Meta-learning with an Adaptive Task Scheduler

Huaxiu Yao[1], Yu Wang[2], Ying Wei[3], Peilin Zhao[4]

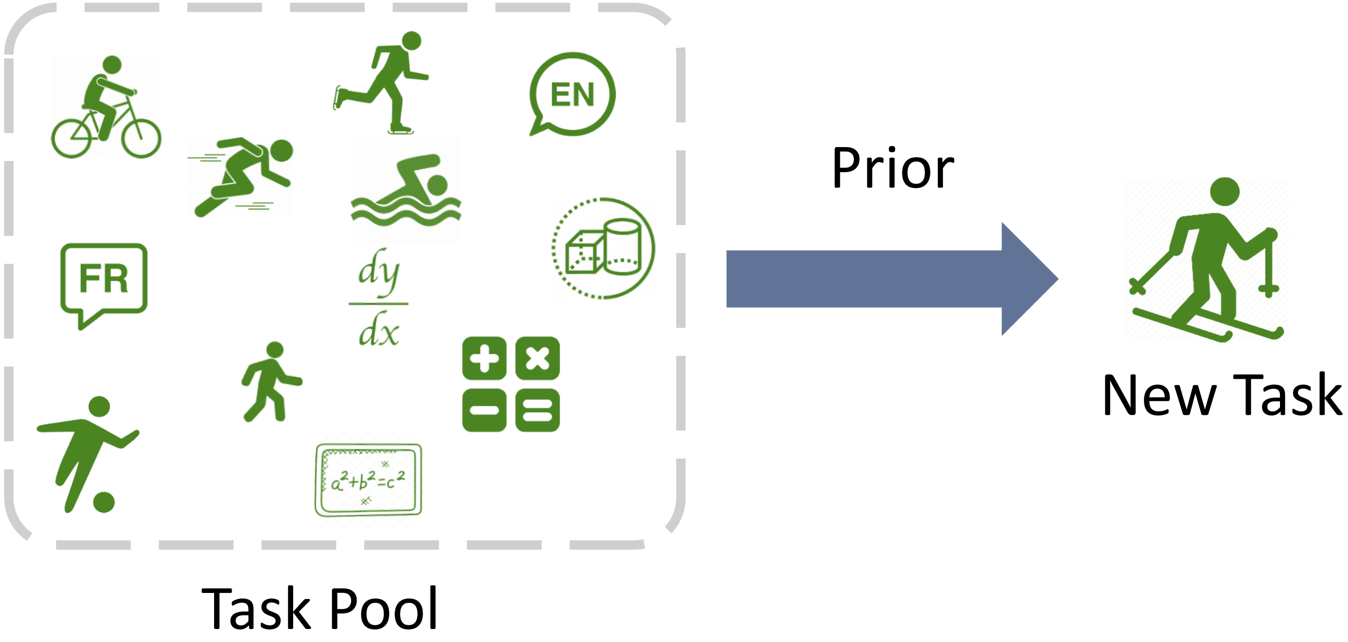
Mehrdad Mahdavi[5], Defu Lian[2], Chelsea Finn[1]

[1]Stanford University, [2]University of Science and Technology of China

[3]City University of Hong Kong, [4]Tencent AI Lab, [5]Pennsylvania State University

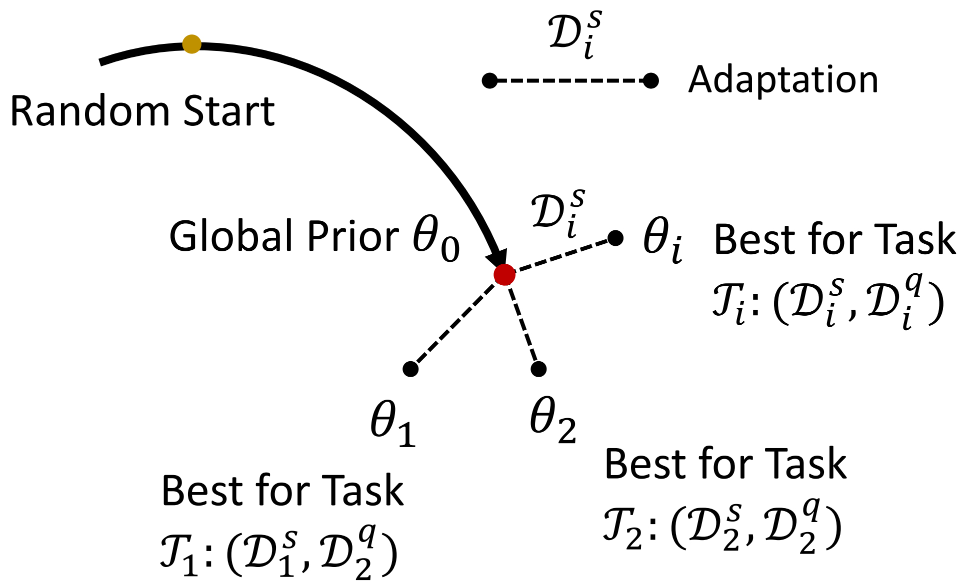
Background: Gradient-based Meta-learning

