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

2019. 07. 19

Data Mining & Quality Analytics Lab.

목 충 협

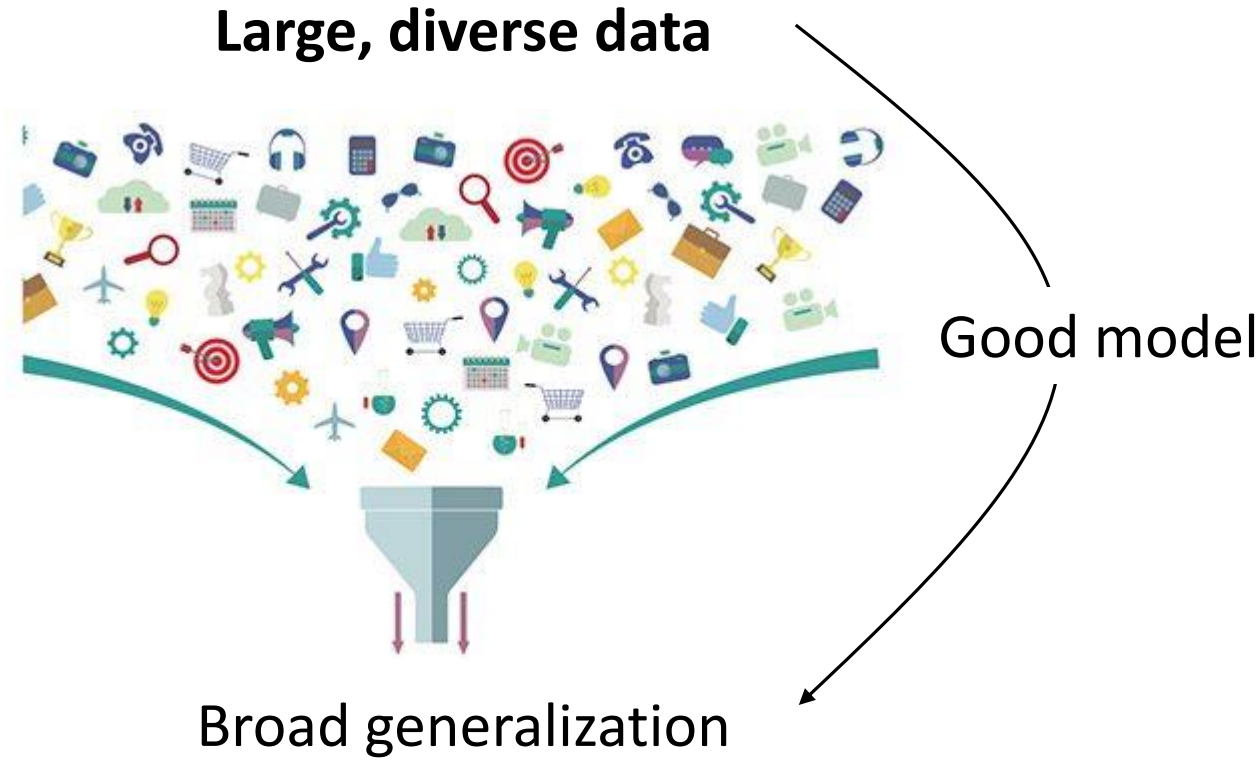
Data Mining
Quality Analytics

목차

- Introduction
- Problem Description
- Meta-Learning Algorithms
- Experiments
- Conclusions

Introduction

- 일반적인 모델 학습의 가정



Introduction

- 하지만 현실은...

필요한 데이터가 부족

클래스가 불균형

레이블 구분이 모호

데이터에 대한 가정을 충족하지 못함

아무리 좋은 모델을 사용해도 학습이 어려움!

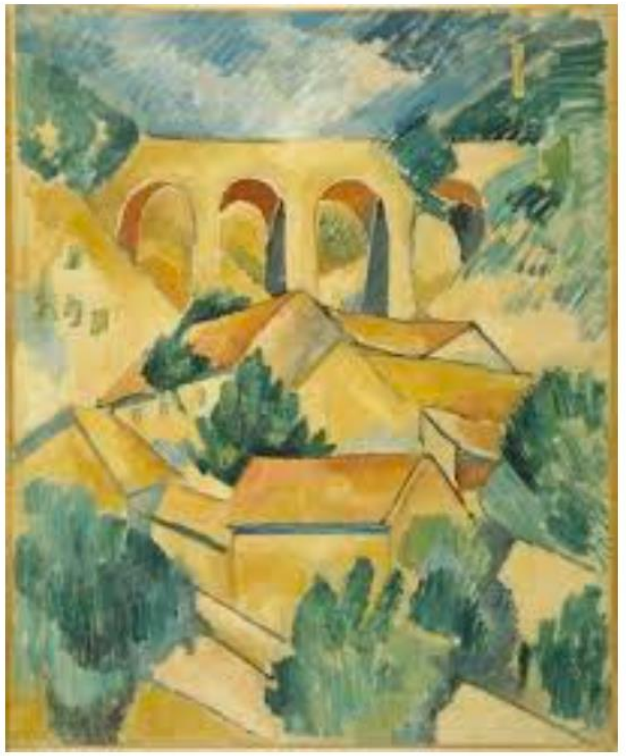
Introduction

- Few-Shot Classification 예제

Train data

Braque | Cezanne

Test data point



By Braque or Cezanne?

Chelsea Finn, ICML 2019 Tutorial

Introduction

데이터에 대한 가정을 충족하지 않은 상황인데 어떻게 알 수 있었을까?

Braque 와 Cezanne를 잘 알지 못해도

몇 개의 그림만으로 그림 간의 유사성을 파악하는 능력을 갖고 있기 때문

이러한 능력은 직접 학습하지 않더라도 과거의 다양한 경험으로부터 학습된 결과

모델도 이러한 능력을 학습시킬 수 있을까?

적은 데이터만 있어도 잘 학습할 수 있는 방법을 학습시킬 수 있을까?

Meta-Learning

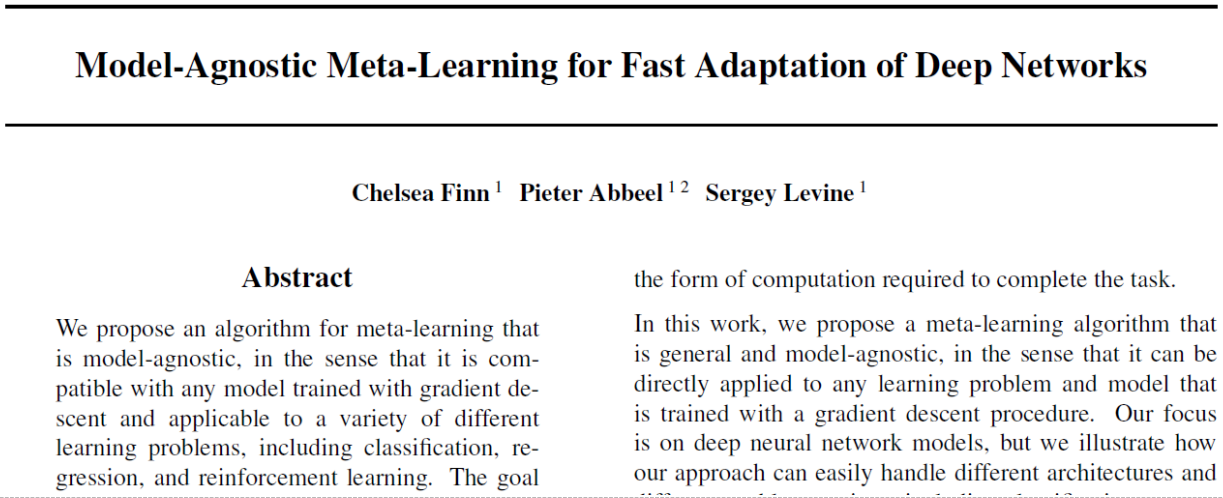
Introduction

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

  ✓ 1저자 : Chelsea Finn

  ✓ 34th International Conference on Machine Learning(ICML) 2017

  ✓ 인용 수 높음(686회), 후속 논문들이 많음



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

Chelsea Finn [1] **Pieter Abbeel** [1 2] **Sergey Levine** [1]

