A Brief Introduction to Meta Learning

by
Anurag Roy
and
Soham Poddar

Indian Institute of Technology, Kharagpur

June 18th, 2020

Overview

Where are we in Machine Learning?

- As Deep Learning became popular, machine learning models have reached remarkable heights.
- Deep Learning models can be applied to virtually any task

Where are we in Machine Learning?

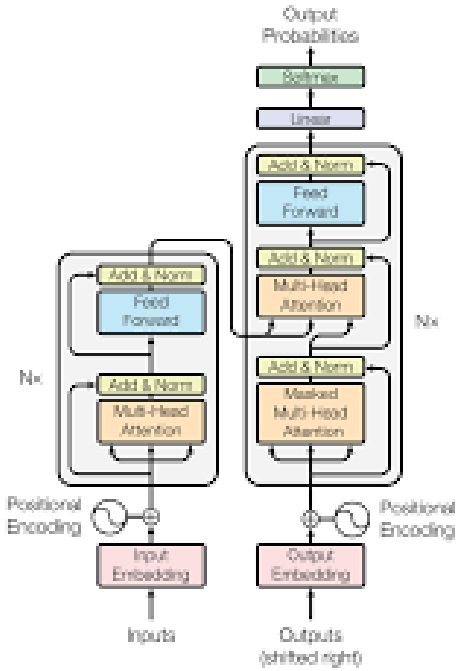


Figure 1: The Transformer - model architecture.

(a) Transformer Architecture

(b) Deep learning models designed for the ImageNet challenge

Figure: Recent deep learning models achieve performance similar to (and sometimes surpassing) human level performance in various areas.

Not enough Labelled Data!

- Modern Deep Learning Models need lots of data that are labelled into different classes!!
- Dealing with this Few-shot learning problem:
  - How does a baby learn different tasks? What if we incorporate skills learnt from other tasks?

- Like in:
  - Transfer Learning
  - Multi-Task Learning

Transfer Learning

- Aimed at transferring knowledge of a source task to a target task.
- Example:

Figure: Learning Bicycle helps to learn riding a Motorcycle

Transfer Learning

- Aimed at transferring knowledge of a source task to a target task.

- Example:

Gloss	Latin	Catalan	Occitan	French	Italian
"window"	FENESTRA	*finestra*	*fenèstra*	*fenêtre*	*finestra*
"to eat"	MANDVCĀRE	*menjar*	*manjar*	*manger*	*mangiare*
"morning"	MATVTĪNVS	*matí*	*matin*	*matin*	*mattina*
"to speak"	PARABOLĀRE	*parlar*	*parlar*	*parler*	*parlare*
"table"	TABVLA	*taula*	*taula*	*table*	*tavola*

Figure: Knowing a Romance Language helps in learning another easily.

What is a task?

- A task can be any well-defined machine learning problem.

Figure: Classifying images of cats and dogs

Figure: Playing Chess, Language Translation

Multi Task Learning

- Learns "general" properties (hyperparameters or weights) by learning a number of different tasks simultaneously.
  - Reduces parameters (for faster training and less overfit)
  - Improving performance on tasks with less data, using underlying common structures present in the data.

- Examples:
  - Learning different science subjects in the same grade in school
  - Learning a new language: Learning to Speak and Read together.

Meta Learning

Can we learn to learn?

Meta Learning in action

By Braque or Cezanne?

Image Courtesy: ICML 2019 Meta Learning Tutorial

Meta Learning in action

How did we guess it pretty easily from just 3 examples?

- Have seen numerous images over the years
- Learnt to recognize features of objects and patterns in images over years

What is Meta Learning?

The Basic Premise:

Given experience on previous tasks learn a **new task** more quickly and/or more proficiently.

How is it different from Transfer learning?

- **Transfer Learning:** No concept of meta-learner, model parameters of another task(s) used in it.
- **Meta-Learning**: Has a meta-learner which is evaluated with the help of some tasks.

How is it different from Multi-Task learning?

- **Multi-Task Learning:** Performs several tasks simultaneously. No Concept of adding a new Task.
- **Meta-Learning:** Learns to learn new tasks by training on previous tasks.

Overview

Motivation

- This is a very timely topic that can solve several practical problems.
- Classifying dogs and cats are fun but how are these complex models being applied to **real** problems?

In Computer Vision and Graphics

- Object Detection and segmentation
  - Few-shot multi-class classification in image recognition tasks.
  - Few-shot object segmentation is important due to the cost of obtaining pixel-wise labeled images in this domain.

