Model-Agnostic Meta-Learning
Universality, Inductive Bias, and Weak Supervision

Chelsea Finn
Why Learn to Learn?

- effectively reuse data on other tasks
- replace manual engineering of architecture, hyperparameters, etc.
- learn to quickly adapt to unexpected scenarios (inevitable failures, long tail)
- learn how to learn with weak supervision
<table>
<thead>
<tr>
<th>Problem Domains:</th>
<th>Approaches:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- few-shot classification &amp; generation</td>
<td>- recurrent networks</td>
</tr>
<tr>
<td>- hyperparameter optimization</td>
<td>- learning optimizers or update rules</td>
</tr>
<tr>
<td>- architecture search</td>
<td>- learning initial parameters &amp;</td>
</tr>
<tr>
<td>- faster reinforcement learning</td>
<td>architecture</td>
</tr>
<tr>
<td>- domain generalization</td>
<td>- acquiring metric spaces</td>
</tr>
<tr>
<td>- learning structure</td>
<td>- Bayesian models</td>
</tr>
<tr>
<td>- ...</td>
<td>- ...</td>
</tr>
</tbody>
</table>

What is the meta-learning problem statement?
The Meta-Learning Problem

Supervised Learning:

Inputs: $\mathbf{x}$  
Outputs: $\mathbf{y}$  
Data: $\{(\mathbf{x}, \mathbf{y})_i\}$  

$\mathbf{y} = f(\mathbf{x}; \theta)$

Meta-Supervised Learning:

Inputs: $\mathcal{D}_{\text{train}}$, $\mathbf{x}_{\text{test}}$  
Outputs: $\mathbf{y}_{\text{test}}$  
Data: $\{\mathcal{D}_i\}$  

$\mathbf{y}_{\text{test}} = f(\mathcal{D}_{\text{train}}, \mathbf{x}_{\text{test}}; \theta)$

Why is this view useful?
Reduces the problem to the design & optimization of $f$. 

Chelsea Finn, UC Berkeley
Design of $f$?

Recurrent network (LSTM, NTM, Conv)

$y_{test} = f(D_{train}, x_{test}; \theta)$

Santoro et al. '16, Duan et al. '17, Wang et al. '17, Munkhdalai & Yu '17, Mishra et al. '17, …
Design of $f$?

### Recurrent network
(LSTM, NTM, Conv)

$$y_{test} = f(D_{train}, x_{test}; \theta)$$

Santoro et al. '16, Duan et al. '17, Wang et al. '17, Munkhdalai & Yu '17, Mishra et al. '17, …

### Learned optimizer
(often uses recurrence)

$$y_{test} = f(x_{test}; g(D_{train}; \theta))$$

Schmidhuber et al. '87, Bengio et al. '90, Hochreiter et al. '01, Li & Malik '16, Andrychowicz et al. '16, Ha et al. '17, Ravi & Larochelle '17, …

---

Chelsea Finn, UC Berkeley
Design of $f$?

Recurrent network (LSTM, NTM, Conv)

$y_{test} = f(D_{train}, x_{test}; \theta)$


Learned optimizer (often uses recurrence)

$y_{test} = f(x_{test}; g(D_{train}; \theta))$


These approaches are general and quite powerful.

What happens when the task is very different? Or very little meta-training?

Impose Structure


Can we build a general meta-learning algorithm that interpolates between learning from scratch and few-shot learning?
**Model-Agnostic Meta-Learning**

**Key idea**: Train over many tasks, to learn parameter vector $\theta$ that transfers

**In-distribution task**: k-shot learning

**Base case**: learning from scratch

**Related but out-of-distribution task**: somewhere in between

\[
\min_{\theta} \sum_{\text{tasks}} \mathcal{L}_v(\theta - \alpha \nabla_\theta \mathcal{L}_{tr}(\theta))
\]

