Meta-Learning with Self-Improving Momentum Target

Jihoon Tack[1], Jongjin Park[1], Hankook Lee[1], Jaeho Lee[2], Jinwoo Shin[1]

Korea Advanced Institute of Science and Technology (KAIST)[1], Pohang University of Science and Technology (POSTECH)[2]

KAIST

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

arXiv

Github

NEURAL INFORMATION PROCESSING SYSTEMS

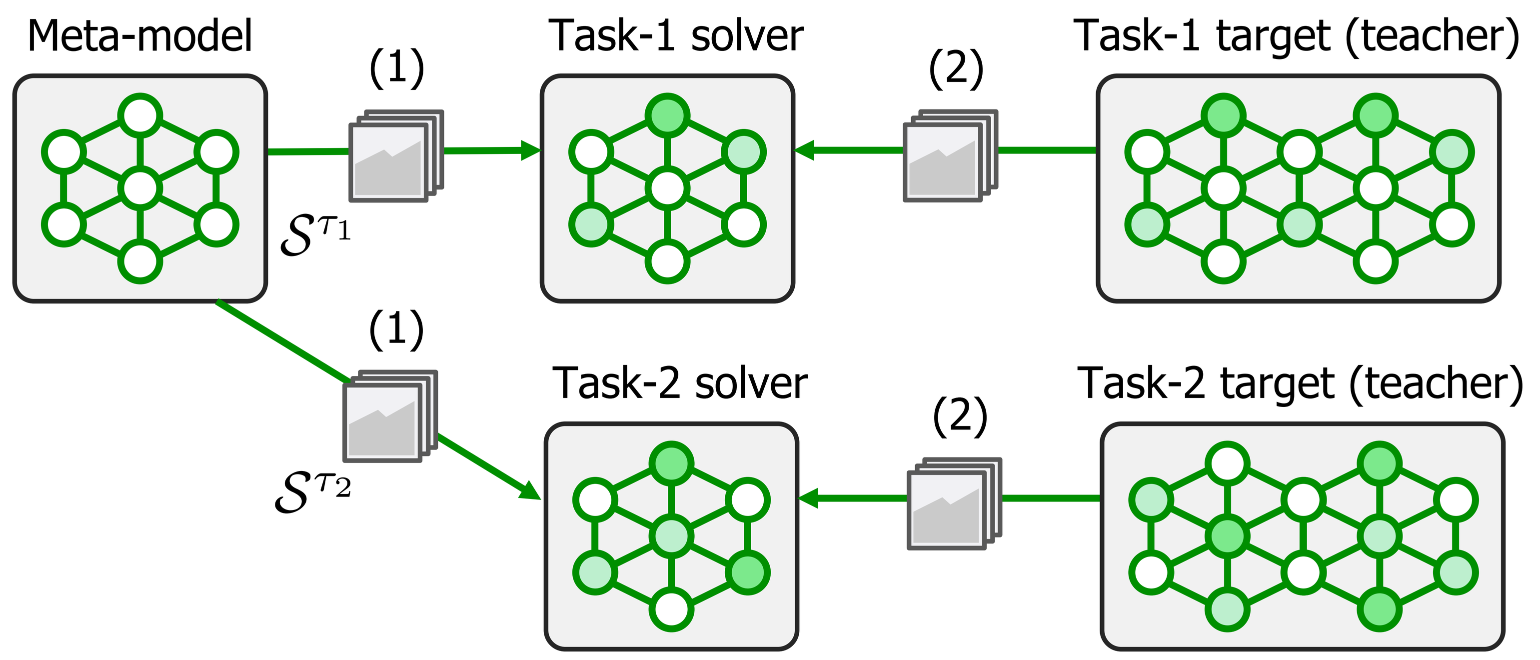
TL;DR. We propose a meta-learning algorithm to generate a target model from which we distill the knowledge to the meta-model, forming a virtuous cycle of improvements

Introduction

Meta-learning

- Extracting and utilizing the **knowledge from the distribution of tasks** to better solve a relevant task

To learn a meta-model, one (1) **adapt** an appropriate solver for each task with the given support set and (2) **evaluate** with
(a) a given query set
(b) **a target model** → recently shown effectiveness