In Computer Vision and Graphics

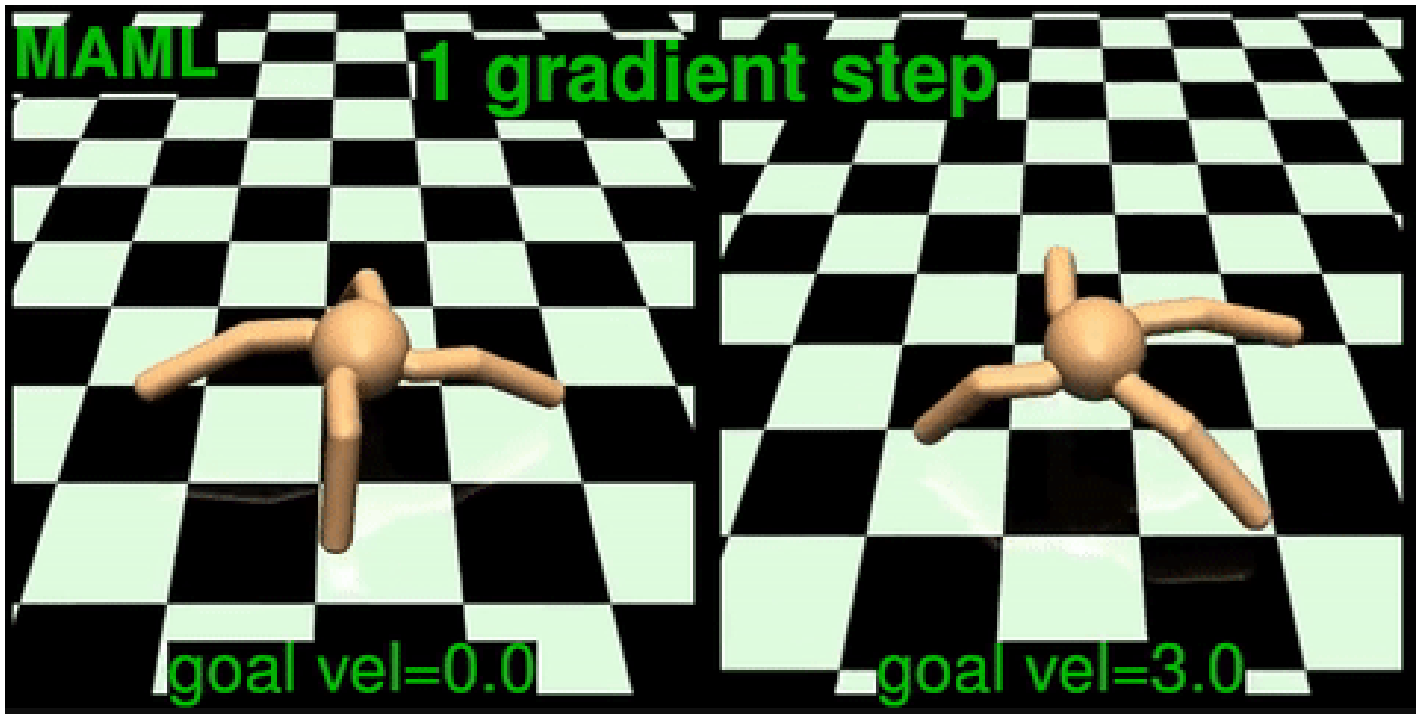
- Object Detection and segmentation
  - Few-shot multi-class classification in image recognition tasks.
  - Few-shot object segmentation is important due to the cost of obtaining pixel-wise labeled images in this domain.
- Image and Video Generation
  - A meta-learner is used to generate multiple views of an object from just a single image

In Computer Vision and Graphics

- Object Detection and segmentation
  - Few-shot multi-class classification in image recognition tasks.
  - Few-shot object segmentation is important due to the cost of obtaining pixel-wise labeled images in this domain.
- Image and Video Generation
  - A meta-learner is used to generate multiple views of an object from just a single image
- Landmark Prediction
  - Find the location of skeleton key points within an image, such as such as joints in human or robot images.

Reinforcement Learning and Robotics

- Knowledge Transfer to Robots: walking, navigating and object pick/place - subroutines of cleaning up a room.