**Fine-tuning**: $\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$
Design of $f$?

**Recurrent network** (LSTM, NTM, Conv)

$y_{\text{test}} = f(D_{\text{train}}, x_{\text{test}}; \theta)$


**Learned optimizer** (often uses recurrence)

$y_{\text{test}} = f(x_{\text{test}}; g(D_{\text{train}}; \theta))$


**Impose Structure**


**MAML** (learned initialization)

$y_{\text{test}} = f(x_{\text{test}}; \theta - \alpha \nabla_\theta \mathcal{L}(D_{\text{train}}))$

Finn et al. ’17, Grant et al. ’17, Reed et al. ’17, Li et al. ’17, …

Chelsea Finn, UC Berkeley
Theoretical & Empirical Questions

1. What happens when MAML faces **out-of-distribution tasks**?
2. How **expressive** are deep representations + gradient descent?
3. Can we interpret MAML in a **probabilistic framework**?
4. Can we use MAML to learn from **weak supervision**?
How well can methods generalize to similar, but extrapolated tasks?

The world is non-stationary.

Omniglot image classification

MAML  TCML, MetaNetworks

Finn & Levine ’17 (under review)
How well can methods generalize to similar, but extrapolated tasks?

The world is non-stationary.

**Sinusoid curve regression**

Takeaway: Strategies learned with MAML consistently generalize better to out-of-distribution tasks
Theoretical & Empirical Questions

1. What happens when MAML faces out-of-distribution tasks?
2. How expressive are deep representations + gradient descent?
3. Can we interpret MAML in a probabilistic framework?
4. Can we use MAML to learn from weak supervision?
Universal Function Approximation Theorem
Hornik et al. ’89, Cybenko ’89, Funahashi ‘89

A neural network with one hidden layer of finite width can approximate any continuous function.
\[ y = f(x; \theta) \]
“universal function approximator”

How can we define a notion of universality / expressive power for meta-learning?
\[ y_{\text{test}} = f(D_{\text{train}}, x_{\text{test}}; \theta) \]
“universal learning procedure approximator”

Recurrent network
\[ y_{\text{test}} = f(D_{\text{train}}, x_{\text{test}}; \theta) \]
Learned optimizer
\[ y_{\text{test}} = f(x_{\text{test}}, g(D_{\text{train}}; \theta)) \]

With sufficient depth, both are universal learning procedure approximators.

Are we losing expressive power when using MAML?
How expressive is MAML?

\[ y_{test} = f(x_{test}; \theta - \alpha \nabla_{\theta} \mathcal{L}(D_{train}) ) \]

Assumptions:
- cross entropy or mean-squared error loss
- datapoints \( x_i \) in training dataset are unique

Result: For a sufficiently deep \( f_\theta \),

\[ f(x_{test}; \theta - \alpha \nabla_{\theta} \mathcal{L}(D_{train})) \]

is a universal learning procedure approximator.

[It can approximate any function of \( D_{train} \) \( x_{test} \)]

Why is this interesting?
MAML has both benefits of inductive bias and expressive power.
Theoretical & Empirical Questions

1. What happens when MAML faces out-of-distribution tasks?
2. How expressive is deep representation + gradient descent?
3. Can we interpret MAML in a probabilistic framework?
4. Can we use MAML to learn from weak supervision?
Can we interpret MAML in a probabilistic framework?

meta-learning $\approx$ learning a prior

*Bayesian concept learning*

[Tenenbaum ’99, Fei-Fei et al. ’03, Lawrence & Platt ’04, …]

formulate few-shot learning as probabilistic inference problem

+ can effectively generalize from limited evidence

- hard to scale to complex models
Can we interpret MAML in a probabilistic framework?

Bayesian meta-learning approach

Task-specific parameters

\[ \phi_j \rightarrow x_{jn} \rightarrow \theta \]

Meta-parameters

\[ \theta \]

How to compute MAP estimate?

\[
\max_{\theta} \prod_j p(D^{(j)}_{\text{train}} | \theta) \\
= \prod_j \int p(D_{\text{train}}^{(j)} | \phi_j)p(\phi_j | \theta) d\phi_j \\
\approx \prod_j p(D_{\text{train}}^{(j)} | \hat{\phi}_j)p(\hat{\phi}_j | \theta)
\]

How to compute MAP estimate?

