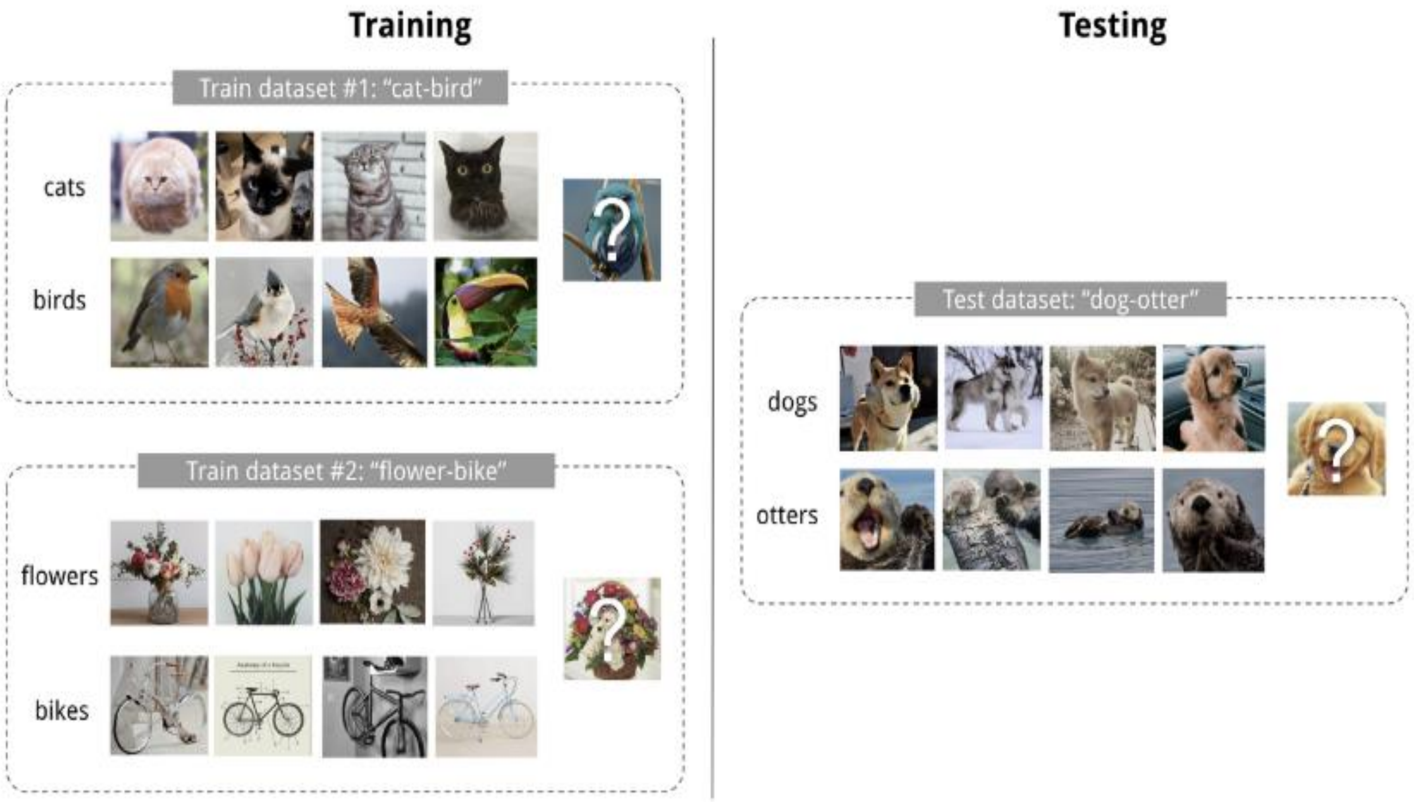
N-way K-shot

- Classes : N
- Examples: K



Meta Learning: Learning to learn

- In the meta-learning framework, we *learn how to learn* to classify given a set of *training tasks* and evaluate using a set of t*est tasks.*
- In other words, we use one set of classification problems to help solve other unrelated sets.