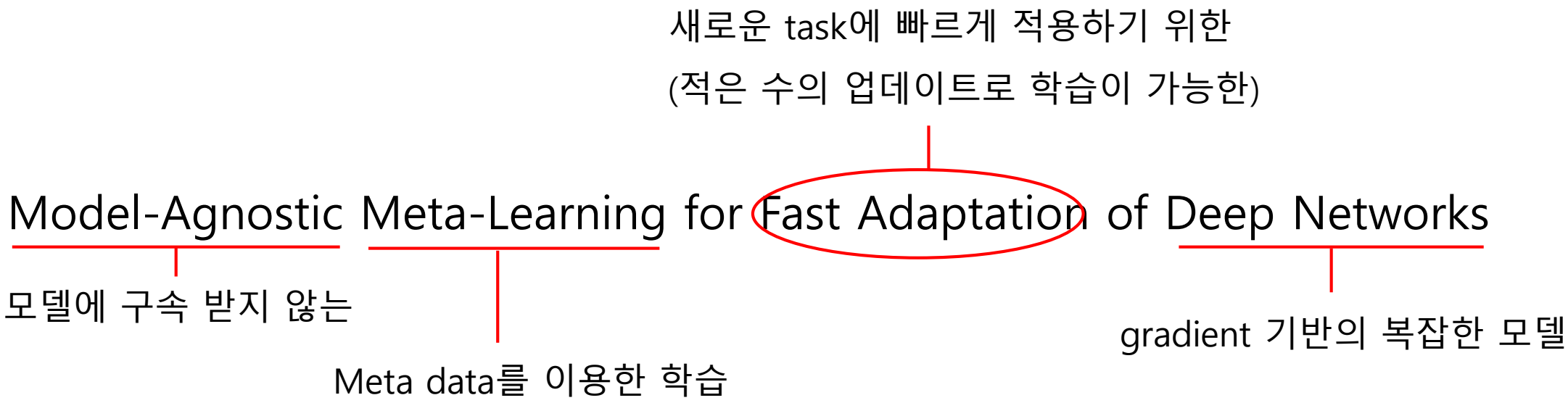
Abstract

We propose an algorithm for meta-learning that is model-agnostic, in the sense that it is compatible with any model trained with gradient descent and applicable to a variety of different learning problems, including classification, regression, and reinforcement learning. The goal

the form of computation required to complete the task.

In this work, we propose a meta-learning algorithm that is general and model-agnostic, in the sense that it can be directly applied to any learning problem and model that is trained with a gradient descent procedure. Our focus is on deep neural network models, but we illustrate how our approach can easily handle different architectures and

Introduction

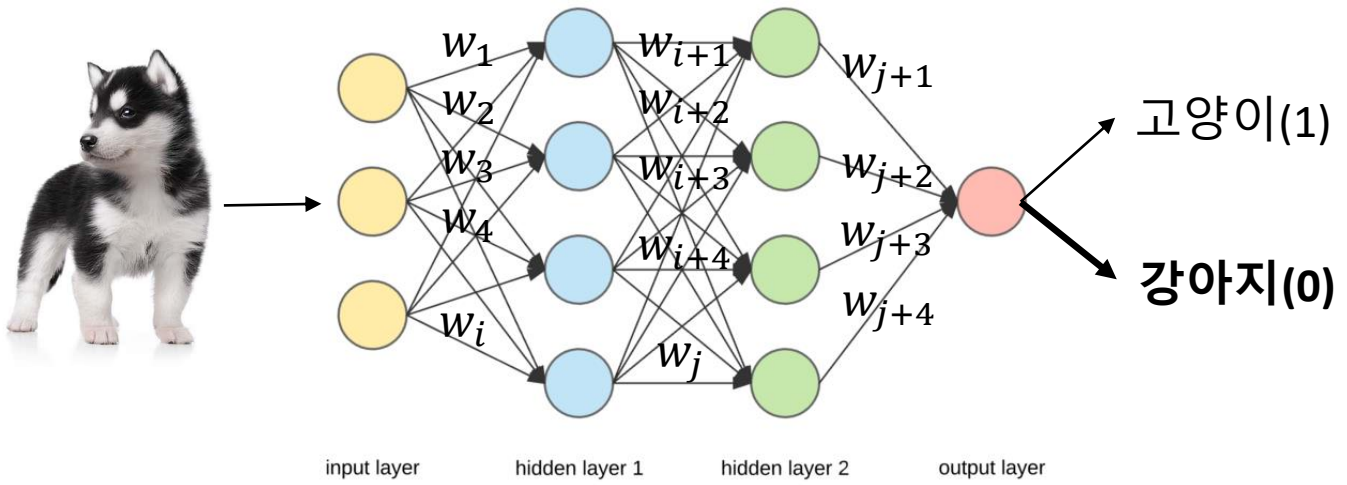


Problem Description

- 지도 학습

  ✓ 학습 데이터(D)가 주어졌을 때 최적의 파라미터(ø)를 구하는 것

  Ex. 개와 고양이 사진을 분류하는 인공신경망 모델



$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)\}$$

(x_1, y_1) = ( , 고양이)

(x_2, y_2) = ( , 강아지)

$$\emptyset = (w_1, w_2, w_3, \dots)$$

https://blog.goodaudience.com/artificial-neural-networks-explained-436fcf36e75

Problem Description

- 지도 학습

✓ 학습 데이터(D)가 주어졌을 때 최적의 파라미터(ø)를 구하는 것

Train data가 적다면?(2개)

$D = \{(x_1, y_1), (x_2, y_2)\}, \emptyset = (w_1, w_2, w_3, \dots)$

Train data(D)

(x_1, y_1) = ( , 고양이)

(x_2, y_2) = ( , 강아지)