- Unsupervised meta-RL variants aim to perform meta-training without manually specified reward

Image Courtesy: BAIR blog on Meta Learning

Language and Speech

- Language Modelling
  - Filling in missing words in texts
  - Language Translation
  - Quickly adapting to new personas in dialogue tasks
  - English to SQL program synthesis

Language and Speech

- Language Modelling
  - Filling in missing words in texts
  - Language Translation
  - Quickly adapting to new personas in dialogue tasks
  - English to SQL program synthesis
- Speech Recognition
  - Automatic speech recognition for low-resource languages
  - Cross-Accent adaptation
  - Optimising models for individual speakers

Model Optimization

- Network Architecture Search
  - Hyperparameter optimization
  - Produce compact and efficient models

Model Optimization

- Network Architecture Search
  - Hyperparameter optimization
  - Produce compact and efficient models
- Network Compression
  - Reducing the required memory for training contemporary networks for use in embedded systems.
  - Quantization and pruning are topical research areas,

Model Optimization

- Network Architecture Search
  - Hyperparameter optimization
  - Produce compact and efficient models
- Network Compression
  - Reducing the required memory for training contemporary networks for use in embedded systems.
  - Quantization and pruning are topical research areas,
- Learning from Data with Label Noise
  - Meta-learning methods have addressed robust learning to noisy label training

In Medical Domain

- COVID-19 outbreak demonstrates the need for robust models that learn from relatively few examples and rapidly adapt to changes in distribution.

- Progress in the medical domain is especially relevant given the global shortage of pathologists

In Medical Domain

- COVID-19 outbreak demonstrates the need for robust models that learn from relatively few examples and rapidly adapt to changes in distribution.
- Progress in the medical domain is especially relevant given the global shortage of pathologists
- Tasks such as medical image classification, drug discovery and protein synthesis, where data is often scarce.

Challenges in Meta Learning

- Generalization
  - Fitting a meta-learner to wide variety of tasks.
  - Generalizing to new task with very different distribution than the meta-trained tasks

Challenges in Meta Learning

- Generalization
  - Fitting a meta-learner to wide variety of tasks.
  - Generalizing to new task with very different distribution than the meta-trained tasks
- Multi-modality of task
  - In theory, tasks can be multimodal
  - Practically, most models at the moment are still unimodal.

Challenges in Meta Learning

- Generalization
  - Fitting a meta-learner to wide variety of tasks.
  - Generalizing to new task with very different distribution than the meta-trained tasks
- Multi-modality of task
  - In theory, tasks can be multimodal
  - Practically, most models at the moment are still unimodal.
- Computation Cost
  - Naive implementations can be quite expensive!

Overview

Supervised Learning

Q: What is supervised Learning?

$$\phi^* = \underset{\phi}{\text{argmax}} \log P(\phi|D)$$

Here the D is the training dataset and ϕ is the set of model parameters.

Supervised Learning

Q: What is supervised Learning?

$$\phi^* = \underset{\phi}{\text{argmax}} \log P(\phi|D)$$

Here the D is the training dataset and ϕ is the set of model parameters.

For Few-shot learning problems D is small.

Q: How can we incorporate prior knowledge?

What is Meta Learning?

- $D = \{(x_1, y_1), \cdots, (x_k, y_k)\}$
  - Example: 5-way Image classification data

- $D_{meta-train} = \{D_1, D_2, \cdots, D_n\}$
- $D_i = \{(x_1^i, y_1^i), \cdots, (x_k^i, y_k^i)\}$
  - Example: 5-way Image classification data, but with different set of objects in each D_i

What is Meta Learning?

- $D = \{(x_1, y_1), \cdots, (x_k, y_k)\}$
  - Example: 5-way Image classification data

- $D_{meta-train} = \{D_1, D_2, \cdots, D_n\}$
- $D_i = \{(x_1^i, y_1^i), \cdots, (x_k^i, y_k^i)\}$
  - Example: 5-way Image classification data, but with different set of objects in each D_i

- A model learning from D and $D_{meta-train}$ will try to find optimal ϕ.

$$\phi^* = \underset{\phi}{\text{argmax}} \log P(\phi | D, D_{meta-train})$$

What do the datasets contain?

- $D_{meta-train}$ consists of a set of task specific datasets.
- D contains the data for the primary task.

