Definition of Meta Learning

- What is Meta Learning / Learning to Learn?
  - Go beyond train from samples from a single distribution.
  - Distribution over tasks, so model has to “learn to learn” when a new task is presented.

“... a system that improves or discovers a learning algorithm”
Hochreiter et al, ‘01
Datasets: Omniglot

- To make progress, we need datasets / metrics!

Lake et al, 2013, 2015
Datasets: Mini-ImageNet

- To make progress, we need datasets / metrics!

Vinyals et al, 2016
Datasets: Beyond

- To make progress, we need datasets / metrics!

Reinforcement learning
Given a small amount of experience

How? learn to learn many other tasks

Solve a new task

fig. from Duan et al. ’17

Yan Duan, Marcin Andrychowicz, Bradly Stadie, Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, Wojciech Zaremba (2017)
1. Sample label set $L$ from $T$

$L = \text{[Pinscher, Golden Retriever, Husky, German Shepherd]}$
Training Setup: An “Episode”

1. Sample label set $\mathbf{L}$ from $\mathbf{T}$
2. Sample a few images as support set $\mathbf{S}$ from $\mathbf{L}$
3. Sample a few images as batch $\mathbf{B}$ from $\mathbf{L}$

$L = \{\text{Pinscher, Golden Retriever, Husky, German Shepherd}\}$
Training Setup: An “Episode”

1. Sample label set $\mathbf{L}$ from $\mathbf{T}$
2. Sample a few images as support set $\mathbf{S}$ from $\mathbf{L}$
3. Sample a few images as batch $\mathbf{B}$ from $\mathbf{L}$
4. Optimize batch, Go to 1

$$\theta = \arg \max_\theta E_{\mathbf{L} \sim \mathbf{T}} \left[ E_{\mathbf{S} \sim \mathbf{L}, \mathbf{B} \sim \mathbf{L}} \left[ \sum_{(x,y) \in \mathbf{B}} \log P_\theta (y | x, \mathbf{S}) \right] \right].$$

$L = \{\text{Pinscher, Golden Retriever, Husky, German Shepherd}\}$
Contrasting with Supervised Learning

\[ \theta = \arg \max_{\theta} \left[ \mathbb{E}_B \left[ \sum_{(x,y) \in B} \log P_{\theta}(y|x) \right] \right] . \]
Contrasting with Supervised Learning

\[ \theta = \arg \max_{\theta} \mathbb{E}_{L \sim T} \left[ E_{S \sim L, B \sim L} \left[ \sum_{(x, y) \in B} \log P_{\theta}(y|x, S) \right] \right]. \]
Contrasting with Supervised Learning

\[ \theta = \arg \max_{\theta} E_{L \sim T} \left[ E_{S \sim L, B \sim L} \left[ \sum_{(x, y) \in B} \log P_{\theta}(y|x, S) \right] \right] . \]
Meta Learning Models Taxonomy

Model Based

- Santoro et al. ’16
- Duan et al. ’17
- Wang et al. ‘17
- Munkhdalai & Yu ‘17
- Mishra et al. ‘17

Metric Based

- Koch ’15
- Vinyals et al. ‘16
- Snell et al. ‘17
- Shyam et al. ‘17
- Sung et al. ‘17

Optimization Based

- Schmidhuber ’87, ’92
- Bengio et al. ’90, ’92
- Hochreiter et al. ’01
- Li & Malik ’16
- Andrychowicz et al. ’16
- Ravi & Larochelle ‘17
- Finn et al. ‘17

Adapted from Finn ‘17
Model Based Meta Learning

\[ P_\theta(y|x, S) = f_\theta(x, S) \]
Metric Based Meta Learning

\[ P_\theta(y \mid x, S) = \sum_{(x_i, y_i) \in S} k_\theta(x, x_i) y_i \]

\[ Y = \sum_{i=1}^{4} \frac{e^{c(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{4} e^{c(f(\hat{x}), g(x_j))}} y_i \]
Metric Based Meta Learning

Matching Networks, Vinyals et al, NIPS 2016

\[ P_\theta(y|x, S) = \sum_{(x_i, y_i) \in S} k_\theta(x, x_i, S) y_i \]
Optimization Based Meta Learning

\[ P_{\theta}(y \mid x, S) = f_{\theta}(S)(x) \]

\[ \theta(S) = g_{\theta_g}(\theta_0, \{\nabla_{\theta_0} L(x_i, y_i)\}_{(x_i, y_i) \in S}) \]

Figure Credit: Hugo Larochelle
Examples of Optimization Based Meta Learning

Finn et al, 17

\[ \theta = \theta_0 - \eta \sum_{(x_i, y_i) \in S} \nabla_{\theta_0} L(x_i, y_i) \]

Ravi et al, 17

\[ \theta_t = f_t \circ \theta_{t-1} + i_t \circ \nabla_{\theta_{t-1}} L(x_t, y_t) \]

\[ P_{\theta}(y|x, S) = f_{\theta(S)}(x) \]

\[ \theta(S) = g_{\theta_g}(\theta_0, \{\nabla_{\theta_0} L(x_i, y_i)\}_{(x_i, y_i) \in S}) \]
## Progress on Mini-ImageNet

<table>
<thead>
<tr>
<th>Model</th>
<th>Optim</th>
<th>Metric</th>
<th>FT</th>
<th>5-way Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1-shot</td>
</tr>
<tr>
<td><strong>Metric</strong></td>
<td>→</td>
<td>→</td>
<td>N</td>
<td>43.56 ± 0.84%</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>→</td>
<td>→</td>
<td>N</td>
<td>49.21 ± 0.96%</td>
</tr>
<tr>
<td><strong>Optim</strong></td>
<td>→</td>
<td>→</td>
<td>N</td>
<td>43.44 ± 0.77%</td>
</tr>
<tr>
<td><strong>Optim</strong></td>
<td>→</td>
<td>→</td>
<td>Y</td>
<td>48.70 ± 1.84%</td>
</tr>
<tr>
<td><strong>Metric</strong></td>
<td>→</td>
<td>→</td>
<td>N</td>
<td>49.42 ± 0.78%</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>→</td>
<td>→</td>
<td>N</td>
<td>51.38 ± 0.82%</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>→</td>
<td>→</td>
<td>Y</td>
<td>55.71 ± 0.99%</td>
</tr>
<tr>
<td><strong>Metric</strong></td>
<td>→</td>
<td>→</td>
<td>N</td>
<td><strong>57.02 ± 0.92%</strong></td>
</tr>
</tbody>
</table>

Table from Sung et al, 17
Summing Up

Model Based

\[ P_{\theta}(y|x, S) = f_{\theta}(x, S) \]

Metric Based

\[ P_{\theta}(y|x, S) = \sum_{(x_i, y_i) \in S} k_{\theta}(x, x_i) y_i \]

Optimization Based

\[ \theta(S) = g_{\theta_0}(\theta_0, \{ \nabla_{\theta_0} L(x_i, y_i) \}_{(x_i, y_i) \in S}) \]
Future Work

- Combining Model / Metric / Optimization based approaches
  - Reed et al, 2017

- Meta-Meta-Meta... learning
  - Tasks need to be related / from same distribution

- What are the right inductive biases?
  - Spatial invariance $\rightarrow$ convolution
  - Temporal sequences $\rightarrow$ recurrence
  - Learning $\rightarrow$ gradients?
Thanks!! Questions??

@OriolVinyalsML
NIPS, December 2017