Test data



일반적인 지도학습으로는 학습이 불가능

Meta-learning을 통해 학습된 모델을 사용하면 적은 데이터로도 충분한 학습이 가능

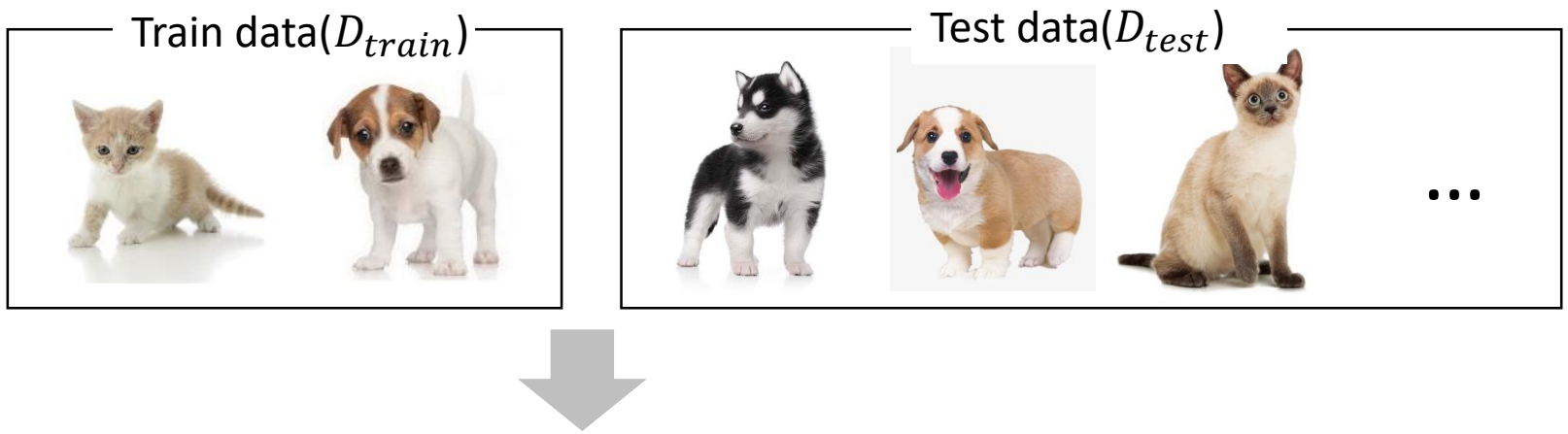
Problem Description

- 적은 데이터의 정의

  ✓ **N-way, k-shot**

  Classes : N 개

  Examples : k 개



2-way 1-shot classification

Train data는 총 n x k 개의 data point(x, y)로 이루어져 있음

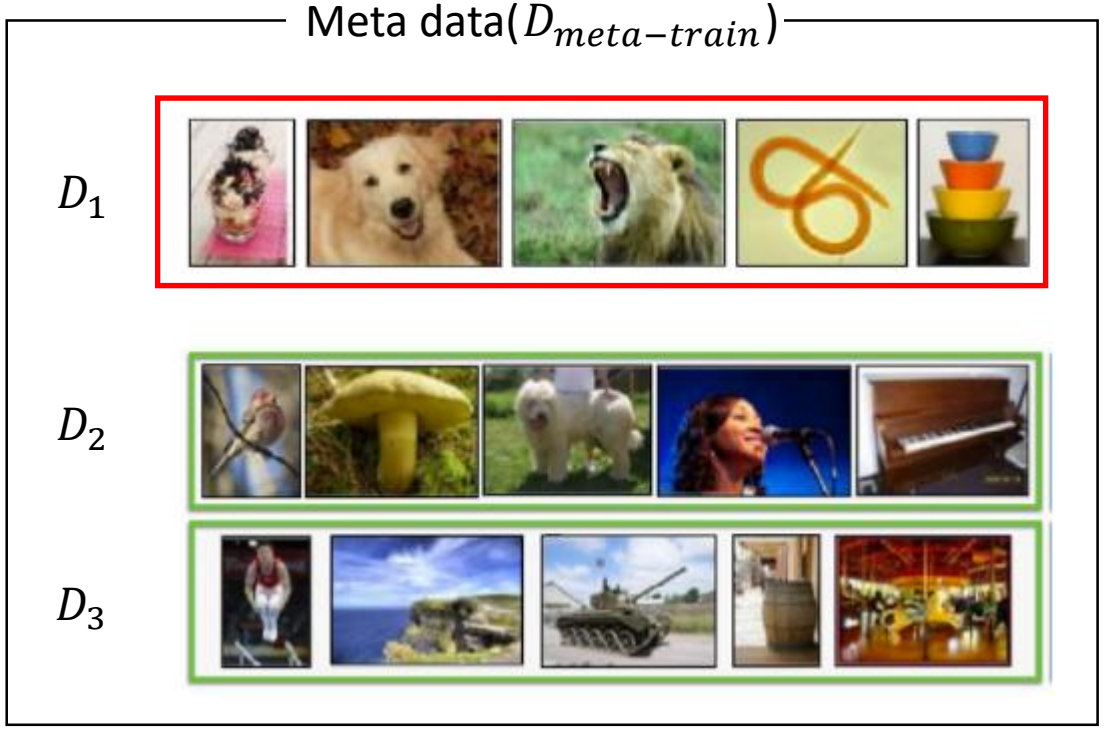
Test data의 경우 개수는 크게 상관이 없으며 test의 label은 학습에 사용하지 않음

Problem Description

- Meta-learning에 사용되는 데이터

✓ $\boldsymbol{D_{meta-train}}$는 train data와 비슷한 Task를 할 수 있는 다양한 데이터셋들로 이루어져 있음

$$D_{meta-train} = (D_1, D_2, D_3, \dots)$$



✓ 다양한 class들의 데이터가 가능

Train data의 class가 개, 고양이 라도

$D_{meta-train}$의 class는 사자, 사람, 전차 등 가능

✓ 비슷한 task의 예시

기존 task(T) : Train data를 이용하여 개와 고양이를 구분할 수 있는 파라미터(ø)를 학습시키는 것

Meta-task(T_1) : D_1을 이용하여 그릇과 사자를 구분할 수 있는 파라미터(ø)를 학습시키는 것

Image : Ravi and Larochelle, 2017

Problem Description

- 정리

  ✓ 적은 train data로 학습을 해야 하는 상황

  ✓ Train data와 비슷하지만 다른 데이터셋($D_{meta-train}$)을 사용하는 상황

  ✓ 비슷한 상황을 갖는 Multi-task learning과는 어떤 것이 다를까?

Problem Description

- Multi-task learning vs Meta-learning

✓ 다른 데이터셋($D_{meta-train}$)을 사용하는 것은 동일한 상황

Multi-task 관점

Task($T_1, T_2, \ldots$) 별 최적의 파라미터 $\emptyset_i$가 모두 동일

즉, 하나의 파라미터를 공유하는 하나의 큰 모델이 모든 task를 해결할 수 있다

$D_{meta-train}$를 이용하여 $\emptyset$를 학습

Meta-learning 관점

Task($T_1, T_2, \ldots$) 별 최적의 파라미터 $\emptyset_i$가 모두 다름

$D_{meta-train}$로 task 별 $\emptyset_i$들을 바로 학습하는 것은 의미가 없음(데이터 특성에 따라 $\emptyset_i$가 달라짐)

$D_{meta-train}$를 이용하여 데이터 특성과 $\emptyset_i$의 사이의 정보를(θ)를 학습

추후 새로운 데이터가 들어오면, θ를 이용하여 더 나은 학습이 가능

Meta-learning Algorithms

- Meta-learning Approach

✓ θ를 어떻게 사용하는지에 따라 접근 방법이 달라짐

- Metric-based approach
  1. $D_{meta-train}$를 이용하여 저차원 공간에 임베딩 방법(θ)을 학습
  2. 새로운 데이터 D 가 들어오면, 저차원 공간에 임베딩하여 가장 가까운 클래스로 분류

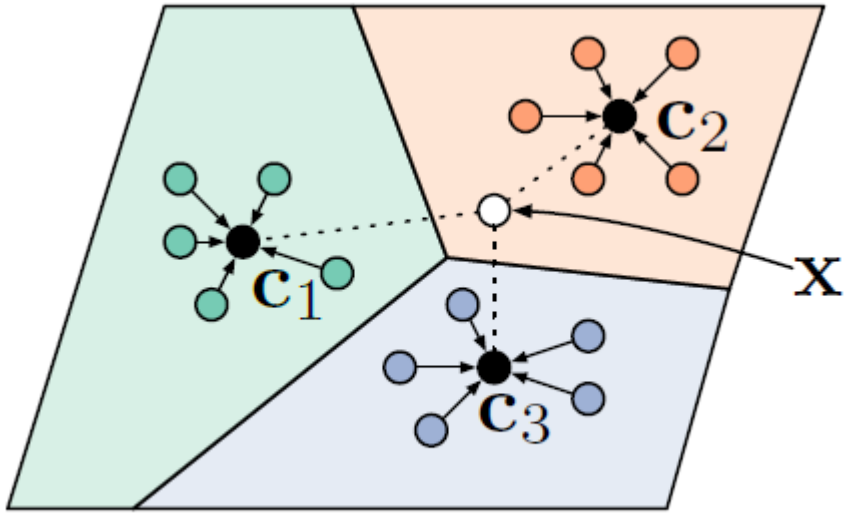
- Optimization-based approach
  1. $D_{meta-train}$를 이용하여 효율적인 update 방법(θ)을 배워
  2. 새로운 데이터 D가 들어오면 빠른 학습(adaptation)이 가능