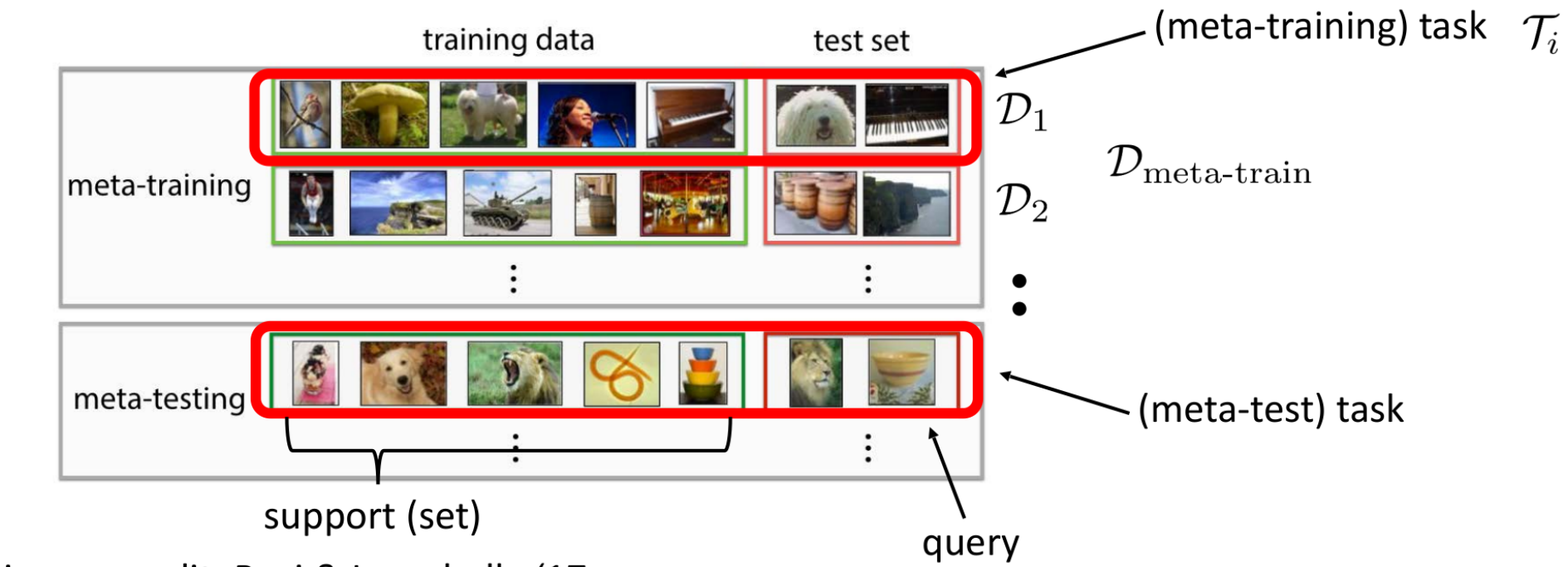


image credit: Ravi & Larochelle '17

Meta Learning: Learning meta parameters

- Q. Do we have to keep $D_{meta-train}$ forever?
  - No!
- Introduce a set of parameters θ that will represent the information in the prior data $D_{meta-train}$.

The complete Framework

- Our main objective is to learn θ which helps create a good ϕ for a new task.
- The complete meta-learning framework consists of two main steps:
  - Learn meta parameters (Meta Training) : $\theta^* = \underset{\theta}{\text{argmax}} \log P(\phi | D_{meta-train})$
  - Adaptation to a new task : $\phi_* = \underset{\phi}{\text{argmax}} \log P(\phi | D, \theta^*)$

How is it trained?