Optimal model parameter

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}[\mathcal{L}_\theta(\mathcal{D})]$$

Each task consists of a dataset D.

https://www.borealisai.com/en/blog/tutorial-2-few-shot-learning-and-meta-learning-i/

Meta Learning vs Multi-task Learning

Multi-task Learning standpoint

- Optimal parameters $\emptyset_{\mathrm{i}}$ for each Task($T1$, $T2$, …) are same.
- 하나의 파라미터를 공유하는 하나의 큰 모델이 모든 task를 해결함.

Meta Learning standpoint

- Optimal parameters $\emptyset_{\mathrm{i}}$ for each Task($T1$, $T2$, …) are different.
- 데이터 특성과 $\emptyset_{\mathrm{i}}$ 사이의 정보(θ)를 학습하고 추후 새로운 데이터에 대해 θ를 이용.

Model-Agnostic Meta-Learning

Problem Definition

- Model f parameterized by θ
  - $\mathrm{a} = \mathrm{f}(\mathrm{x})$: mapping function
  - P(T): tasks distribution
  - $T = \{L(x_1, a_1, \ldots, x_H, a_H), q(x_1), q(x_{t+1}|x_t, a_t), H\}$
    - Supervised learning: H=1
    - K-shot learning: K samples drawn from $\mathrm{q_i}$

L : loss function
$\mathrm{q}(\mathrm{x}_1)$: a distribution over initial observations
$\mathrm{q}(\mathrm{x}_{t+1}|\mathrm{x}_t, \mathrm{a}_t)$: a transition distribution

Model-Agnostic Meta-Learning

- Method
  - For task T_I model's parameter θ become

Fixed as a hyperparameter or meta-learned

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

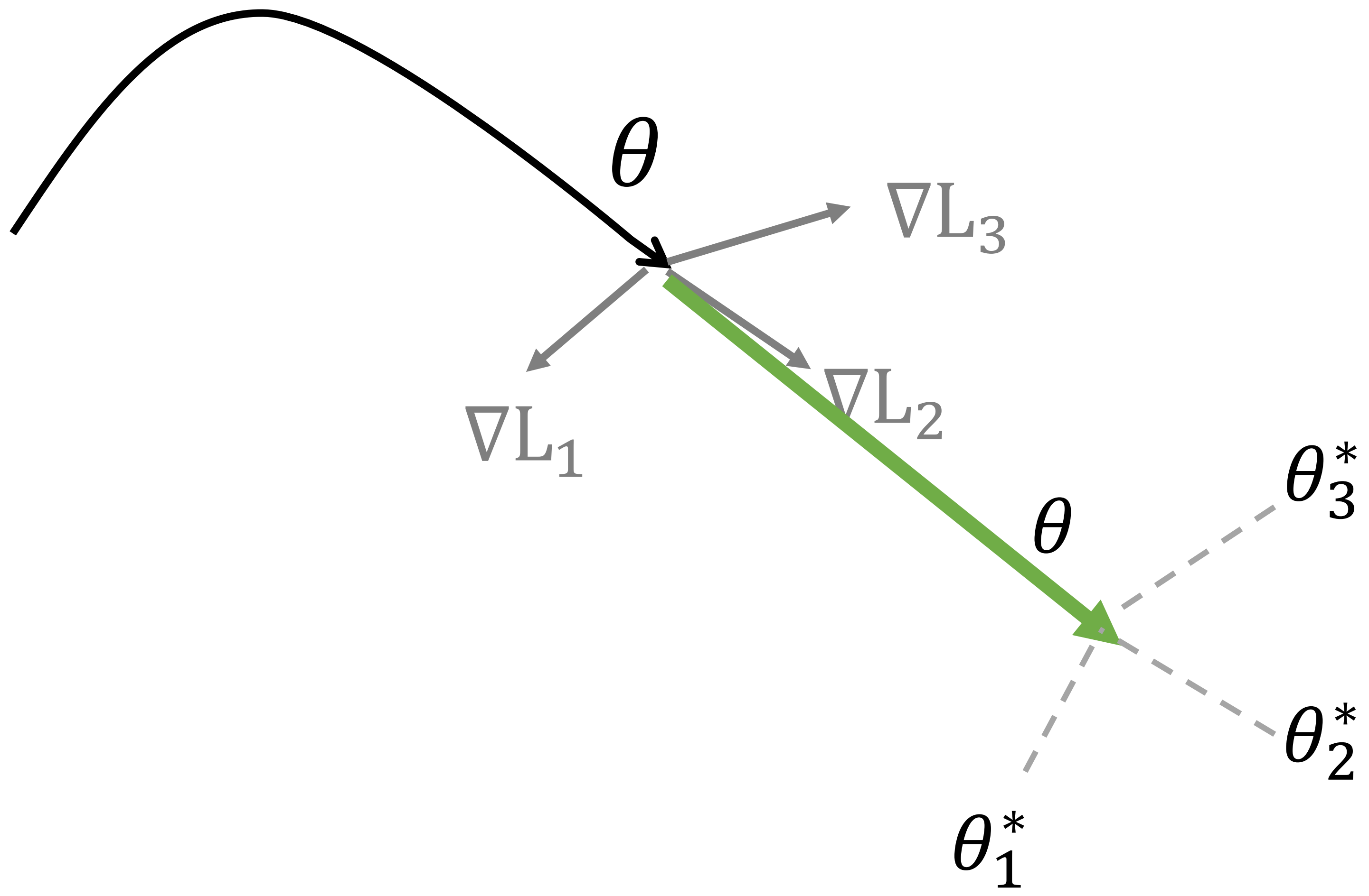
  - Multiple gradient update also is extendable