- Bayesian, …

Meta-learning Algorithms

- Metric-based approach
  - ✓ Idea : 1. $D_{meta-train}$를 이용하여 저차원으로 임베딩 시키는 f_θ를 학습
    2. 임베딩 차원에서의 거리를 이용하여 가장 가까운 class로 할당
  - ✓ Prototypical Networks for Few-shot Learning

3-way 5-shot example



(a) Few-shot

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_{\boldsymbol{\phi}}(\mathbf{x}_i)$$

$$p_{\boldsymbol{\phi}}(y = k \mid \mathbf{x}) = \frac{\exp(-d(f_{\boldsymbol{\phi}}(\mathbf{x}), \mathbf{c}_k))}{\sum_{k'} \exp(-d(f_{\boldsymbol{\phi}}(\mathbf{x}), \mathbf{c}_{k'}))}$$

Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." *Advances in Neural Information Processing Systems*. 2017

Meta-learning Algorithms

- Meta-learning Approach

✓ θ를 어떻게 사용하는지에 따라 접근 방법이 달라짐

- Metric-based approach
  1. $D_{meta-train}$를 이용하여 저차원 공간에 임베딩 방법(θ)을 학습
  2. 새로운 데이터 D 가 들어오면, 저차원 공간에 임베딩하여 가장 가까운 클래스로 분류

- Optimization-based approach
  1. $D_{meta-train}$를 이용하여 효율적인 update 방법(θ)을 배워
  2. 새로운 데이터 D가 들어오면 빠른 학습(adaptation)이 가능

- Bayesian, …

Meta-learning Algorithms

- Optimization-based Approach
  - ✓ Idea : 1. $D_{meta-train}$를 이용하여 θ를 구하고
    2. 새로운 D와 θ를 이용하여 새로운 Task의 $\emptyset$를 빠르게 구함

Adaptation : $\underset{\emptyset}{\operatorname{argmax}} \log p(\emptyset|\theta, D)$

Meta-learning : $\underset{\theta}{\operatorname{argmax}} \log p(\theta|D_{meta-train})$

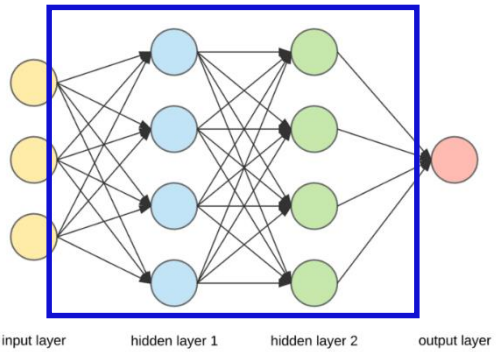
Meta-learning Algorithms

- Optimization-based Approach
  - ✓ Idea : 1. $D_{meta-train}$를 이용하여 θ를 구하고
    2. 새로운 D와 θ를 이용하여 새로운 Task의 $\emptyset$**를 빠르게 구함**

Adaptation : $\underset{\emptyset}{\text{argmax}} \log p(\emptyset|\theta, D)$

Meta-learning : $\underset{\theta}{\text{argmax}} \log p(\theta|D_{meta-train})$

Notation review



$\emptyset$: 인공신경망의 weights

https://blog.goodaudience.com/artificial-neural-networks-explained-436fcf36e75

Meta-learning Algorithms

- Optimization-based Approach
  - ✓ Idea : 1. $D_{meta-train}$를 이용하여 θ를 구하고
    2. 새로운 D와 θ를 이용하여 새로운 Task의 $\emptyset$**를 빠르게 구함**

$$\text{Adaptation} : \underset{\emptyset}{\operatorname{argmax}} \log p(\emptyset|\theta, D)$$

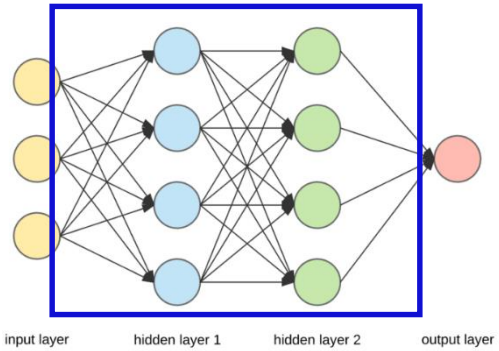
$$\text{Meta-learning} : \underset{\theta}{\operatorname{argmax}} \log p(\theta|D_{meta-train})$$

  - ✓ How?
    - θ를 $\emptyset$의 weight initialization으로 정의

$$\text{Adaptation} : \emptyset_i \leftarrow \theta - \alpha \nabla_\theta L(\theta, D_i^{tr})$$

$$\text{Meta-learning} : \theta \leftarrow \theta - \beta \nabla_\theta \sum L\left(\emptyset_i, D_i^{test}\right)$$

Notation review



$\emptyset$: 인공신경망의 weights

https://blog.goodaudience.com/artificial-neural-networks-explained-436fcf36e75

Meta-learning Algorithms

- Optimization-based Approach
  - ✓ $\emptyset_i \leftarrow \theta - \alpha \nabla_\theta L(\theta, D_i^{tr})$

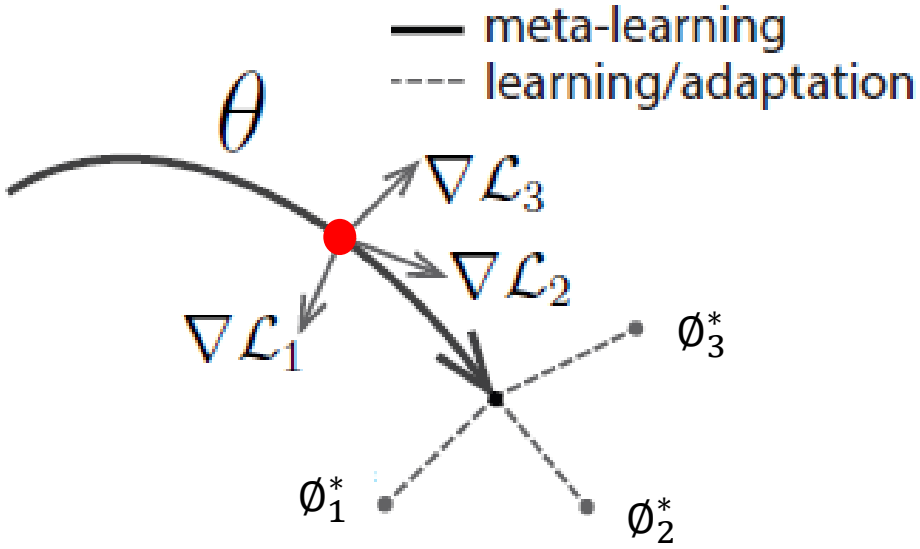
    θ를 $\emptyset_i$의 weight initialization으로 사용

    D_i^{tr}의 양이 적기 때문에 적은 update 만으로 θ -> $\emptyset_i$

  - ✓ $\theta \leftarrow \theta - \beta \nabla_\theta \sum L(\emptyset_i, D_i^{test})$