Meta-learning



Gradient-based meta-learning

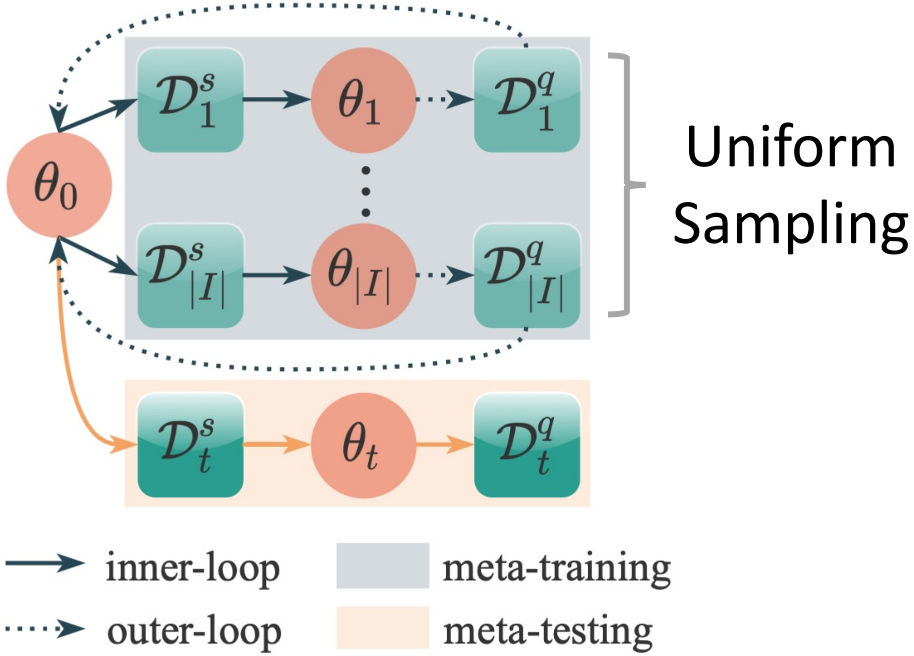
ML model: f with initial parameter θ_0



$\mathcal{D}_1^s$: 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

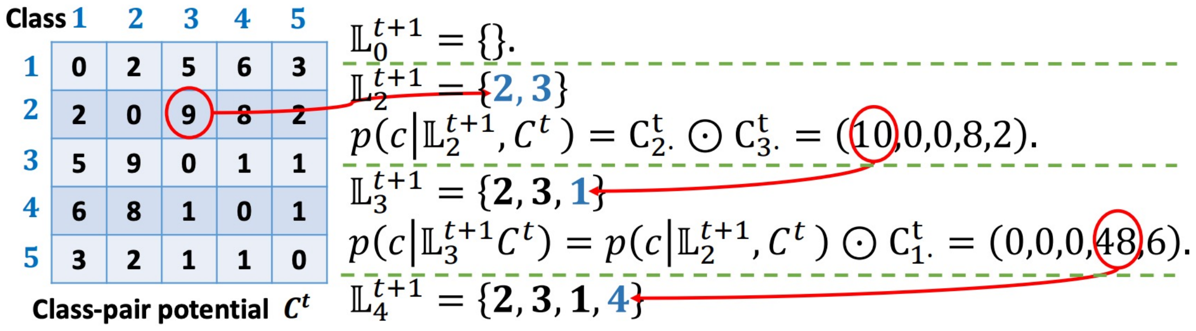
- Each assay is a task
- Noisy tasks caused by improper measurement

Some tasks are less valuable or contain noises

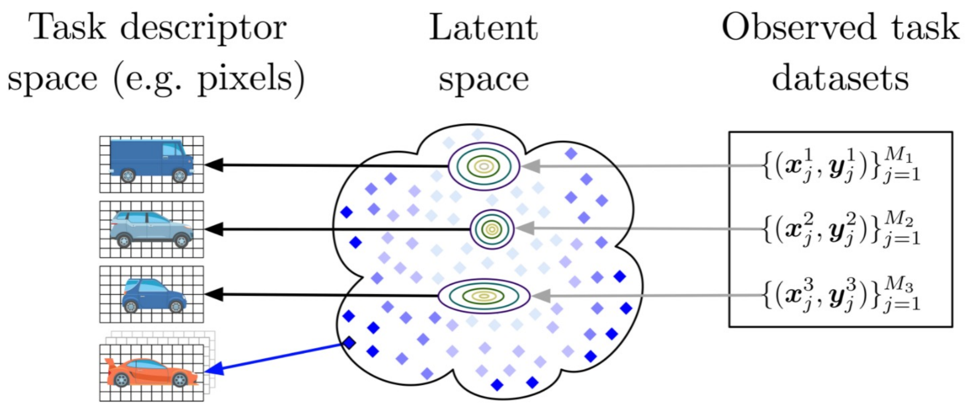
Require non-uniform sampling

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]