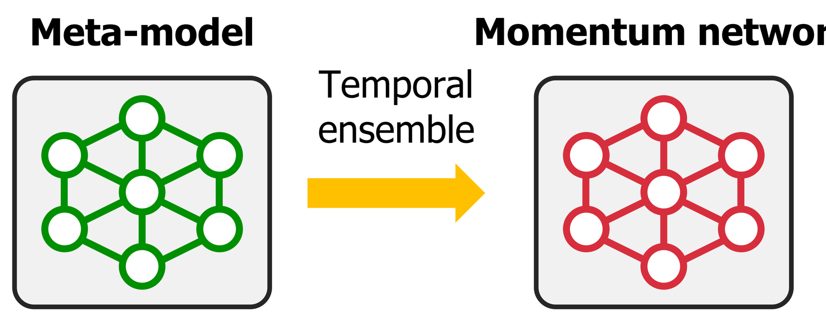


Q. How can we train target models for **every task**…?
- This suffers from **computation** and **storage** burden!

A. Lets generate target models from a **better meta-learner**!

Temporal Ensemble

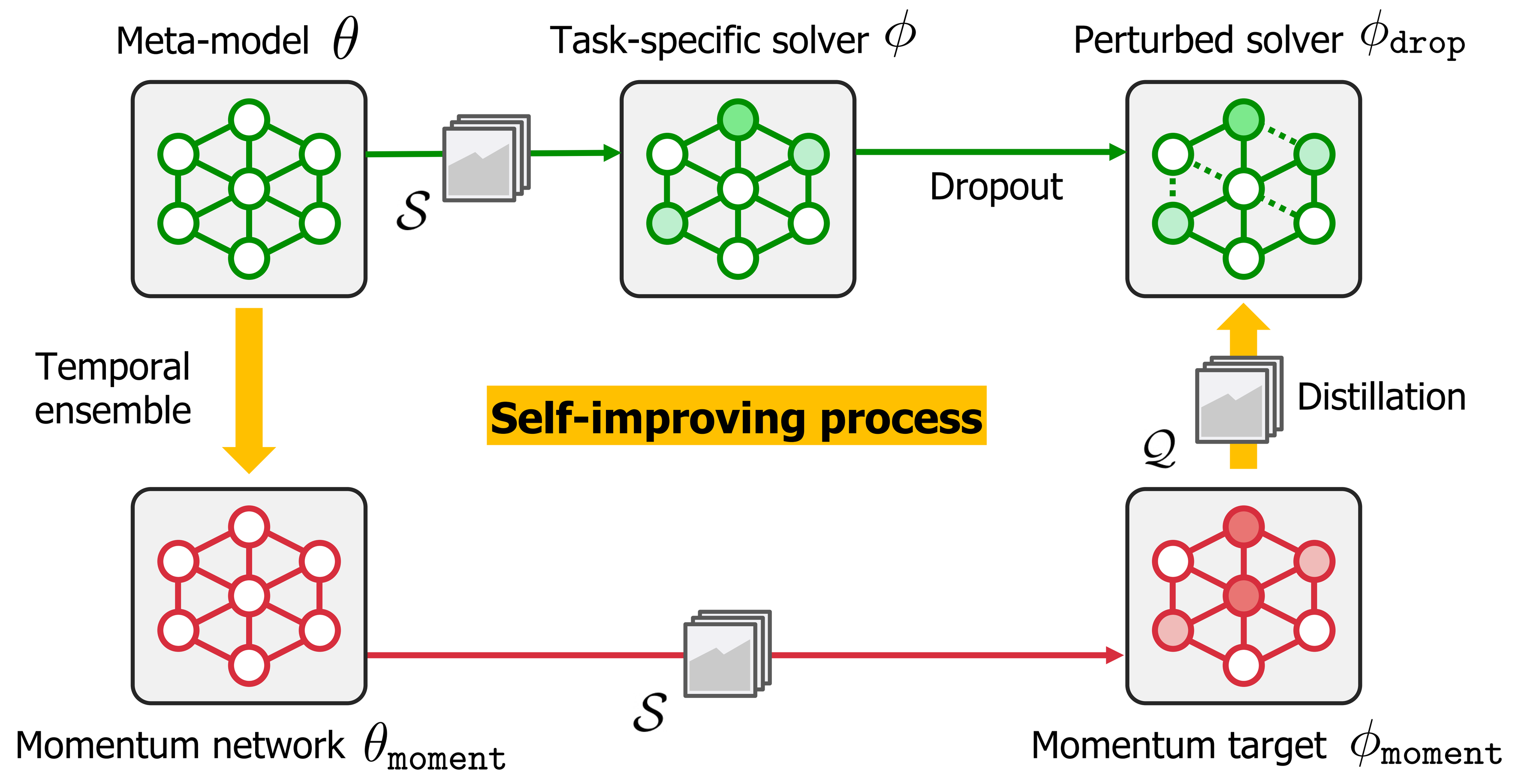
We find that the **temporal ensemble** of the meta-model is a better meta-learner, i.e., a better adaptation performance



$$\theta_{\texttt{moment}} \leftarrow \eta \cdot \theta_{\texttt{moment}} + (1-\eta)\cdot\theta$$