    $L(\emptyset_i, D_i^{test})$ 가 최소인 경우는 $L(\emptyset_i^*, D_i^{test})$

    즉 $\emptyset_i = \emptyset_i^*$가 되는 방향으로 θ를 업데이트



  - ✓ 즉, 적은 update 만으로 $\emptyset_i^*$를 구할 수 있는 θ를 찾는 것이 meta-learning의 목적

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Meta-learning Algorithms

- Optimization-based Approach
  - ✓ $\emptyset_i \leftarrow \theta - \alpha \nabla_\theta L(\theta, D_i^{tr})$

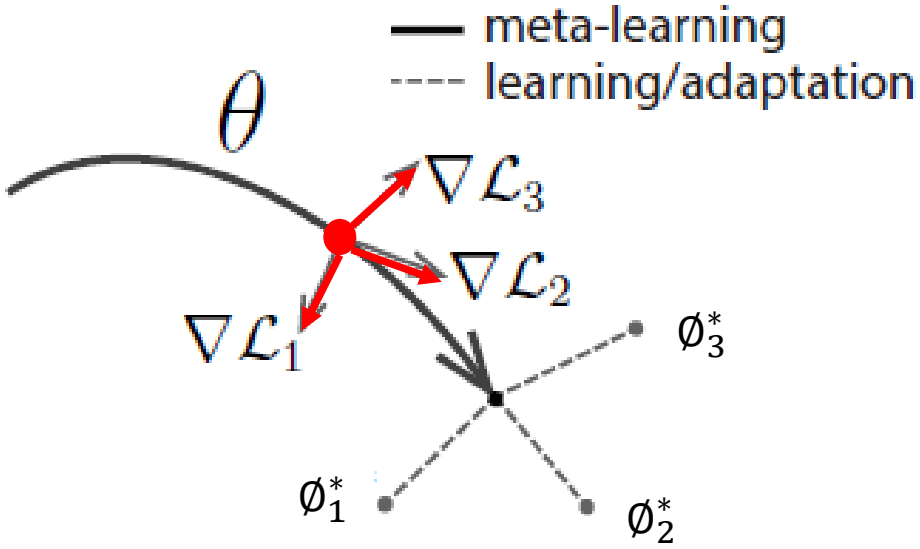
    θ를 $\emptyset_i$의 weight initialization으로 사용

    D_i^{tr}의 양이 적기 때문에 적은 update 만으로 θ -> $\emptyset_i$

  - ✓ $\theta \leftarrow \theta - \beta \nabla_\theta \sum L(\emptyset_i, D_i^{test})$

    $L(\emptyset_i, D_i^{test})$ 가 최소인 경우는 $L(\emptyset_i^*, D_i^{test})$

    즉 $\emptyset_i = \emptyset_i^*$가 되는 방향으로 θ를 업데이트



  - ✓ 즉, 적은 update 만으로 $\emptyset_i^*$를 구할 수 있는 θ를 찾는 것이 meta-learning의 목적

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Meta-learning Algorithms

- Optimization-based Approach
  - ✓ $\emptyset_i \leftarrow \theta - \alpha \nabla_\theta L(\theta, D_i^{tr})$

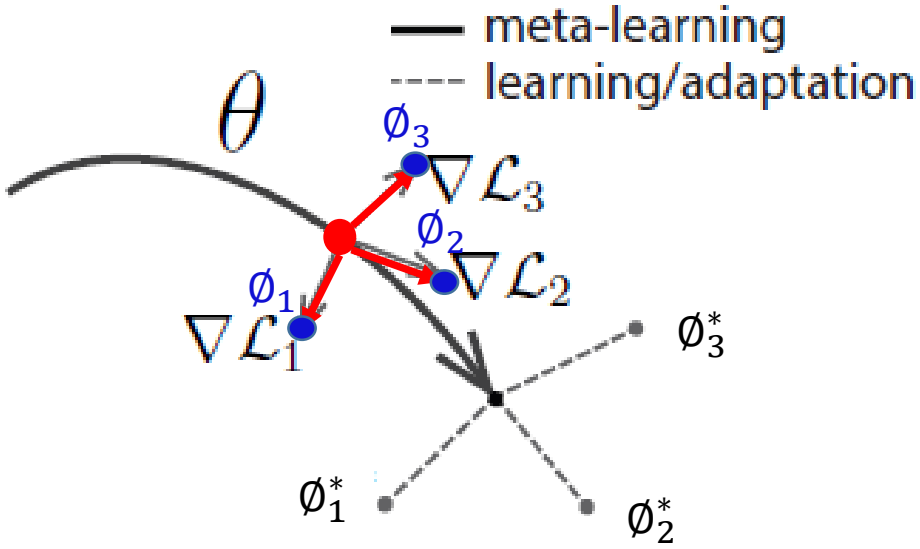
    θ를 $\emptyset_i$의 weight initialization으로 사용

    D_i^{tr}의 양이 적기 때문에 적은 update 만으로 θ -> $\emptyset_i$

  - ✓ $\theta \leftarrow \theta - \beta \nabla_\theta \sum L(\emptyset_i, D_i^{test})$

    $L(\emptyset_i, D_i^{test})$ 가 최소인 경우는 $L(\emptyset_i^*, D_i^{test})$

    즉 $\emptyset_i$ = $\emptyset_i^*$가 되는 방향으로 θ를 업데이트



  - ✓ 즉, 적은 update 만으로 $\emptyset_i^*$를 구할 수 있는 θ를 찾는 것이 meta-learning의 목적

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Meta-learning Algorithms

- Optimization-based Approach
  - ✓ $\emptyset_i \leftarrow \theta - \alpha \nabla_\theta L(\theta, D_i^{tr})$

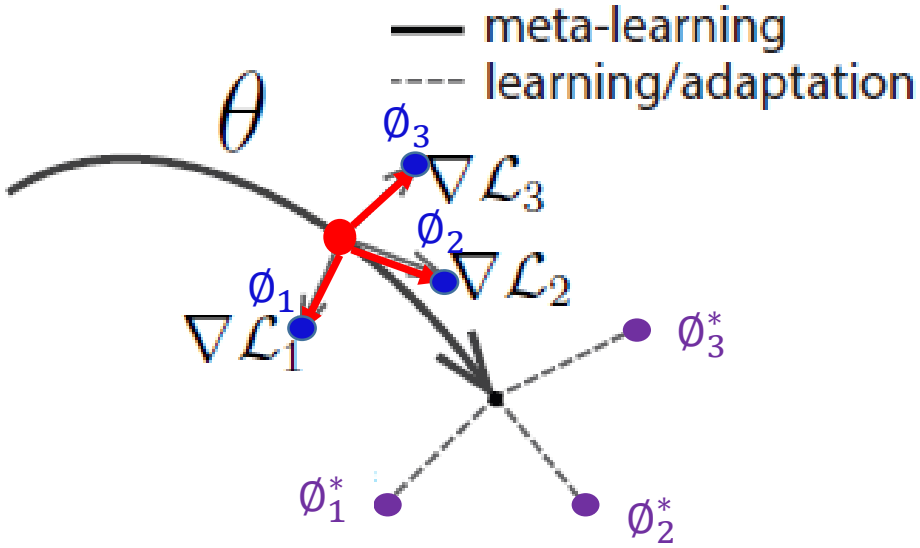
    θ를 $\emptyset_i$의 weight initialization으로 사용

    D_i^{tr}의 양이 적기 때문에 적은 update 만으로 θ -> $\emptyset_i$

  - ✓ $\theta \leftarrow \theta - \beta \nabla_\theta \sum L(\emptyset_i, D_i^{test})$

    $L(\emptyset_i, D_i^{test})$ 가 최소인 경우는 $L(\emptyset_i^*, D_i^{test})$

    즉 $\emptyset_i$ = $\emptyset_i^*$가 되는 방향으로 θ를 업데이트



  - ✓ 즉, 적은 update 만으로 $\emptyset_i^*$를 구할 수 있는 θ를 찾는 것이 meta-learning의 목적

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Meta-learning Algorithms

- Optimization-based Approach
  - ✓ $\emptyset_i \leftarrow \theta - \alpha \nabla_\theta L(\theta, D_i^{tr})$

