Optimization-Based Meta-Learning

CS 330
Course Reminders

HW1 due next Weds (10/6).

Some project idea suggestions to be posted today.
Plan for Today

Recap
- Meta-learning problem & black-box meta-learning

Optimization Meta-Learning
- Overall approach
- Compare: optimization-based vs. black-box
- Challenges & solutions
- Case study of land cover classification (time-permitting)

Goals for by the end of lecture:
- Basics of optimization-based meta-learning techniques (& how to implement)
- Trade-offs between black-box and optimization-based meta-learning
Problem Settings Recap

**Multi-Task Learning**
Solve multiple tasks $\mathcal{T}_1, \ldots, \mathcal{T}_T$ at once.

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\theta, \mathcal{D}_i)$$

**Transfer Learning**
Solve target task $\mathcal{T}_b$ after solving source task $\mathcal{T}_a$ by transferring knowledge learned from $\mathcal{T}_a$

---

**The Meta-Learning Problem**
Given data from $\mathcal{T}_1, \ldots, \mathcal{T}_n$, quickly solve new task $\mathcal{T}_{test}$

In transfer learning and meta-learning:
generally impractical to access prior tasks

In all settings: tasks must share structure.
Example Meta-Learning Problem

5-way, 1-shot image classification (MiniImagenet)

Given 1 example of 5 classes:

Classify new examples

Can replace image classification with: regression, language generation, skill learning, any ML problem
Black-Box Adaptation

How else can we represent $\phi_i = f_{\theta}(D^{tr}_i)$?

What if we treat it as an **optimization** procedure?

**+ expressive**

**- challenging optimization problem**

$$y^{ts} = f_{\text{black-box}}(D^{tr}_i, x^{ts})$$
Plan for Today

Recap
- Meta-learning problem & black-box meta-learning

Optimization Meta-Learning
- Overall approach
- Compare: optimization-based vs. black-box
- Challenges & solutions
- Case study of land cover classification (time-permitting)

} Part of Homework 2!
Black-Box Adaptation Optimization-Based Adaptation

\[ f_\theta \]

\[ \phi_i \]

\[ x_{ts} \]

\[ y_{ts} \]

\[ \mathcal{D}_i^{tr} \]
Black-Box Adaptation Optimization-Based Adaptation

Key idea: embed optimization inside the inner learning process

Why might this make sense?
Recall: Fine-tuning

**Fine-tuning**

\[ \phi \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, D_{tr}) \]

- pre-trained parameters
- training data for new task
- (typically for many gradient steps)

Universal Language Model Fine-Tuning for Text Classification. Howard, Ruder. ‘18

Figure 3: Validation error rates for supervised and semi-supervised ULMFiT vs. training from scratch with different numbers of training examples on IMDb, TREC-6, and AG (from left to right).

Fine-tuning less effective with very small datasets.
Optimization-Based Adaptation

**Meta-learning** \( \min \theta \sum_{\text{task } i} \mathcal{L}(\theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, D_{i}^{tr}), D_{i}^{ts}) \)

**Fine-tuning** \( \phi \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, D_{i}^{tr}) \)

**Key idea:** Over many tasks, learn parameter vector \( \theta \) that transfers via fine-tuning

Finn, Abbeel, Levine. Model-Agnostic Meta-Learning. ICML 2017
Optimization-Based Adaptation

\[ \min_{\theta} \sum_{\text{task } i} \mathcal{L}(\theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, D_{i}^{tr}), D_{i}^{ts}) \]

\( \theta \) parameter vector being meta-learned

\( \phi_{i}^{*} \) optimal parameter vector for task i

Model-Agnostic Meta-Learning

Finn, Abbeel, Levine. Model-Agnostic Meta-Learning. ICML 2017
Optimization-Based Adaptation

**Key idea:** Acquire $\phi_i$ through optimization.

**General Algorithm:**

1. Sample task $\mathcal{T}_i$ (or mini batch of tasks)
2. Sample disjoint datasets $\mathcal{D}^{\text{tr}}_i, \mathcal{D}^{\text{test}}_i$ from $\mathcal{D}_i$
3. Compute $\phi_i \leftarrow f_\theta(\mathcal{D}^{\text{tr}}_i)$, Optimize $\phi_i \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{\text{tr}}_i)$
4. Update $\theta$ using $\nabla_\theta \mathcal{L}(\phi_i, \mathcal{D}^{\text{test}}_i)$

---

Black box approach  Optimization-based approach

Do we get higher-order derivatives with more inner gradient steps?