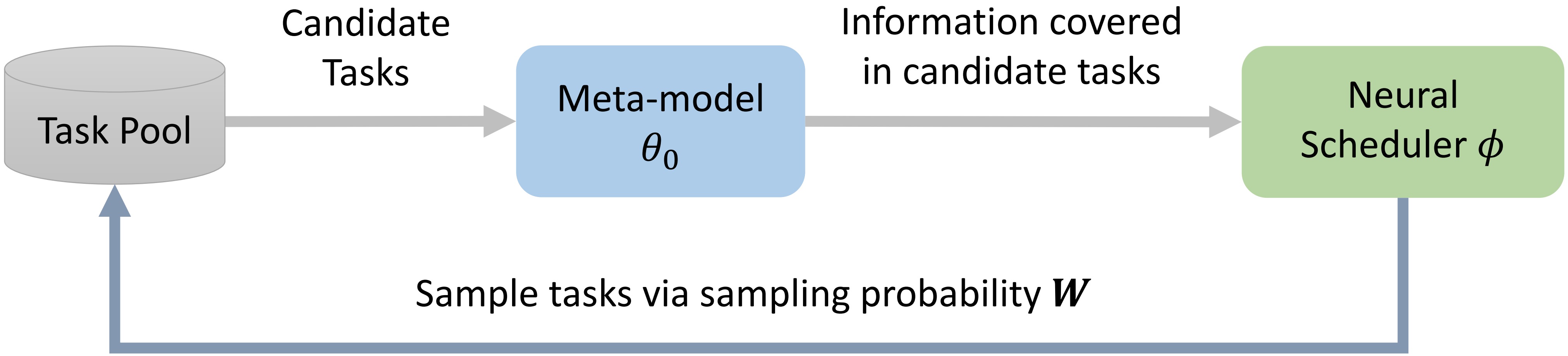


+ Benefit meta-learning process with a task scheduler

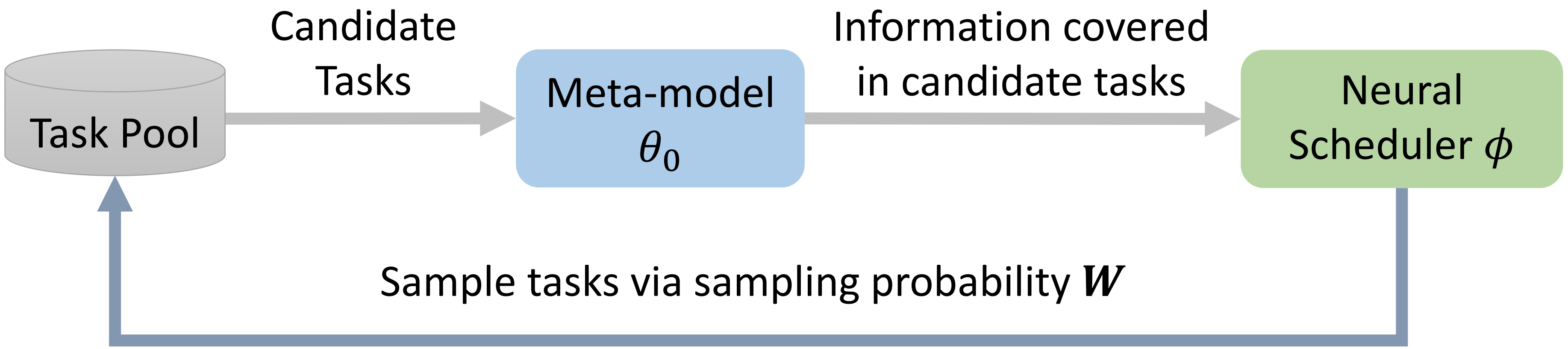
- Require manually strategy design
- The task scheduler can not adapt to the learning progress of the meta-model

Adaptive Task Sampling (ATS)

Goal: determining task sampling probability via a neural scheduler



Meta-model-related Factors



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

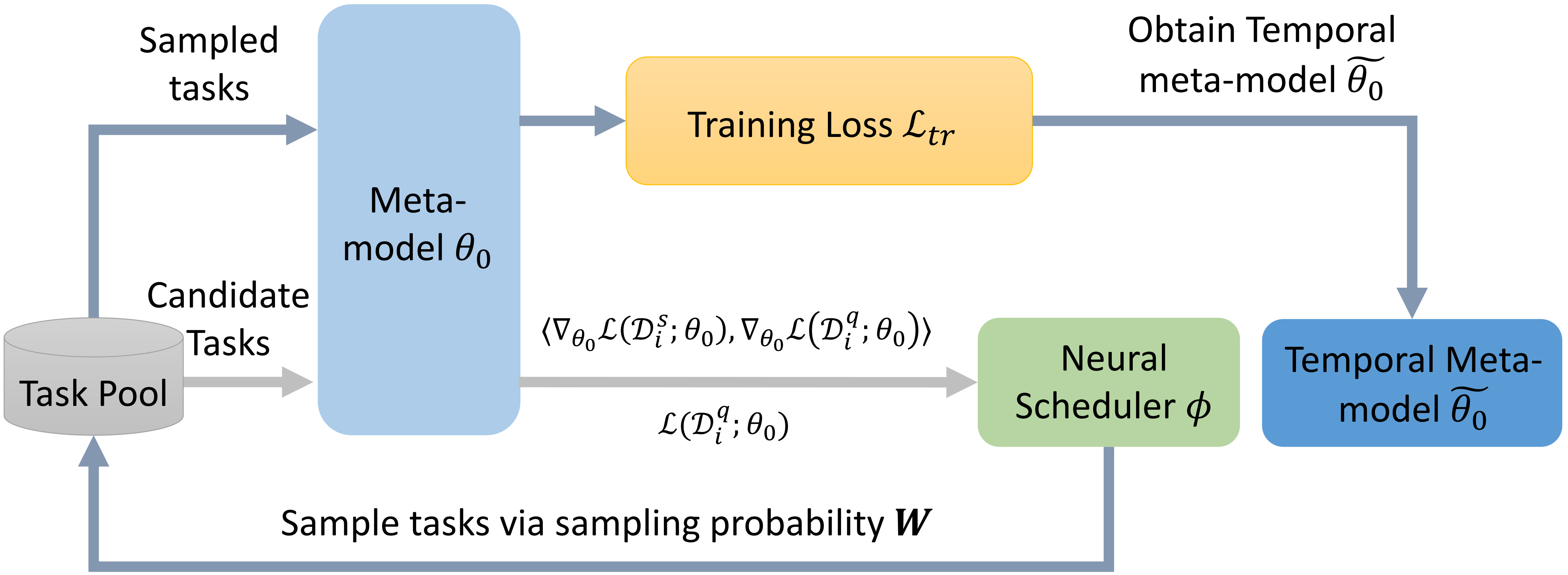
- large query losses + large gradient similarities $\longrightarrow$ true hard tasks
- Large query losses + small gradient similarities $\longrightarrow$ tasks with noise

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

How to Optimize?