- Adaptation: $\phi_* = \underset{\phi}{\text{argmax}} \log P(\phi|D, \theta^*)$

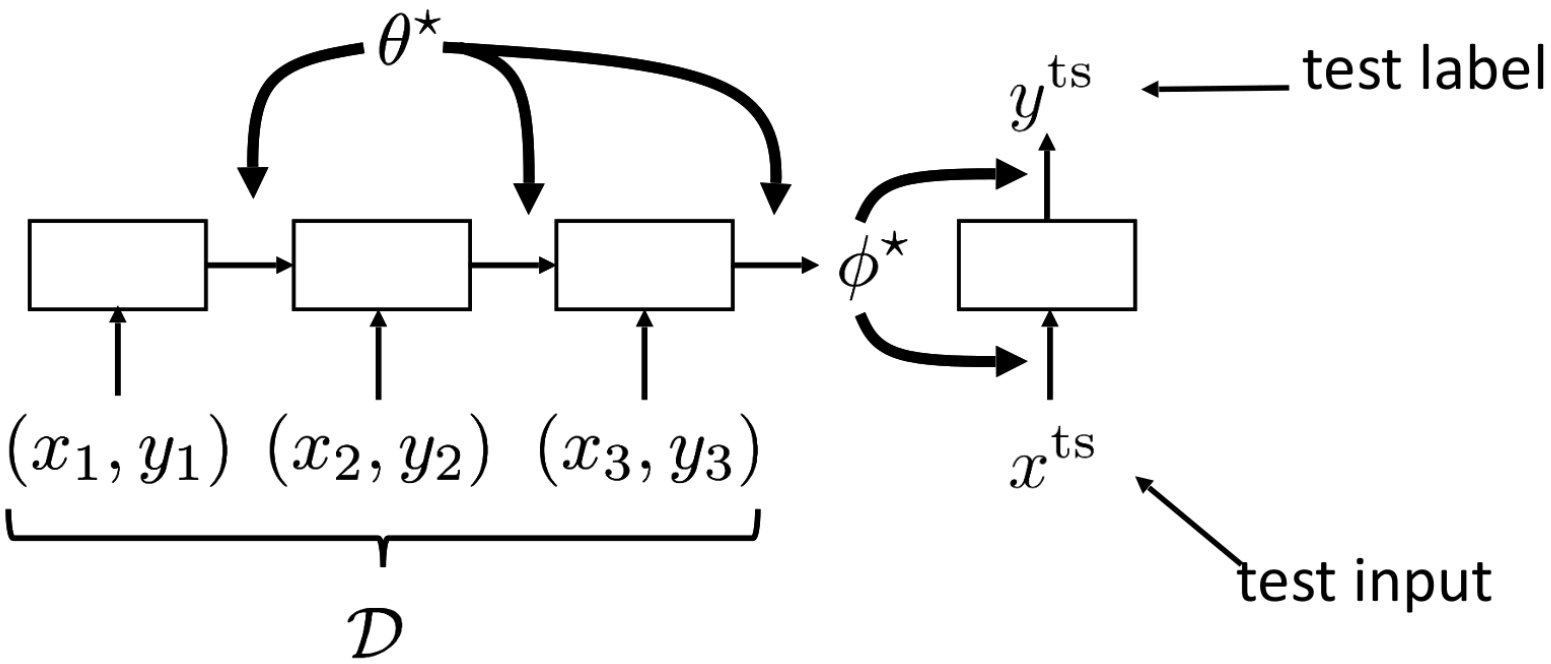


Image Courtesy: ICML 2019 Meta Learning Tutorial

How is it trained?

- Meta Training : $\theta^* = \underset{\theta}{\text{argmax}} \log P(\phi | D_{meta-train})$