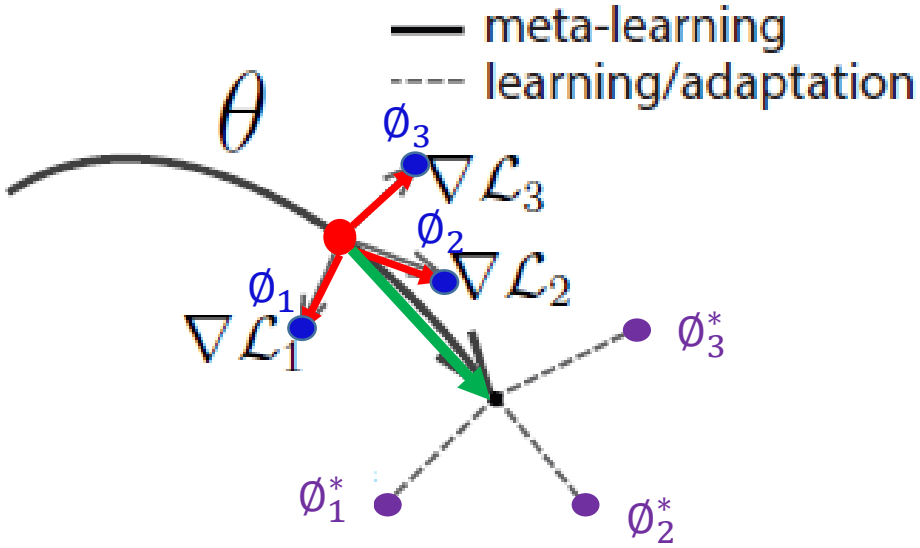
    θ를 $\emptyset_i$의 weight initialization으로 사용

    D_i^{tr}의 양이 적기 때문에 적은 update 만으로 θ -> $\emptyset_i$

  - ✓ $\theta \leftarrow \theta - \beta \nabla_\theta \sum L(\emptyset_i, D_i^{test})$