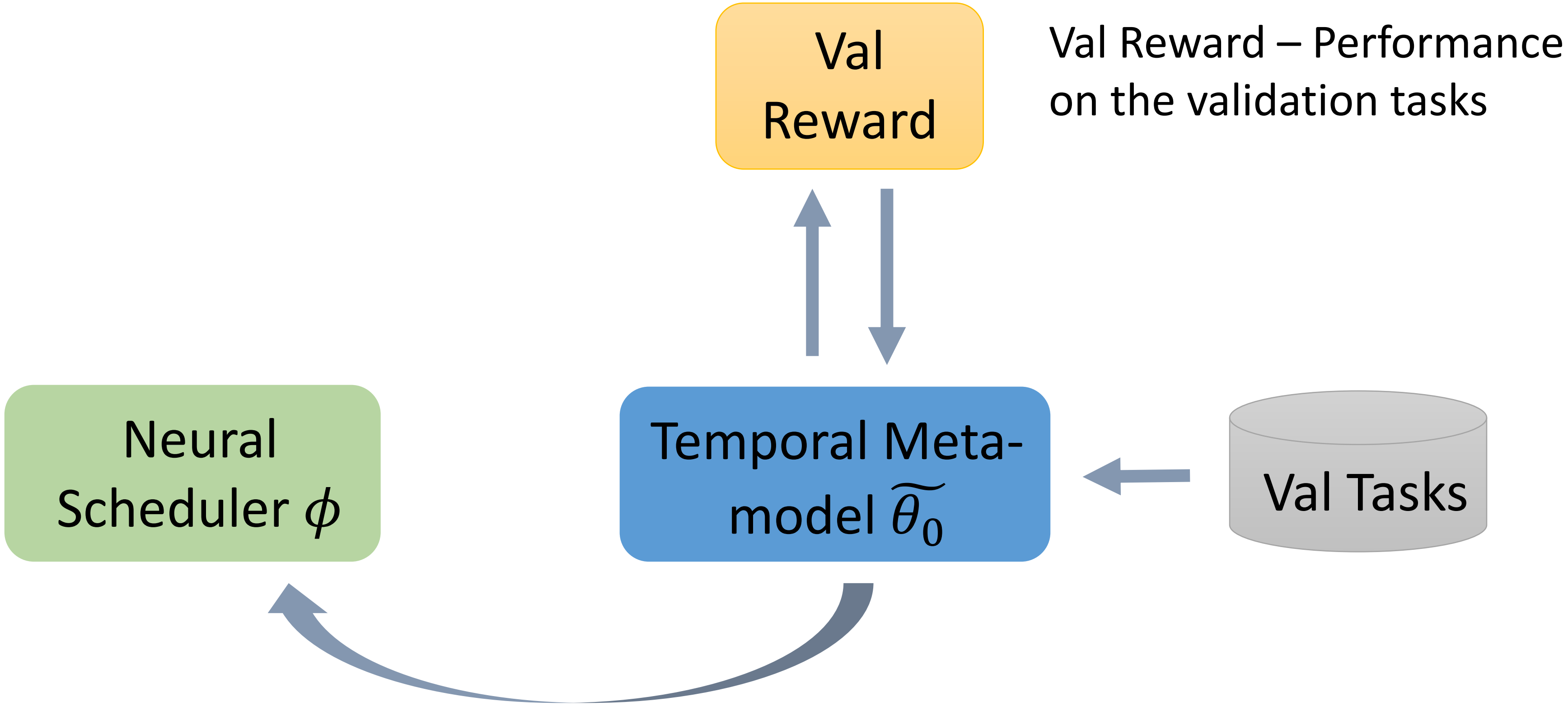
Optimize neural scheduler and meta-model alternatively

Step 1: Obtain the temporal meta-model



How to Optimize?