Do we need to compute the full Hessian? 😱

---

*-> whiteboard*
Plan for Today

Recap
- Meta-learning problem & black-box meta-learning

Optimization Meta-Learning
- Overall approach
- **Compare:** optimization-based vs. black-box
- Challenges & solutions
- Case study of land cover classification (time-permitting)

} Part of Homework 2!
Optimization vs. Black-Box Adaptation

Black-box adaptation

**general form:** \( y^{ts} = f_{\text{black-box}}(D_i^{tr}, x^{ts}) \)

Model-agnostic meta-learning

\[
y^{ts} = f_{\text{MAML}}(D_i^{tr}, x^{ts}) \\
= f_{\phi_i}(x^{ts})
\]

where \( \phi_i = \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, D_i^{tr}) \)

MAML can be viewed as **computation graph**, with embedded gradient operator

**Note:** Can mix & match components of computation graph

Learn initialization but replace gradient update with learned network

\[
\phi_i = \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, D_i^{tr})
\]

Ravi & Larochelle ICLR ’17

(actually precedes MAML)

This **computation graph view** of meta-learning will come back again!
Optimization vs. Black-Box Adaptation

How well can learning procedures generalize to similar, but extrapolated tasks?

Omniglot image classification

Does this structure come at a cost?

Finn & Levine ICLR ’18
Black-box adaptation

\[ y^{ts} = f_{\text{black-box}}(D_i^{tr}, x^{ts}) \]

Optimization-based (MAML)

\[ y^{ts} = f_{\text{MAML}}(D_i^{tr}, x^{ts}) \]

Does this structure come at a cost?

For a sufficiently deep network,
MAML function can approximate any function of \( D_i^{tr}, x^{ts} \)

Finn & Levine, ICLR 2018

Assumptions:
- nonzero \( \alpha \)
- loss function gradient does not lose information about the label
- datapoints in \( D_i^{tr} \) are unique

Why is this interesting?
MAML has benefit of inductive bias without losing expressive power.
Plan for Today

Recap
- Meta-learning problem & black-box meta-learning

*Optimization Meta-Learning*  
- Overall approach  
- Compare: optimization-based vs. black-box  
- **Challenges & solutions**  
- Case study of land cover classification (time-permitting)
Optimization-Based Adaptation

**Challenges.** Bi-level optimization can exhibit instabilities.

**Idea:** Automatically learn inner vector learning rate, tune outer learning rate  
(Li et al. Meta-SGD, Behl et al. AlphaMAML)

**Idea:** Optimize only a subset of the parameters in the inner loop  
(Zhou et al. DEML, Zintgraf et al. CAVIA)

**Idea:** Decouple inner learning rate, BN statistics per-step  
(Antoniou et al. MAML++)

**Idea:** Introduce context variables for increased expressive power.  
(Finn et al. bias transformation, Zintgraf et al. CAVIA)

**Takeaway:** a range of simple tricks that can help optimization significantly
Optimization-Based Adaptation

**Challenges.** Backpropagating through many inner gradient steps is compute- & memory-intensive.

**Idea:** [Crudely] approximate \( \frac{d\phi_i}{d\theta} \) as identity (Finn et al. first-order MAML ‘17, Nichol et al. Reptile ‘18)

Surprisingly works for simple few-shot problems, but (anecdotally) not for more complex meta-learning problems.

**Idea:** Only optimize the *last layer* of weights.

*ridge regression, logistic regression* (Bertinetto et al. R2-D2 ’19)

*support vector machine* (Lee et al. MetaOptNet ’19)

\( \rightarrow \) leads to a closed form or convex optimization on top of meta-learned features

**Idea:** Derive meta-gradient using the implicit function theorem (Rajeswaran, Finn, Kakade, Levine. Implicit MAML ’19)

\( \rightarrow \) compute full meta-gradient *without differentiating through optimization path*
Optimization-Based Adaptation

Can we compute the meta-gradient \textit{without differentiating through the optimization path}? 