- Meta-objective

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)})$$

Model-Agnostic Meta-Learning

Intuition



Model-Agnostic Meta-Learning

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters

randomly initialize θ
while not done **do**
  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
  for all $\mathcal{T}_i$ **do**
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
    Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
  end for
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
end while

Model-Agnostic Meta-Learning

- Regression
  - Few-shot Regression: the goal is to predict the outputs of a continuous-valued function from only a few datapoints sampled from that function.

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

  - Using mean-squared error(MSE)
- Classification
  - Using cross-entropy loss

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

Experimental Evaluation

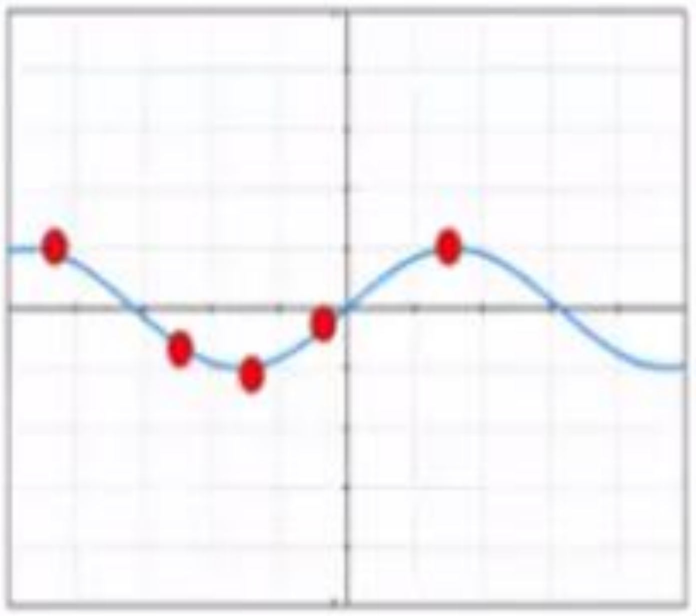
1) Can MAML enable fast learning of new tasks?

2) Can MAML be used for meta-learning in multiple different domains, including supervised regression, classification, and reinforcement learning?

3) Can a model learned with MAML continue to improve with additional gradient updates and/or examples?

