Meta-learning with an Adaptive Task Scheduler

Huaxiu Yao¹, Yu Wang², Ying Wei³, Peilin Zhao⁴
Mehrdad Mahdavi⁵, Defu Lian², Chelsea Finn¹

¹Stanford University, ²University of Science and Technology of China
³City University of Hong Kong, ⁴Tencent AI Lab, ⁵Pennsylvania State University
Background: Gradient-based Meta-learning

Meta-learning

Task Pool

Task Pool

Gradient-based meta-learning

ML model: $f$ with initial parameter $\theta_0$

Prior

New Task

Random Start

Global Prior $\theta_0$

Adaptation

Best for Task $\mathcal{T}_1: (\mathcal{D}_1^s, \mathcal{D}_1^q)$

Best for Task $\mathcal{T}_2: (\mathcal{D}_2^s, \mathcal{D}_2^q)$

$\mathcal{D}_1^s$: support set of task $\mathcal{T}_i$

$\mathcal{D}_1^q$: query set of task $\mathcal{T}_i$
Uniform Task Sampling

Ideal Scenario

Real Scenario

Drug discovery
• Each assay is a task
• Noisy tasks caused by improper measurement

Some tasks are less valuable or contain noises

Require non-uniform sampling
Non-adaptive task schedulers

Adjusting class sampling strategies
[Liu et al. 2020]

Ranking tasks based on the amount of their information [Sæmundsson et al. 2020]

+ Benefit meta-learning process with a task scheduler
- Require manually strategy design
- The task scheduler can not adapt to the learning progress of the meta-model
Adaptive Task Sampling (ATS)

Goal: determining task sampling probability via a neural scheduler

Task Pool → Candidate Tasks → Meta-model $\theta_0$ → Information covered in candidate tasks → Neural Scheduler $\phi$

Sample tasks via sampling probability $W$
Information covered in candidate tasks – Two meta-model-related factors

1. Loss $\mathcal{L}(\mathcal{D}_i^q, \theta_i^{(0/k)})$ on the query set
2. Gradient similarity between the support and query sets

Motivation
- large query losses + large gradient similarities $\rightarrow$ true hard tasks
- large query losses + small gradient similarities $\rightarrow$ tasks with noise

$$w_i^{(k)} = g \left( \mathcal{L}(\mathcal{D}_i^q, \theta_i^{(k)}), \left\langle \nabla_{\theta_0^{(k)}} \mathcal{L}(\mathcal{D}_i^s, \theta_0^{(k)}), \nabla_{\theta_0^{(k)}} \mathcal{L}(\mathcal{D}_i^q, \theta_0^{(k)}) \right\rangle ; \phi^{(k)} \right)$$
How to Optimize?

Optimize neural scheduler and meta-model alternatively

Step 1: Obtain the temporal meta-model
How to Optimize?

Step 2: Use validation tasks to optimize the neural scheduler

Use REINFOCE to Update $\phi$

$$\phi^{(k+1)} \leftarrow \phi^{(k)} - \gamma \nabla_{\phi^{(k)}} \log P(W^{(k)}; \phi^{(k)})(\frac{1}{N_v} \sum_{i=1}^{N_v} R_i^{(k)} - b)$$
How to Optimize?

Step 3: Update meta-model $\theta_0$

Sampled tasks → Meta-model $\theta_0$ → Update $\theta_0$ → Training Loss $\mathcal{L}_{tr}$ → Neural Scheduler $\phi$ → Sample tasks via sampling probability $W$

$\langle \nabla_{\theta_0} \mathcal{L}(\mathcal{D}_i^s; \theta_0), \nabla_{\theta_0} \mathcal{L}(\mathcal{D}_i^q; \theta_0) \rangle$

$\mathcal{L}(\mathcal{D}_i^q; \theta_0)$
Overall Framework

- Sampled tasks
- Meta-model $\theta_0$
- Training Loss $L_{tr}$
- Update $\theta_0$
- Neural Scheduler $\phi$
- Obtain Temporal meta-model $\tilde{\theta}_0$
- Val Reward
- Temporal Meta-model $\tilde{\theta}_0$
- Val Tasks
- Sample tasks via sampling probability $W$
- Obtain meta-model related factors
- Optimize neural scheduler
- Optimize meta-model
How does ATS Improves Meta-training Process?