\textbf{Idea}: Derive meta-gradient using the implicit function theorem

(Rajeswaran, Finn, Kakade, Levine. Implicit MAML)

\begin{itemize}
    \item Memory and computation trade-offs
    \item Allows for second-order optimizers in inner loop
    \item Also useful for hyper parameter optimization
\end{itemize}

(e.g. Lorraine, Vicol, Duvenaud et al. Optimizing Millions of Hyperparameters by Implicit Differentiation ‘20)
Optimization-Based Adaptation

**Challenges.** How to choose architecture that is effective for inner gradient step?

**Idea:** Progressive neural architecture search + MAML  
(Kim et al. Auto-Meta)
- finds highly non-standard architecture (deep & narrow)  
- different from architectures that work well for standard supervised learning

MinImageNet, 5-way 5-shot  
MAML, basic architecture: 63.11%  
MAML + AutoMeta: 74.65%
Optimization-Based Adaptation

**Key idea**: Acquire \( \phi_i \) through optimization.

**Takeaways**: Construct *bi-level optimization* problem.

+ positive inductive bias at the start of meta-learning
+ tends to extrapolate better via structure of optimization
+ maximally expressive with sufficiently deep network
+ model-agnostic (easy to combine with your favorite architecture)
  - typically requires second-order optimization
  - usually compute and/or memory intensive
Plan for Today

Recap
- Meta-learning problem & black-box meta-learning

**Optimization Meta-Learning**
- Overall approach
- Compare: *optimization-based* vs. *black-box*
- Challenges & solutions
- **Case study of land cover classification** (time-permitting)

} Part of Homework 2!
Case Study

Meta-Learning for Few-Shot Land Cover Classification

Marc Rußwurm$^{1,*}$, Sherrie Wang$^{2,3,*}$, Marco Körner$^1$, and David Lobell$^2$

$^1$Technical University of Munich, Chair of Remote Sensing Technology
$^2$Stanford University, Center on Food Security and the Environment
$^3$Stanford University, Institute for Computational and Mathematical Engineering

CVPR 2020 EarthVision Workshop

Problem: Map land covering from satellite images

Applications in global urban planning, climate change research

Challenges: Labeling data is expensive.
Different regions look different & have different land use proportions
Framing land cover mapping as a meta-learning problem

Different tasks: different regions of the world
Goal: Segment/classify images from a new region with a small amount of data

Croplands from four countries.
Framing land cover mapping as a meta-learning problem

**Goal:** Segment/classify images from a new region with a small amount of data

**SEN12MS dataset** (Schmitt et al. 2019)

Geographic meta-data provided

Example 2-way 2-shot classification task
Framing land cover mapping as a meta-learning problem

Goal: Segment/classify images from a new region with a small amount of data

DeepGlobe dataset (Demir et al. 2018)

No geographic metadata, used clustering to guess region
Evaluation

Meta-training data: \{\mathcal{D}_1, \ldots, \mathcal{D}_T\}  
Meta-test time: small amount of data from new region: \mathcal{D}^{tr}_j
(meta-test training set / meta-test support set)

Random init: Train from scratch on \mathcal{D}^{tr}_j

Compare:
- Pre-train on meta-training data \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_T, fine-tune on \mathcal{D}^{tr}_j
- MAML on meta-training data \{\mathcal{D}_1, \ldots, \mathcal{D}_T\}, adapt with \mathcal{D}^{tr}_j

SEN12MS dataset
DeepGlobe dataset

More visualizations and analysis in the paper!
Plan for Today

Recap
- Meta-learning problem & black-box meta-learning

Optimization Meta-Learning
- Overall approach
- Compare: optimization-based vs. black-box
- Challenges & solutions
- Case study of land cover classification (time-permitting)

Goals for by the end of lecture:
- Basics of optimization-based meta-learning techniques (& how to implement)
- Trade-offs between black-box and optimization-based meta-learning

} Part of Homework 2!
Roadmap for upcoming lectures

Next week:
- **Monday:** Non-parametric few-shot learners, comparison of approaches
- **Wednesday:** Advanced (but important!) meta-learning topics
- **Following Monday:** Bayesian meta-learning

Week 4: Start of reinforcement learning topics [project proposals due]
Course Reminders

HW1 due next Weds (10/6).

Some project idea suggestions to be posted today.