Experiments_Regression

- Sine wave experiments
  - Meta Training (700000)
    - Amplitude[0.1, 5.0]
    - Phase[0, π]
    - K points sampled from [-0.5, 5.0]
  - Meta Testing
    - K samples from a sine wave
  - Evaluation
    - Mean squared error for 600 points



θ_i'

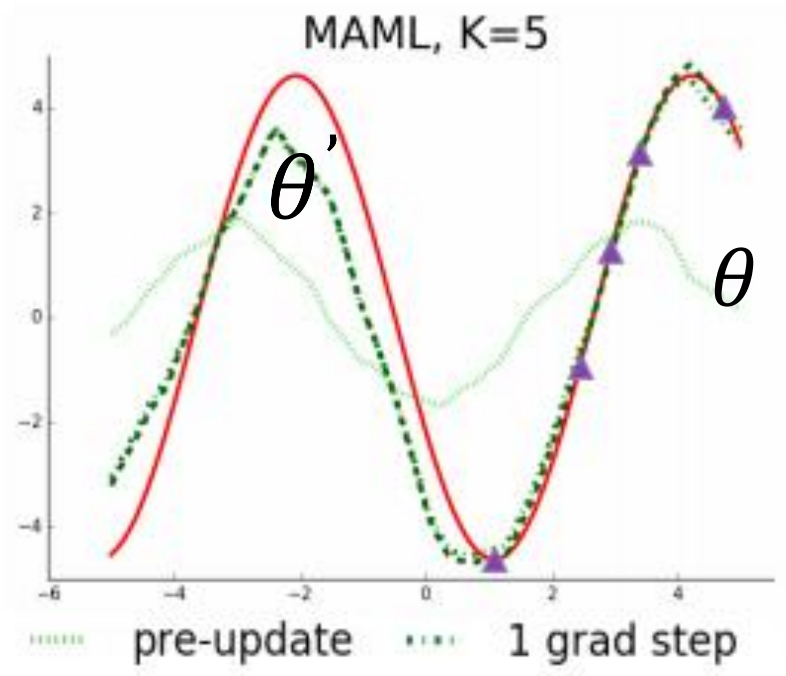
Experiments_Regression

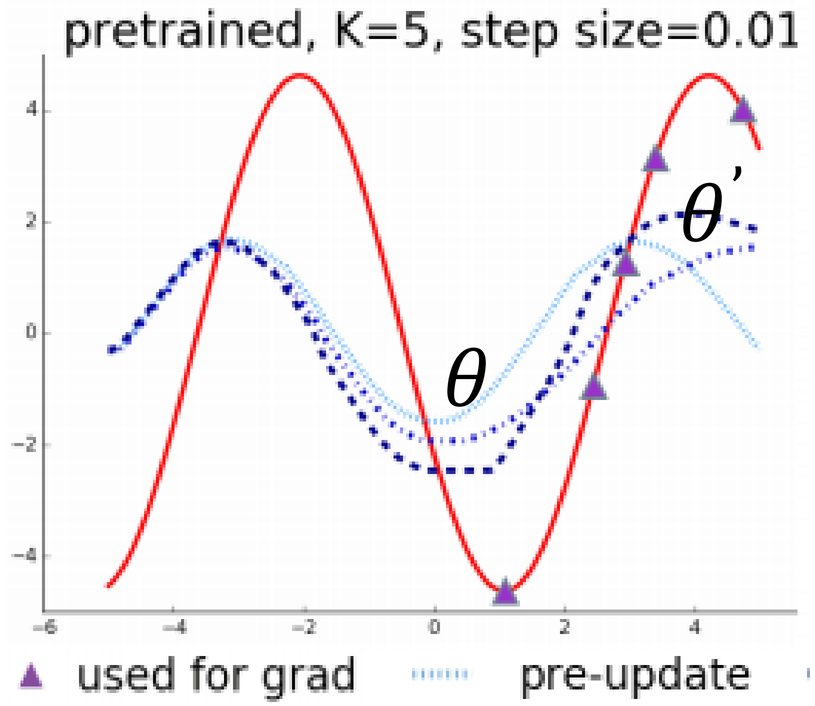
- Sine wave experiments
  - Meta Training (700000)
    - Amplitude[0.1, 5.0]
    - Phase[0, π]
    - K points sampled from [-0.5, 5.0]