Gradient descent with early stopping = MAP inference under Gaussian prior with mean at initial parameters [Santos ‘96]

(exact in linear case, approximate in nonlinear case)

MAML approximates hierarchical Bayesian inference. [Grant et al. ‘17]
Theoretical & Empirical Questions

1. What happens when MAML faces out-of-distribution tasks?
2. How expressive is deep representation + gradient descent?
3. Can we interpret MAML in a probabilistic framework?
4. Can we use MAML to learn from weak supervision?
Learning to Learn from Weak Supervision

Meta-Supervised Learning:

Inputs: $D_{\text{train}}$, $x_{\text{test}}$  
Outputs: $y_{\text{test}}$  
Data: $\{D_i\}$

During meta-training: access full supervision for each task
During meta-testing: only use weakly-supervised datapoints

With MAML: $\min_\theta \sum L_v(\theta - \alpha \nabla_\theta L_{\text{tr}}(\theta))$

Key insight: inner loss can be different than outer loss
Weak Supervision Results

- **Learning from positive examples**
  Grant, Finn, Peterson, Abbott, Levine, Darrell, Griffiths, NIPS ‘17 CIAI workshop

- **One-shot Imitation from human video**
  (in preparation, with Yu, Abbeel, Levine)
Typical Objective of Few-Shot Learning

**Image recognition**
Given 1 example of 5 classes:
Classify new examples

**Human Concept Learning**
Given 1 *positive example*:
Classify new examples:

Beyond how humans learn, this setting is also more interesting.
Human Concept Learning

Given 1 positive example:

Classify new examples:

- only positive examples
- both positive & negatives

Why does this make sense?

MAML approximates hierarchical Bayesian inference

Concept Acquisition through Meta-Learning (CAML)

Grant et al. ’17 (NIPS CIAI workshop)
Few-Shot Image Classification from Positive Examples

MinImagenet dataset

- **MatchNets** (Vinyals et al., 2016)
- **ProtoNets** (Snell et al., 2017)
- **CAML** (Our method)

Grant et al. '17 (NIPS CIAI workshop)
One-Shot **Visual** Imitation Learning

**Goal:** Given one visual demonstration of a new task, learn a policy

Visual imitation is expensive.

*behavior cloning / supervised learning*

Rahmanizadeh et al. '17  Zhang et al. '17

learns from raw pixels,
but requires many demonstrations

**Through meta-learning:** reuse data from other tasks/objects/environments

No direct supervision signal in video of human.

Yu*, Finn*, et al. (in prep.)
One-Shot Visual Imitation from Humans

\[ \mathcal{L} = \sum_{t} ||\pi_{\theta}(o_t) - a_{t}^*||^2 \]

\textit{meta-training time} \quad \min_{\theta} \sum_{\text{tasks}} \mathcal{L}_v(\theta - \alpha \nabla_{\theta} \mathcal{L}_{tr}(\theta))

\textit{meta-training tasks}

\textit{val demo} (robot demo) \quad \textit{training demo} (video of human)

\textit{meta-test time} \quad \theta' \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta)

\textit{demo of meta-test task} (video of human)

Yu*, Finn*, et al. (in prep.)
On-going work: One-shot imitation from human video

input human demo

resulting policy

Yu*, Finn*, et al. (in prep.)
Takeaways

• Meta-learning can be seen as learning a function
  \( D_{\text{train}} \xrightarrow{\text{x}_{\text{test}}} y_{\text{test}} \)

• Embedding gradient descent provides beneficial inductive bias while maintaining universality

• MAML is equivalent to empirical Bayes

• Can learn how to learn from “weak” supervision
  From 1 positive example: From a video of a human:

  ![Image of a lion](lion.png)  ![Image of a human](human.png)
Collaborators

Sergey Levine
Pieter Abbeel
Tianhe Yu
Tianhao Zhang
Erin Grant
Josh Abbott
Tom Griffiths
Josh Peterson
Trevor Darrell

Blog post, code, and papers: eecs.berkeley.edu/~cbfinn

Questions?

Chelsea Finn, UC Berkeley