Step 2: Use validation tasks to optimize the neural scheduler

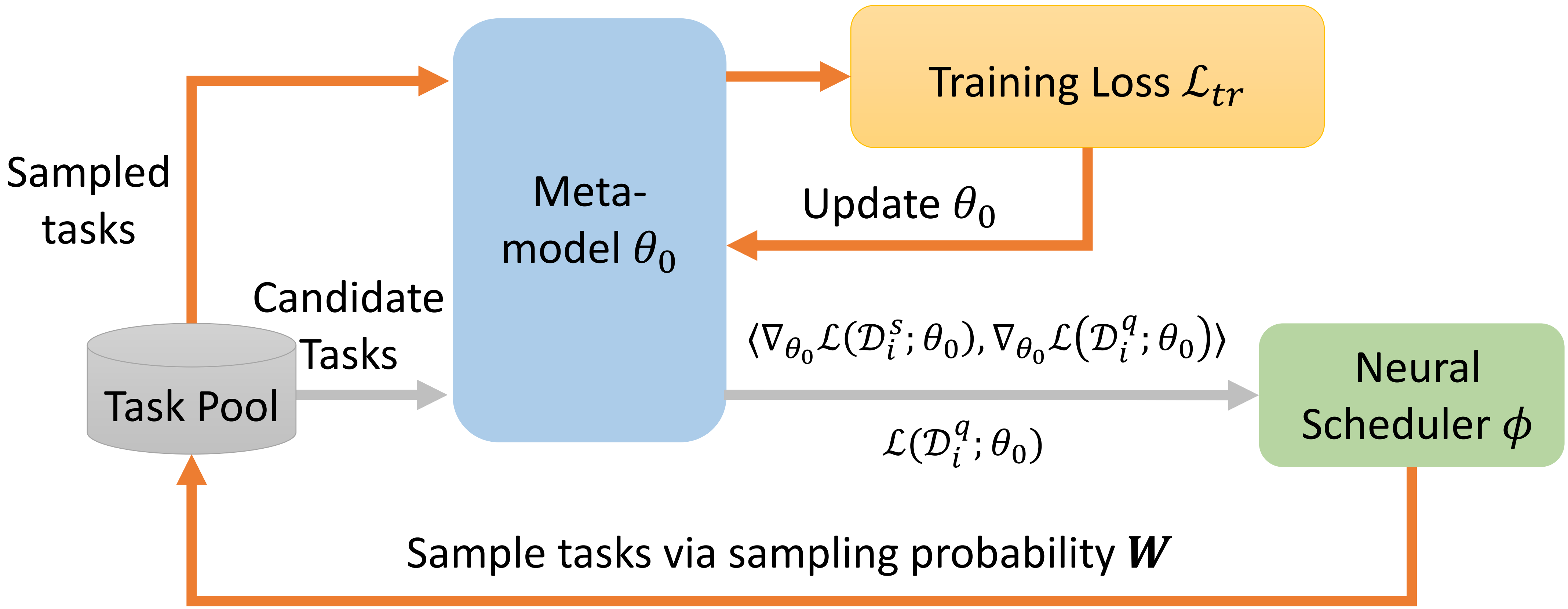


Use REINFOCE to Update ϕ

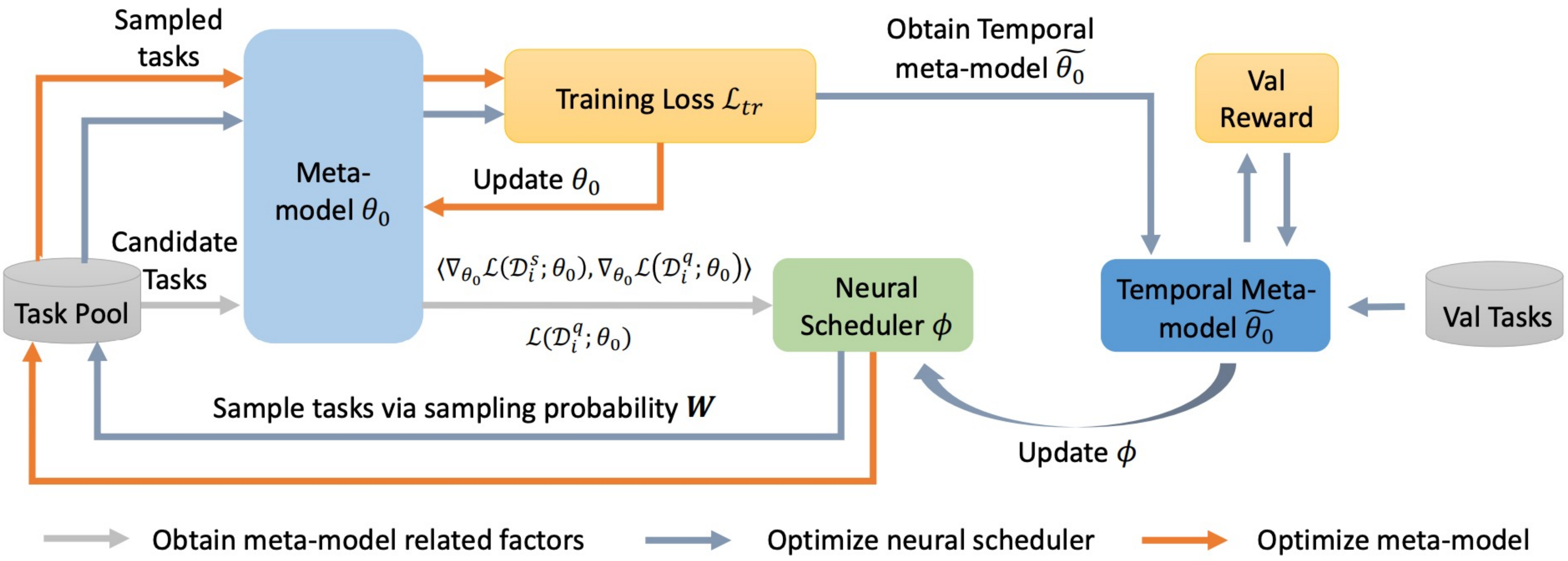
$$\phi^{(k+1)} \leftarrow \phi^{(k)} - \gamma \nabla_{\phi^{(k)}} \log P(\mathbf{W}^{(k)}; \phi^{(k)})\left(\frac{1}{N_v}\sum_{i=1}^{N_v} R_i^{(k)} - b\right)$$

How to Optimize?