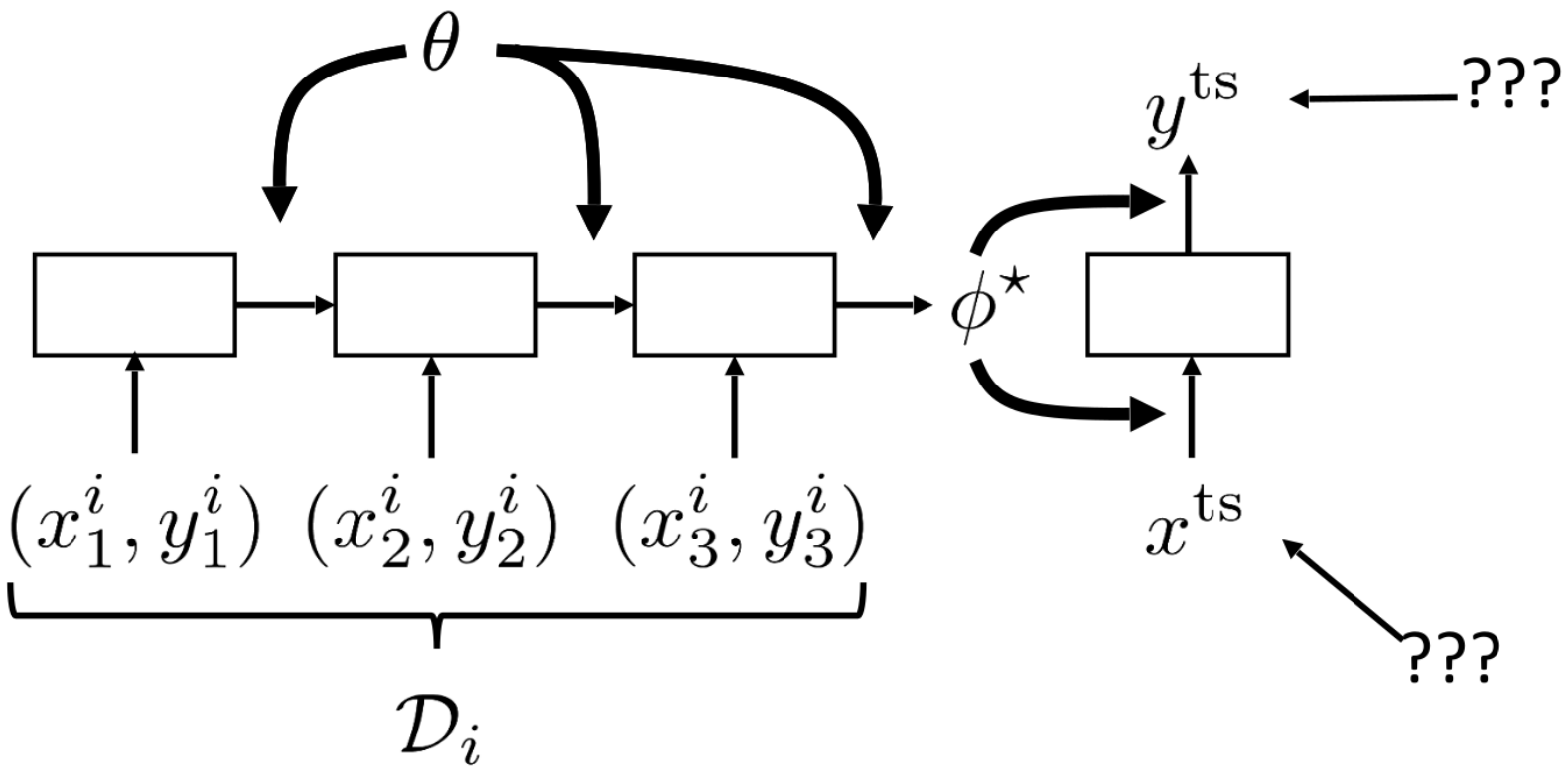


Image Courtesy: ICML 2019 Meta Learning Tutorial

How is the model evaluated during Meta Train?

Obtaining Optimal θ^*:

$$\theta^* = \max_{\theta} \sum_{i=1}^{N} \log P(\phi_i | D_i^{ts})$$

where $\phi_i = f_\theta(D_i^{tr})$

[Note: D_i^{tr} and D_i^{ts} are the train and test splits of the i-th task.]

Overview

Overview of Meta Learning Process

In general the training steps of a meta learning model involves the following general steps:

- Sample task $\mathcal{T}_i$ from a set of N tasks
- Sample disjoint train and test sets D^{tr} and D^{ts} from $D_i \in D_{meta-train}$
- Calculate ϕ^*
- Calculate θ^*

Common Approaches

Most meta-learning approaches can be categorized in three main categories [Vinyals, 2017]:

- Metric Based(a.k.a non parametric)
- Model Based(a.k.a Black Box)
- Optimization Based

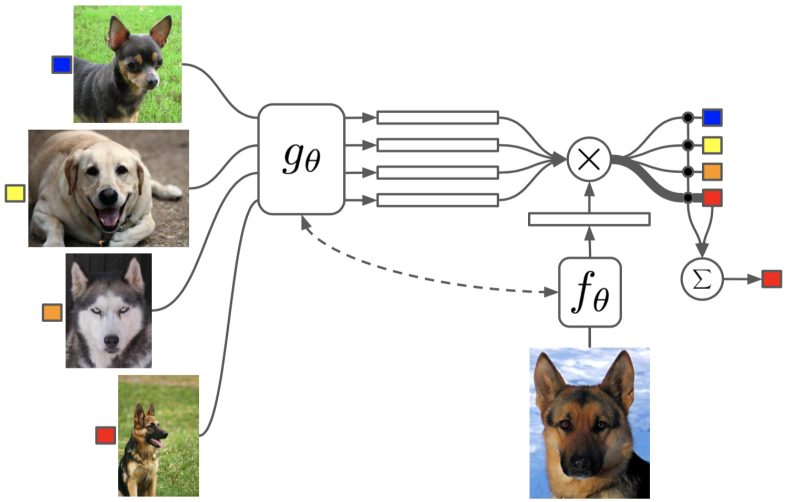
Metric Based Approach

- kNN like idea.
- The predicted probability P_θ of a label y is a weighted sum of lables in the training set.
- This weighting is done using a function $f_\theta()$.
- The objective of metric based meta-learning is to learn the above function.
- Some examples are Siamese Networks [Gregory Koch, 2015], Matching Networks [Vinyals et al., 2016], Prototypical Networks [Snell et al., 2017]