**Proposition 1.** Suppose that \( w = [w_1, \cdots, w_{N_{pool}}] \) denotes the random variable for sampling probabilities, \( L_{\theta_0} = [L(D_1^q; \theta_0), \cdots, L(D_{N_{pool}}^q; \theta_0)] \) denotes the random variable for the loss using the meta-model, and \( \nabla \theta_0 = [(\nabla \theta_0 L(D_1^q; \theta_0), \nabla \theta_0 L(D_1^q; \theta_0)), \cdots, (\nabla \theta_0 L(D_{N_{pool}}^q; \theta_0), \nabla \theta_0 L(D_{N_{pool}}^q; \theta_0))] \) denotes the random variable for the inner product between gradients of the support and query sets with respect to the meta-model. Then the following equation connecting the meta-learning losses with and without the task scheduler holds:

\[
L^w(\theta_0) = L(\theta_0) + \text{Cov}(L_{\theta_0}, w) - \alpha \text{Cov}(\nabla \theta_0, w).
\]

sampling probability negatively correlated with loss + positively correlated with gradient similarity

ATS improves the meta-training loss

**Proposition 2.** With the sampling probability defined as

\[
w^*_i = \frac{e^{-L(D_i^q; \theta_0^*) - \alpha \langle \nabla \theta_0 L(D_i^q; \theta_0^*), \nabla \theta_0 L(D_i^q; \theta_0^*) \rangle}}{\sum_{i=1}^{N_{pool}} e^{-L(D_i^q; \theta_0^*) - \alpha \langle \nabla \theta_0 L(D_i^q; \theta_0^*), \nabla \theta_0 L(D_i^q; \theta_0^*) \rangle}},
\]

the following hold:

\( \forall \theta_0 : \text{Cov}(L_{\theta_0} - \alpha \nabla \theta_0, e^{-(L_{\theta_0^*} - \alpha \nabla \theta_0^*)}) \geq 0, \)

\( L^w(\theta_0) - L^w(\theta_0^*) \geq L(\theta_0) - L(\theta_0^*), \)

\( \forall \theta_0 : \text{Cov}(L_{\theta_0} - \alpha \nabla \theta_0, e^{-(L_{\theta_0^*} - \alpha \nabla \theta_0^*)}) \leq -\text{Var}(L_{\theta_0^*} - \alpha \nabla \theta_0^*), \)

\( L^w(\theta_0) - L^w(\theta_0^*) \leq L(\theta_0) - L(\theta_0^*). \)

---

Speed up training

The minima tends to be flat with better generalization ability
Experiments: Meta-learning with Noise

• Create noisy tasks
  • Add noises on the support set of each task – noisy support set + clean query set

• Two datasets
  • miniImagenet – classify the category of each image
  • Drug – predict the activity of each drug compound (regression)

<table>
<thead>
<tr>
<th>Model</th>
<th>miniImagenet-noisy 5-way 1-shot</th>
<th>miniImagenet-noisy 5-way 5-shot</th>
<th>Drug-noisy mean</th>
<th>Drug-noisy medium</th>
<th>&gt;0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>41.67 ± 0.80%</td>
<td>55.80 ± 0.71%</td>
<td>0.202</td>
<td>0.113</td>
<td>21</td>
</tr>
<tr>
<td>SPL</td>
<td>42.13 ± 0.79%</td>
<td>56.19 ± 0.70%</td>
<td>0.211</td>
<td>0.138</td>
<td>24</td>
</tr>
<tr>
<td>FocalLoss</td>
<td>41.91 ± 0.78%</td>
<td>53.58 ± 0.75%</td>
<td>0.205</td>
<td>0.106</td>
<td>23</td>
</tr>
<tr>
<td>GCP</td>
<td>41.86 ± 0.75%</td>
<td>54.63 ± 0.72%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>PAML</td>
<td>41.49 ± 0.74%</td>
<td>52.45 ± 0.69%</td>
<td>0.204</td>
<td>0.120</td>
<td>24</td>
</tr>
<tr>
<td>DAML</td>
<td>41.26 ± 0.73%</td>
<td>55.46 ± 0.70%</td>
<td>0.197</td>
<td>0.113</td>
<td>24</td>
</tr>
<tr>
<td>ATS (Ours)</td>
<td><strong>44.21 ± 0.76%</strong></td>
<td><strong>59.50 ± 0.71%</strong></td>
<td><strong>0.233</strong></td>
<td><strong>0.152</strong></td>
<td><strong>31</strong></td>
</tr>
</tbody>
</table>

* means the result are significant according to Student’s T-test at level 0.01 compared to SPL
Ablation Study about Meta-model-related Factors

• Sim – gradient similarity
• Loss – loss on the query set
• Reweighting – change sampling probabilities to task weights

<table>
<thead>
<tr>
<th>Ablation Model</th>
<th>miniImagenet-noisy</th>
<th>Drug-noisy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5-way 1-shot</td>
<td>5-way 5-shot</td>
</tr>
<tr>
<td>Random $\phi$</td>
<td>41.95 ± 0.80%</td>
<td>56.07 ± 0.71%</td>
</tr>
<tr>
<td>Rank by Sim/Loss</td>
<td>42.84 ± 0.76%</td>
<td>57.90 ± 0.68%</td>
</tr>
<tr>
<td>$\phi$+Loss</td>
<td>42.45 ± 0.80%</td>
<td>56.65 ± 0.75%</td>
</tr>
<tr>
<td>$\phi$+Sim</td>
<td>42.28 ± 0.82%</td>
<td>56.71 ± 0.72%</td>
</tr>
<tr>
<td>Reweighting</td>
<td>42.19 ± 0.80%</td>
<td>56.48 ± 0.72%</td>
</tr>
<tr>
<td>ATS ($\phi$+Loss+Sim)</td>
<td><strong>44.21 ± 0.76%</strong></td>
<td><strong>59.50 ± 0.71%</strong></td>
</tr>
</tbody>
</table>