    $L(\emptyset_i, D_i^{test})$ 가 최소인 경우는 $L(\emptyset_i^*, D_i^{test})$

    즉 $\emptyset_i$ = $\emptyset_i^*$가 되는 방향으로 θ를 업데이트



  - ✓ 즉, 적은 update 만으로 $\emptyset_i^*$를 구할 수 있는 θ를 찾는 것이 meta-learning의 목적

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Meta-learning Algorithms

- Optimization-based Approach
  0. $D_{meta-train}(D_1, D_2, D_3)$

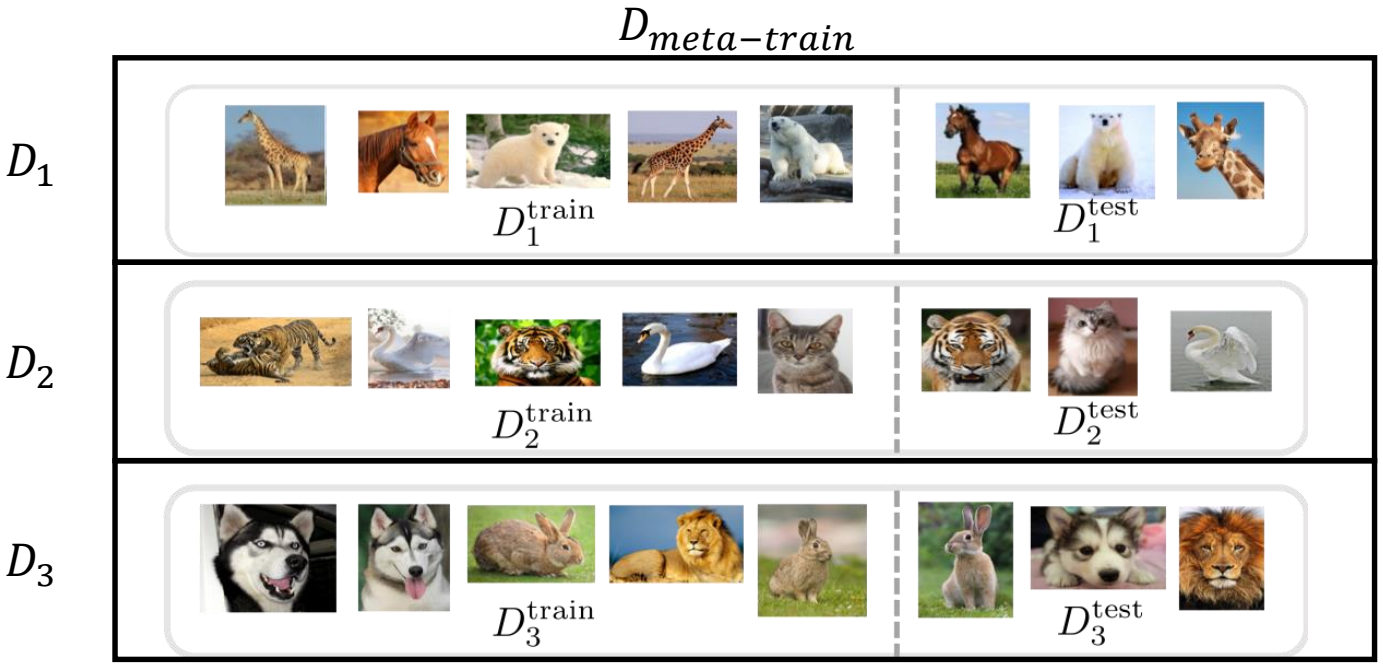


Image : Ravi and Larochelle, 2017

Meta-learning Algorithms

- Optimization-based Approach
  1. D_i 를 D_i^{train}, D_i^{test} 으로 분할

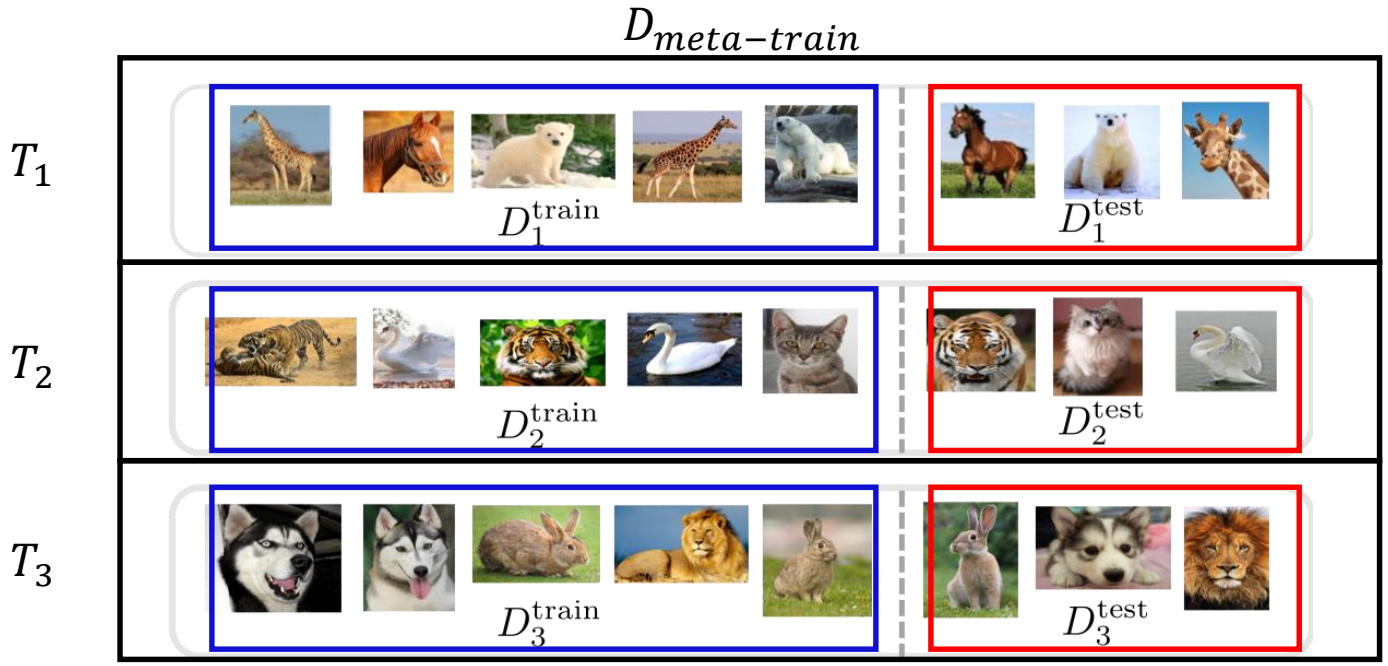


Image : Ravi and Larochelle, 2017

Meta-learning Algorithms

- Optimization-based Approach
  2. θ 와 D_i^{train}를 이용하여 $\emptyset_i$를 구함(모델 학습)

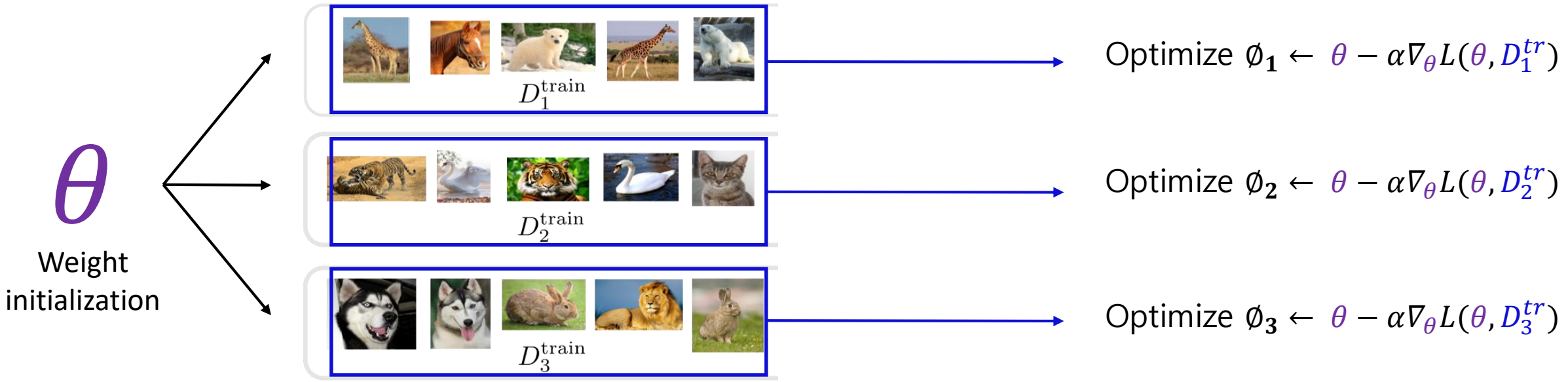


Image : Ravi and Larochelle, 2017

Meta-learning Algorithms

- Optimization-based Approach
  3. $\emptyset_i$와 D_i^{test}을 이용하여 θ update

$$\theta' \longleftarrow \theta - \beta \nabla_\theta \sum L(\emptyset_i, D_i^{test})$$

기존의 인공신경망 학습

$$\emptyset = \emptyset - \beta \nabla_\emptyset \sum L(\emptyset_i, D_i^{test})$$

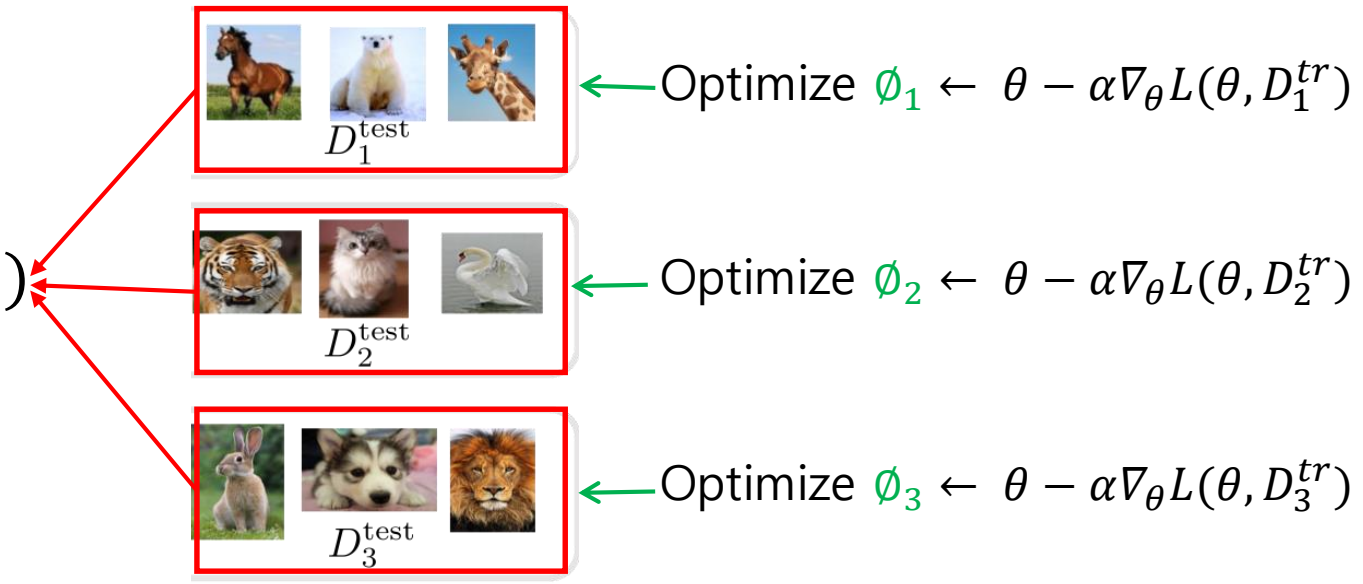


Image : Ravi and Larochelle, 2017

Meta-learning Algorithms

- Optimization-based Approach

  4. θ를 이용하여 새로운 Task에 빠르게 학습(Adaptation)하는 과정

  ex)5-way, 1-shot classification

  New data :$D_{train}^{new}, D_{test}^{new}$

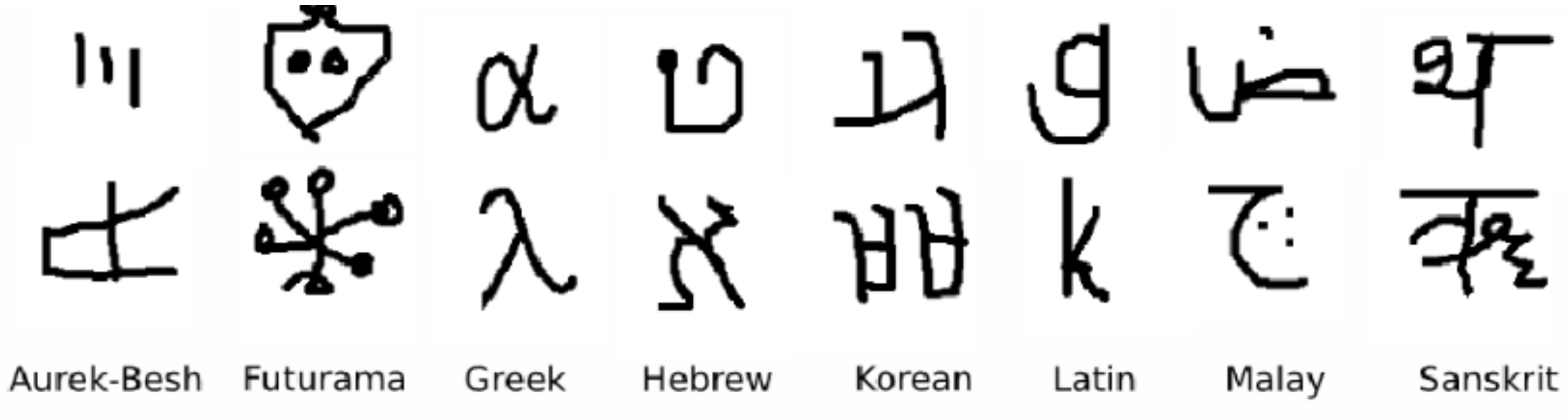
optimize $\emptyset_{new} \leftarrow \theta - \alpha \nabla_{\theta} L(\theta, D_{train}^{new})$