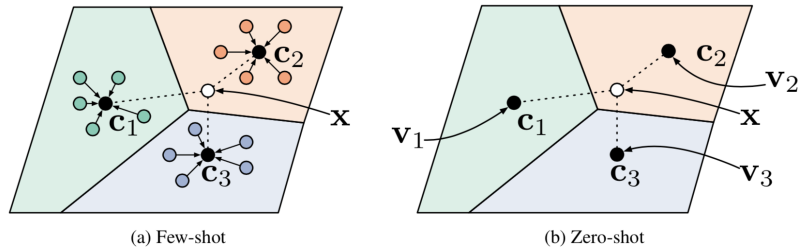
In other words metric based meta-learning finds P_θ as:

$$P_\theta(y|x, D^{tr}) = \sum_{(x_i,y_i)\in D^{tr}} f_\theta(x, x_i)y_i$$

Example



(a) Matching Network



(b) Prototypical Network

Figure: Metric based approach: Example Networks

Image Courtesy: [Snell et al., 2017, Vinyals et al., 2016]

Metric Based Approach: General Algo

Most of the algorithms belonging to this category can be summarized to the following steps:

1. Sample task $\mathcal{T}_i$ from a set of N tasks
2. Sample disjoint train and test sets D^{tr} and D^{ts} from $D_i \in D_{meta-train}$
3. Compute $\hat{y}^{ts} = \sum_{(x_i,y_i)\in D^{tr}} f_\theta(x^{ts}, x_i) y_i$
4. Update θ using $\nabla_\theta \mathcal{L}(\hat{y}^{ts}, y)$

Model Based Approach

- Key idea is to get $P(\phi_i|D_i^{tr}, \theta)$ using the feed-forward pass of a single model $f_\theta(x, D^{tr})$.
- In other words uses Neural Network to get parameters of a neural network (or some sufficient statistic) for a new task.
- f_θ is usually represented by architectures which have good memory (like RNN [Ravi and Larochelle, 2017],
  Meta-Networks [Munkhdalai and Yu, 2017],
  MANN [Santoro et al., 2016])
- Requires large number of meta-train tasks N.
- Less efficient in terms of generalizing to OOD tasks than Optimization based Approaches. [Finn and Levine, 2018]

Model Based Approach: General Algo

Most of the algorithms belonging to this category can be summarized to the following steps:

1. Sample task $\mathcal{T}_i$ from a set of N tasks
2. Sample disjoint train and test sets D^{tr} and D^{ts} from $D_i \in D_{meta-train}$
3. Compute $\phi_i = f_\theta(D_i^{tr})$
4. Update θ using $\nabla_\theta \mathcal{L}(\phi_i, D_i^{tr})$

Optimization Based Approach

- Similar to fine-tuning
- meta-parameters θ act as initializers for task specific parameters ϕ_i
- Involves second-order derivatives in backprop
- Usually memory intensive(e.g. in MAML [Finn et al., 2017])
- Some examples are MAML, FOMAML [Nichol et al., 2018], Reptile [Nichol and Schulman, 2018]