Step 3: Update meta-model θ_0



Overall Framework



How does ATS Improves Meta-training Process?

Proposition 1. *Suppose that* $\mathbf{w} = [w_1, \cdots, w_{N^{pool}}]$ *denotes the random variable for sampling probabilities,* $\boldsymbol{\mathcal{L}}_{\theta_0} = [\mathcal{L}(\mathcal{D}_1^q; \theta_0), \cdots, \mathcal{L}(\mathcal{D}_{N^{pool}}^q; \theta_0)]$ *denotes the random variable for the loss using the meta-model, and* $\boldsymbol{\nabla}_{\theta_0} = [\langle \nabla_{\theta_0}\mathcal{L}(\mathcal{D}_1^s; \theta_0), \nabla_{\theta_0}\mathcal{L}(\mathcal{D}_1^q; \theta_0)\rangle, \cdots, \langle \nabla_{\theta_0}\mathcal{L}(\mathcal{D}_{N^{pool}}^s; \theta_0), \nabla_{\theta_0}\mathcal{L}(\mathcal{D}_{N^{pool}}^q; \theta_0)\rangle]$ *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:*

$$\mathcal{L}^w(\theta_0) = \mathcal{L}(\theta_0) + \mathrm{Cov}(\boldsymbol{\mathcal{L}}_{\theta_0}, \mathbf{w}) - \alpha \mathrm{Cov}(\boldsymbol{\nabla}_{\theta_0}, \mathbf{w}). \tag{10}$$

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*

$$w_i^* = \frac{e^{-\left[\mathcal{L}(\mathcal{D}_i^q;\theta_0^*) - \alpha\left\langle \nabla_{\theta_0}\mathcal{L}(\mathcal{D}_i^s;\theta_0^*), \nabla_{\theta_0}\mathcal{L}(\mathcal{D}_i^q;\theta_0^*)\right\rangle\right]}}{\sum_{i=1}^{B} e^{-\left[\mathcal{L}(\mathcal{D}_i^q;\theta_0^*) - \alpha\left\langle \nabla_{\theta_0}\mathcal{L}(\mathcal{D}_i^s;\theta_0^*), \nabla_{\theta_0}\mathcal{L}(\mathcal{D}_i^q;\theta_0^*)\right\rangle\right]}}, \tag{11}$$

the following hold:

$$\forall\theta_0 : \mathrm{Cov}(\boldsymbol{\mathcal{L}}_{\theta_0} - \alpha\boldsymbol{\nabla}_{\theta_0}, e^{-(\boldsymbol{\mathcal{L}}_{\theta_0^*} - \alpha\boldsymbol{\nabla}_{\theta_0^*})}) \geq 0, \qquad \mathcal{L}^w(\theta_0) - \mathcal{L}^w(\theta_0^*) \geq \mathcal{L}(\theta_0) - \mathcal{L}(\theta_0^*),$$

← Speed up training

$$\forall\theta_0 : \mathrm{Cov}(\boldsymbol{\mathcal{L}}_{\theta_0} - \alpha\boldsymbol{\nabla}_{\theta_0}, e^{-(\boldsymbol{\mathcal{L}}_{\theta_0^*} - \alpha\boldsymbol{\nabla}_{\theta_0^*})}) \leq -\mathrm{Var}(\boldsymbol{\mathcal{L}}_{\theta_0^*} - \alpha\boldsymbol{\nabla}_{\theta_0^*}), \quad \mathcal{L}^w(\theta_0) - \mathcal{L}^w(\theta_0^*) \leq \mathcal{L}(\theta_0) - \mathcal{L}(\theta_0^*).$$

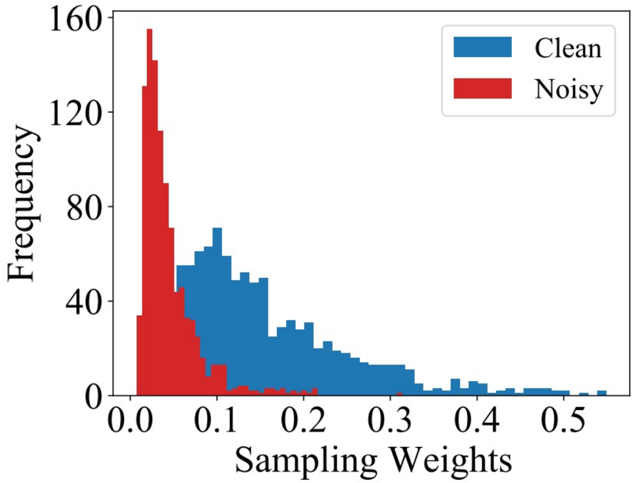
← The minima tends to be flat with better generalization ability

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)

Model	miniImagenet-noisy 5-way 1-shot	miniImagenet-noisy 5-way 5-shot	Drug-noisy mean	Drug-noisy medium	Drug-noisy >0.3
Uniform	$41.67 \pm 0.80\%$	$55.80 \pm 0.71\%$	0.202	0.113	21
SPL	$42.13 \pm 0.79\%$	$56.19 \pm 0.70\%$	0.211	0.138	24
FocalLoss	$41.91 \pm 0.78\%$	$53.58 \pm 0.75\%$	0.205	0.106	23
GCP	$41.86 \pm 0.75\%$	$54.63 \pm 0.72\%$	N/A	N/A	N/A
PAML	$41.49 \pm 0.74\%$	$52.45 \pm 0.69\%$	0.204	0.120	24
DAML	$41.26 \pm 0.73\%$	$55.46 \pm 0.70\%$	0.197	0.113	24
ATS (Ours)	$\mathbf{44.21 \pm 0.76\%}$	$\mathbf{59.50 \pm 0.71\%}$	**0.233***	**0.152***	**31***