  - Meta Testing
    - K samples from a sine wave

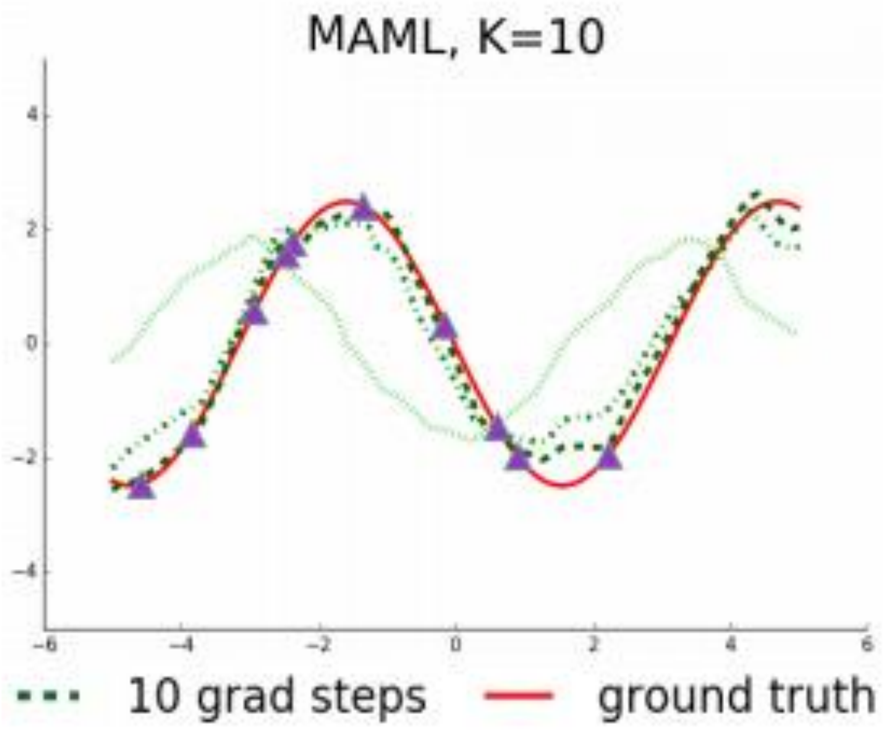
  - Evaluation
    - Mean squared error for 600 points