Example

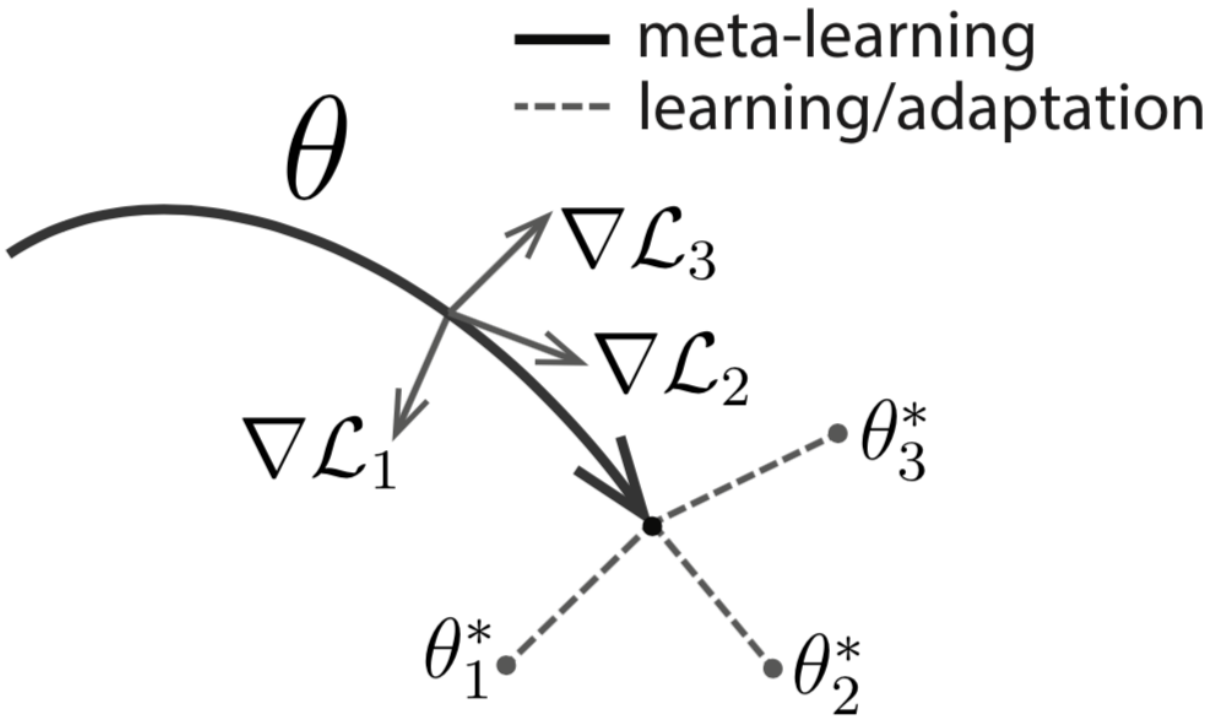


Figure: Optimization Based Network: MAML

Image Courtesy: [Finn et al., 2017]

Optimization Based Approach: General Algo

Most of the algorithms belonging to this category can be summarized to the following steps:

1. Sample task $\mathcal{T}_i$ from a set of N tasks
2. Sample disjoint train and test sets D^{tr} and D^{ts} from $D_i \in D_{meta-train}$
3. Optimize $\phi_i = \theta - \alpha \nabla_\theta \mathcal{L}(\theta, D_i^{tr})$
4. Update θ using $\nabla_\theta \mathcal{L}(\phi_i, D^{tr})$

Overview

Sources

- Meta-Learning in Neural Networks: A Survey
- Stanford CS 330: Deep Multi-Task and Meta Learning
- Lil'Log Meta-Learning: Learning to Learn Fast
- Tutorial 2: few-shot learning and meta-learning
- ICML 2019 Meta-Learning: from Few-Shot Learning to Rapid Reinforcement Learning
- OpenAi blog on Reptile
- Chelsea Finn's Learning to Learn blog at BAIR

References I

Finn, C., Abbeel, P., and Levine, S. (2017).

Model-agnostic meta-learning for fast adaptation of deep networks.

In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Finn, C. and Levine, S. (2018).

Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm.

In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Gregory Koch, Richard Zemel, R. S. (2015).

Siamese neural networks for one-shot image recognition.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2020).

Meta-learning in neural networks: A survey.

arXiv preprint arXiv:2004.05439.

References II

Munkhdalai, T. and Yu, H. (2017).

Meta networks.

In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2554–2563. PMLR.

Nichol, A., Achiam, J., and Schulman, J. (2018).

On first-order meta-learning algorithms.

CoRR, abs/1803.02999.

Nichol, A. and Schulman, J. (2018).

Reptile: a scalable metalearning algorithm.

Ravi, S. and Larochelle, H. (2017).

Optimization as a model for few-shot learning.

In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

References III

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. (2016).

Meta-learning with memory-augmented neural networks.

In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1842–1850. JMLR.org.

Snell, J., Swersky, K., and Zemel, R. (2017).

Prototypical networks for few-shot learning.

In *Advances in Neural Information Processing Systems 30*, pages 4077–4087.

Vinyals, O. (2017).

Model vs optimization meta learning.

In *NIPS 2017*.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., and Wierstra, D. (2016).

Matching networks for one shot learning.

In *Advances in Neural Information Processing Systems 29*, pages 3630–3638.

Acknowledgements

Special Thanks to

- Prof. Niloy Ganguly
- Prof. Saptarshi Ghosh
- Soumi Das

For their valuable feedback