$$\hat{y} = f_{\emptyset_{new}}(x_{test})$$

Image : Ravi and Larochelle, 2017

Experiments

- Classification
  - ✓ Omniglot dataset 사용
    - 50 다른 언어로 된 1623개의 글자로 이루어짐
    - 글자 별 샘플 각각 20개(총 data points : 1623 x 20 개)
    - 샘플은 모두 다른 사람이 작성한 것
    - Data 예시



https://github.com/Goldesel23/Siamese-Networks-for-One-Shot-Learning

Experiments

- Classification
  - ✓ Meta-train : 1200 characters, test : 423 characters
  - ✓ Matching nets 과 동일한 architecture 사용

Omniglot (Lake et al., 2011)	5-way Accuracy		20-way Accuracy	
	1-shot	5-shot	1-shot	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%**

  - ✓ 대부분의 모델들이 높은 정확도를 나타내지만 그 중에서 가장 좋은 성능을 보임

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

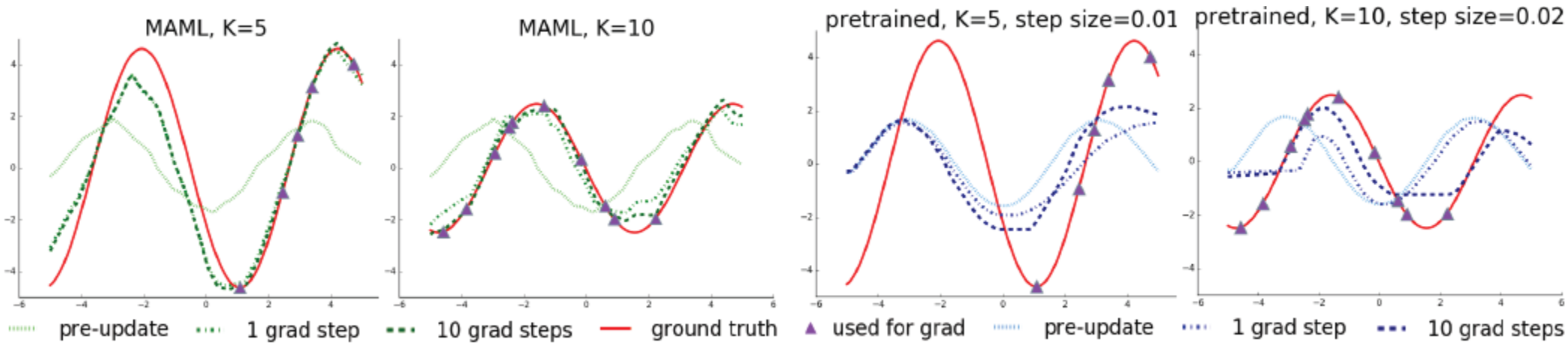
Experiments

- Regression
  - ✓ Meta train : 진폭(0.1~0.5), 위상(0~π) 인 sine 그래프
  - ✓ Test : meta train에 포함되지 않는 sine 그래프
  - ✓ Data point : x : (-0.5~0.5) 에서 샘플링(K개 만큼 사용)
  - ✓ Task : x 가 주어졌을 때 y 를 예측하는 것
  - ✓ Model : 40개 unit(activation function : RELU)의 2 hidden layers를 갖는 인공신경망 모델
  - ✓ Loss function : mean squared error

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Experiments

- Regression

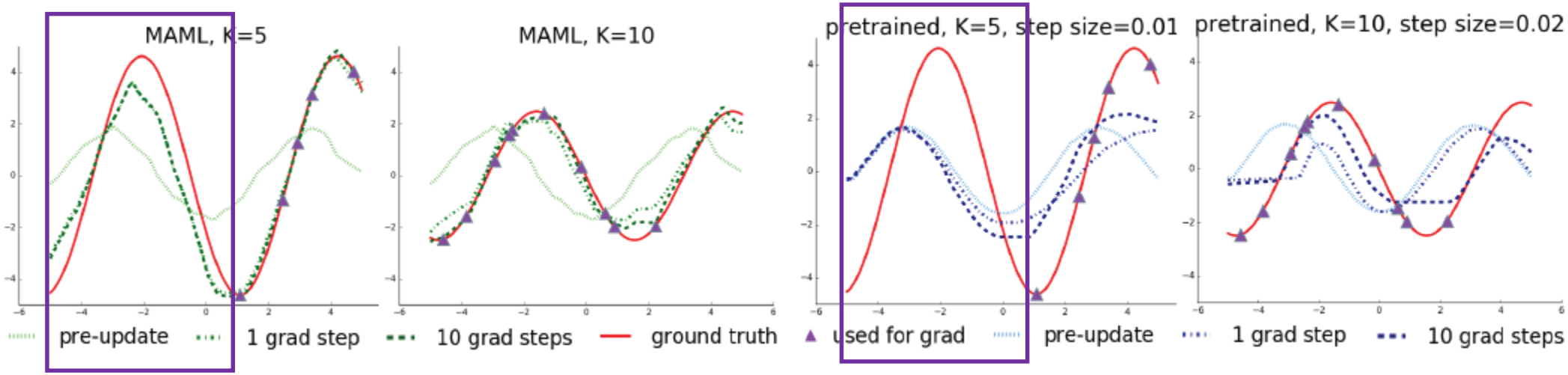


- 연두색 : MAML update 전(θ), 초록색 : MAML update 후($\emptyset$), 빨간색 : 실제 값
- 하늘색 : pretrained model update 전, 파란색 : pretrained model update
- 연두색과 하늘색이 거의 같음 -> MAML의 업데이트 속도가 훨씬 빠름
- MAML 모델의 경우 데이터 포인트가 없는 곳도 예측 가능

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Experiments

- Regression



✓ 연두색 : MAML update 전(θ), 초록색 : MAML update 후($\emptyset$), 빨간색 : 실제 값

✓ 하늘색 : pretrained model update 전, 파란색 : pretrained model update

✓ 연두색과 하늘색이 거의 같음 -> MAML의 업데이트 속도가 훨씬 빠름

✓ MAML 모델의 경우 데이터 포인트가 없는 곳도 예측 가능

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Conclusions

- Contributions

  ✓ 몇 회만의 update 만으로 target task 에서 높은 정확도를 낼 수 있었다. (fast adaptation)

  ✓ Gradient descent만 사용하기 때문에 간단하고, 원래 모델의 구조를 공유하기 때문에 추가적인 모델링 과정이 필요 없음

  ✓ **gradient-based로 학습되는 모든 모델에 적용할 수 있음(regression, classification, reinforcement learning 등)**

감사합니다