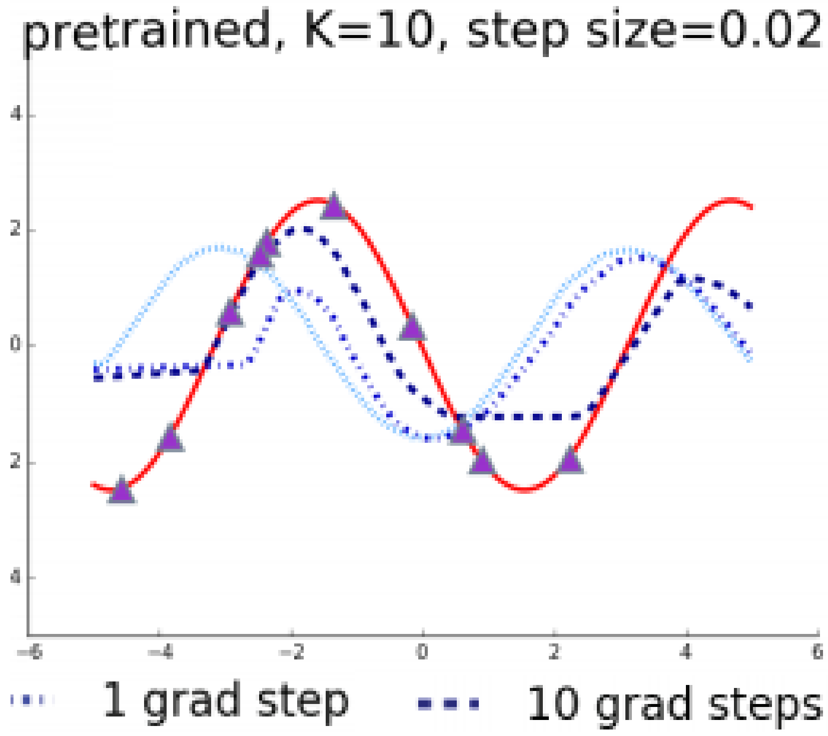
Experiments_Regression





Experiments_Regression





Experiments_Regression

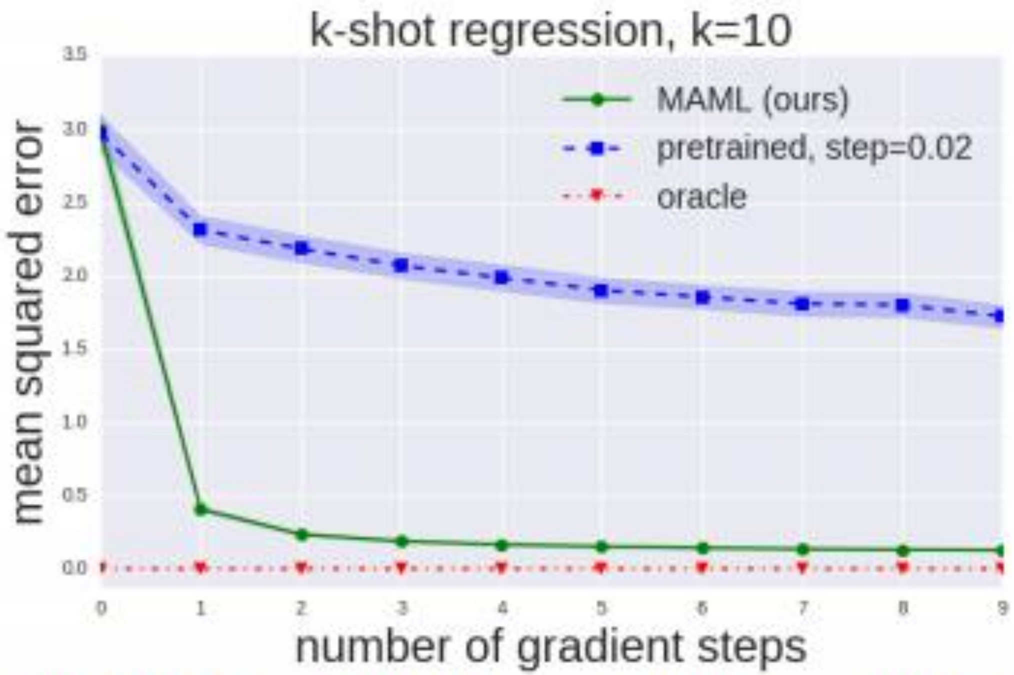


Figure 3. Quantitative sinusoid regression results showing the learning curve at meta test-time. Note that MAML continues to improve with additional gradient steps without overfitting to the extremely small dataset during meta-testing, achieving a loss that is substantially lower than the baseline fine-tuning approach.

Experiments_Classification

- N-way classification
  - Use N class during test with K-shot learning
- Network Architecture
  - 4 modules
    - 3 x 3 convolutions and 64 filters
    - ReLU nonlinearity
    - 2 x 2 max-pooling

Dataset

- Omniglot
  - 1623 characters from 50 alphabets
    - 20 instances each drawn by a different person
  - Training
    - 1200 characters
  - Testing
    - 423 characters

- Mini-ImageNet
  - 80 training classes
  - 20 test classes

Experiments_Classification

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%**

Q & A

Thank You