* 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

Ablation Model	miniImagenet-noisy 5-way 1-shot	miniImagenet-noisy 5-way 5-shot	Drug-noisy mean	Drug-noisy medium	Drug-noisy >0.3
Random ϕ	$41.95 \pm 0.80\%$	$56.07 \pm 0.71\%$	0.204	0.100	22
Rank by Sim/Loss	$42.84 \pm 0.76\%$	$57.90 \pm 0.68\%$	0.181	0.109	22
ϕ+Loss	$42.45 \pm 0.80\%$	$56.65 \pm 0.75\%$	0.212	0.122	27
ϕ+Sim	$42.28 \pm 0.82\%$	$56.71 \pm 0.72\%$	0.214	0.122	29
Reweighting	$42.19 \pm 0.80\%$	$56.48 \pm 0.72\%$	0.217	0.118	28
ATS (ϕ+Loss+Sim)	$\mathbf{44.21 \pm 0.76\%}$	$\mathbf{59.50 \pm 0.71\%}$	**0.233***	**0.152***	**31***

* 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

More noises →

	Noise Ratio	0.2	0.4	0.6	0.8
Image	Uniform	$43.46 \pm 0.82\%$	$42.92 \pm 0.78\%$	$41.67 \pm 0.80\%$	$36.53 \pm 0.73\%$
	BNS	$44.04 \pm 0.81\%$	$43.36 \pm 0.75\%$	$42.13 \pm 0.79\%$	$38.21 \pm 0.75\%$
	ATS (Ours)	$\mathbf{45.55 \pm 0.80\%}$	$\mathbf{44.50 \pm 0.86\%}$	$\mathbf{44.21 \pm 0.76\%}$	$\mathbf{42.18 \pm 0.73\%}$

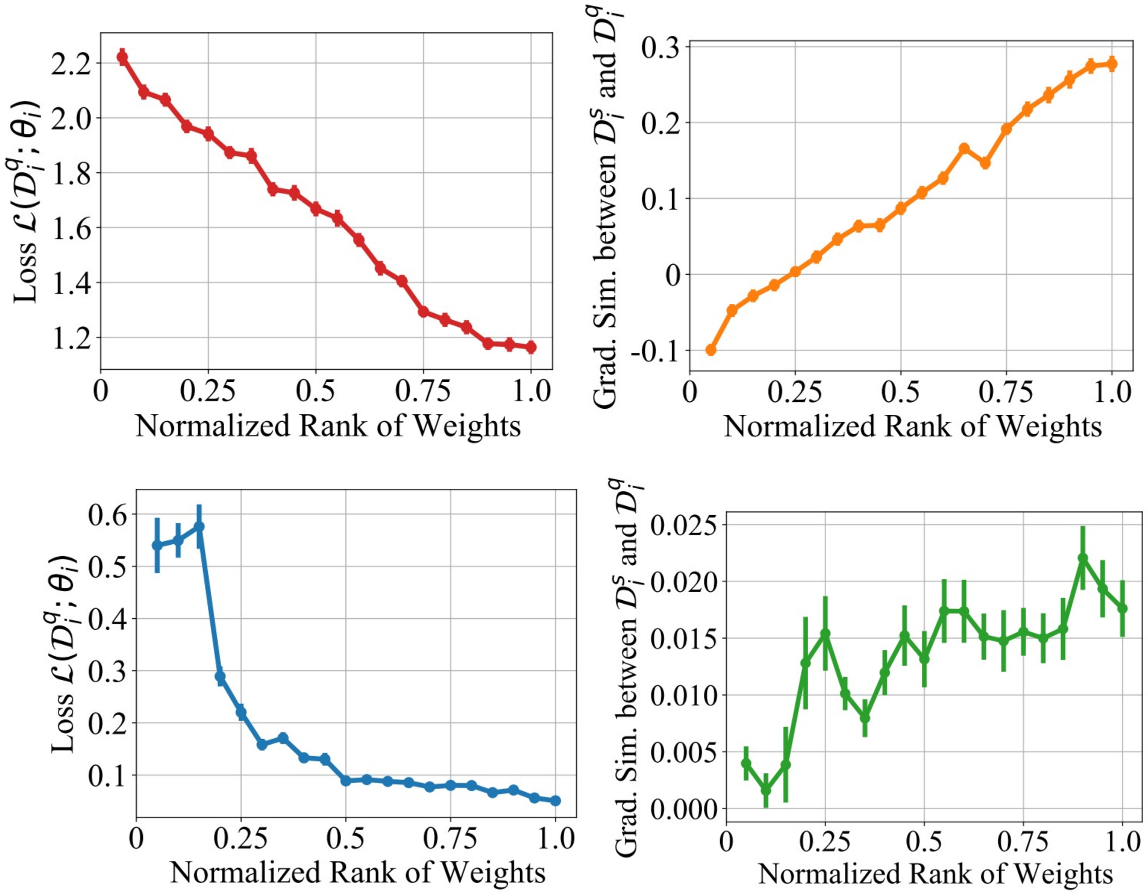
	Noise Scaler	η=2			η=4			η=6			η=8		
Drug	Uniform	0.222	0.139	26	0.202	0.113	21	0.196	0.131	22	0.194	0.100	21
	BNS	0.229	0.136	31	0.211	0.138	24	0.208	0.116	24	0.200	0.101	24
	ATS* (Ours)	**0.235**	**0.160**	**33**	**0.233**	**0.152**	**31**	**0.221**	**0.136**	**28**	**0.219**	**0.133**	**28**