* means the result is significant according to Student’s T-test at level 0.01 compared to Weighting
## Effect of Noise Ratio

More noises ➔ more improvements

<table>
<thead>
<tr>
<th>Noise Ratio</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>43.46 ± 0.82%</td>
<td>42.92 ± 0.78%</td>
<td>41.67 ± 0.80%</td>
<td>36.53 ± 0.73%</td>
</tr>
<tr>
<td>BNS</td>
<td>44.04 ± 0.81%</td>
<td>43.36 ± 0.75%</td>
<td>42.13 ± 0.79%</td>
<td>38.21 ± 0.75%</td>
</tr>
<tr>
<td><strong>ATS (Ours)</strong></td>
<td>45.55 ± 0.80%</td>
<td>44.50 ± 0.86%</td>
<td>44.21 ± 0.76%</td>
<td>42.18 ± 0.73%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noise Scaler</th>
<th>η=2</th>
<th>η=4</th>
<th>η=6</th>
<th>η=8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drug</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>0.222</td>
<td>0.202</td>
<td>0.196</td>
<td>0.194</td>
</tr>
<tr>
<td>BNS</td>
<td>0.229</td>
<td>0.211</td>
<td>0.208</td>
<td>0.200</td>
</tr>
<tr>
<td><em><em>ATS</em> (Ours)</em>*</td>
<td>0.235</td>
<td>0.233</td>
<td>0.221</td>
<td>0.219</td>
</tr>
</tbody>
</table>

* means all results are significant according to Student’s T-test at level 0.01 compared to BNS
Analysis of the Meta-model-related Factors

High losses + low gradient similarities $\rightarrow$ noisy tasks
Experiments: Meta-learning with Limited Budgets

• Goal: identify the most useful tasks
• Datasets
  • miniImagenet – less meta-training classes means less budgets
  • Drug – only 4,100 tasks in the whole dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>miniImagenet-Limited</th>
<th>Drug-Full</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5-way 1-shot</td>
<td>5-way 5-shot</td>
<td>mean</td>
</tr>
<tr>
<td>Uniform</td>
<td>33.61 ± 0.66%</td>
<td>45.97 ± 0.65%</td>
<td>0.233</td>
</tr>
<tr>
<td>SPL</td>
<td>34.28 ± 0.65%</td>
<td>46.05 ± 0.69%</td>
<td>0.232</td>
</tr>
<tr>
<td>FocalLoss</td>
<td>33.11 ± 0.65%</td>
<td>46.12 ± 0.70%</td>
<td>0.229</td>
</tr>
<tr>
<td>GCP</td>
<td>34.69 ± 0.67%</td>
<td>46.86 ± 0.68%</td>
<td>N/A</td>
</tr>
<tr>
<td>PAML</td>
<td>33.64 ± 0.62%</td>
<td>45.01 ± 0.69%</td>
<td>0.238</td>
</tr>
<tr>
<td>DAML</td>
<td>34.83 ± 0.69%</td>
<td>46.66 ± 0.67%</td>
<td>0.227</td>
</tr>
<tr>
<td>ATS (Ours)</td>
<td>35.15 ± 0.67%</td>
<td>47.76 ± 0.68%</td>
<td>0.252*</td>
</tr>
</tbody>
</table>

* means the result is significant according to Student’s T-test at level 0.01 compared to PAML
Analysis of the Meta-model-related Factors

High losses + high gradient similarities → more valuable tasks
Effect of the Budgets

Less meta-training tasks → more improvements

<table>
<thead>
<tr>
<th>Budgets</th>
<th>16</th>
<th>32</th>
<th>48</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>33.61 ± 0.66%</td>
<td>40.48 ± 0.75%</td>
<td>44.07 ± 0.80%</td>
<td>45.73 ± 0.79%</td>
</tr>
<tr>
<td>GCP</td>
<td>34.69 ± 0.67%</td>
<td>41.27 ± 0.74%</td>
<td>44.30 ± 0.79%</td>
<td>45.35 ± 0.81%</td>
</tr>
<tr>
<td>ATS (Ours)</td>
<td><strong>35.15 ± 0.67%</strong></td>
<td><strong>41.68 ± 0.78%</strong></td>
<td><strong>44.89 ± 0.79%</strong></td>
<td><strong>46.27 ± 0.80%</strong></td>
</tr>
</tbody>
</table>
Takeaways & Next

• Adaptive task sampling strategies improves the meta-training process

• Both query loss and gradient similarity are important factors in ATS

• What’s Next?
  • Incorporate task scheduler with sample scheduler
  • Reduce the computational cost
Thanks

Q & A