* 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

miniImagenet

Drug



High losses + low gradient similarities → 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

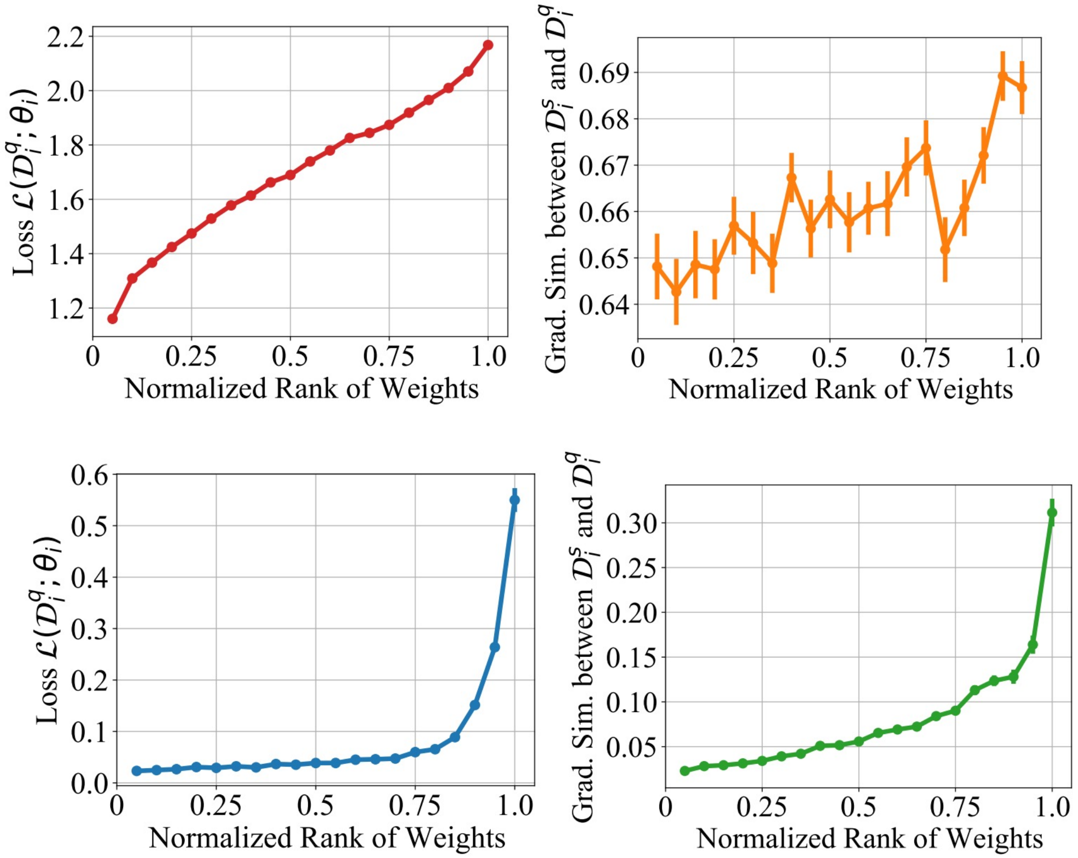
Model	miniImagenet-Limited		Drug-Full		
	5-way 1-shot	5-way 5-shot	mean	medium	>0.3
Uniform	$33.61 \pm 0.66\%$	$45.97 \pm 0.65\%$	0.233	0.140	33
SPL	$34.28 \pm 0.65\%$	$46.05 \pm 0.69\%$	0.232	0.135	29
FocalLoss	$33.11 \pm 0.65\%$	$46.12 \pm 0.70\%$	0.229	0.140	28
GCP	$34.69 \pm 0.67\%$	$46.86 \pm 0.68\%$	N/A	N/A	N/A
PAML	$33.64 \pm 0.62\%$	$45.01 \pm 0.69\%$	0.238	0.144	32
DAML	$34.83 \pm 0.69\%$	$46.66 \pm 0.67\%$	0.227	0.141	28
ATS (Ours)	$\mathbf{35.15 \pm 0.67\%}$	$\mathbf{47.76 \pm 0.68\%}$	**0.252***	**0.179***	**36***

* 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

miniImagenet

Drug



High losses + high gradient similarities → more valuable tasks

Effect of the Budgets

Less meta-training tasks → more improvements

More tasks →

Budgets	16	32	48	64
Uniform	$33.61 \pm 0.66\%$	$40.48 \pm 0.75\%$	$44.07 \pm 0.80\%$	$45.73 \pm 0.79\%$
GCP	$34.69 \pm 0.67\%$	$41.27 \pm 0.74\%$	$44.30 \pm 0.79\%$	$45.35 \pm 0.81\%$
ATS (Ours)	$\mathbf{35.15 \pm 0.67\%}$	$\mathbf{41.68 \pm 0.78\%}$	$\mathbf{44.89 \pm 0.79\%}$	$\mathbf{46.27 \pm 0.80\%}$

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