SiMT: Self-improving Momentum Target

We propose **Self-improving Momentum Target (SiMT)**



- **Momentum target**: Generating the target model from the momentum network to teach the original model
- **Self-improving process**: Improving the meta-model recursively improves the momentum network
- **(+) Asymmetricity**: preventing the momentum target and the solver from becoming too similar, stabilizes the training

Flexibility of SiMT. SiMT can be used over various backbone meta-learning methods, including gradient- and model-based.

$$\underbrace{(1-\lambda)\cdot\mathcal{L}(\phi_{\texttt{drop}},\mathcal{Q})}_{\textit{Learning from the query set}} + \lambda\cdot\underbrace{\mathcal{L}_{\texttt{teach}}(\phi_{\texttt{drop}},\phi_{\texttt{target}},\mathcal{Q})}_{\textit{Learning from the momentum target}}$$

Experimental Results

Main results. SiMT shows the effectiveness for three parts:
(1) few-shot regression, (2) few-shot classification, and
(3) meta-reinforcement learning. *See the paper for other results*

Model	Method	mini-ImageNet		tiered-ImageNet	
		1-shot	5-shot	1-shot	5-shot
Conv4 [55]	MAML [10]	47.33±0.45	63.27±0.14	50.19±0.21	66.05±0.19
	MAML [10] + SiMT	**51.49**±**0.18**	**68.74**±**0.12**	**52.51**±**0.21**	**69.58**±**0.11**
	ANIL [36]	47.71±0.47	63.13±0.43	49.57±0.04	66.34±0.28
	ANIL [36] + SiMT	**50.81**±**0.56**	**67.99**±**0.19**	**51.66**±**0.26**	**68.88**±**0.08**
	MetaSGD [31]	50.66±0.18	65.55±0.54	52.48±1.22	71.06±0.20
	MetaSGD [31] + SiMT	**51.70**±**0.80**	**69.13**±**1.40**	**52.98**±**0.07**	**71.46**±**0.12**
	ProtoNet [45]	47.97±0.29	65.16±0.67	51.90±0.55	71.51±0.25
	ProtoNet [45] + SiMT	**51.25**±**0.55**	**68.71**±**0.35**	**53.25**±**0.27**	**72.69**±**0.27**
ResNet-12 [34]	MAML [10]	52.66±0.60	68.69±0.33	57.32±0.59	73.78±0.27
	MAML [10] + SiMT	**56.28**±**0.63**	**72.01**±**0.26**	**59.72**±**0.22**	**74.40**±**0.90**
	ANIL [36]	51.80±0.59	68.38±0.20	57.52±0.68	73.50±0.35
	ANIL [36] + SiMT	**54.44**±**0.27**	**69.98**±**0.66**	**58.18**±**0.31**	**75.59**±**0.50**
	MetaSGD [31]	54.95±0.11	70.65±0.43	58.97±0.89	76.37±0.11
	MetaSGD [31] + SiMT	**55.72**±**0.96**	**74.01**±**0.79**	**61.03**±**0.05**	**78.04**±**0.48**
	ProtoNet [45]	52.84±0.21	68.35±0.29	61.16±0.17	79.94±0.20
	ProtoNet [45] + SiMT	**55.84**±**0.57**	**72.45**±**0.32**	**62.01**±**0.42**	**81.82**±**0.12**

Comparison with other target models [1,2]

Method	mini-ImageNet		tiered-ImageNet	
	1-shot	5-shot	1-shot	5-shot
MAML [10]	47.33±0.45	63.27±0.14	50.19±0.21	66.05±0.19
MAML [10] + Bootstrap [16]	48.68±0.33	68.45±0.40	49.34±0.26	68.84±0.37
MAML [10] + SiMT	**51.49**±**0.18**	**68.74**±**0.12**	**52.51**±**0.21**	**69.58**±**0.11**
ANIL [36]	47.71±0.47	63.13±0.43	49.57±0.04	66.34±0.28
ANIL [36] + Bootstrap [16]	47.74±0.44	65.16±0.04	48.85±0.34	66.09±0.07
ANIL [36] + SiMT	**50.81**±**0.56**	**67.99**±**0.19**	**51.66**±**0.26**	**68.88**±**0.08**

Method	1-shot train cost (GPU hours)	mini-ImageNet		tiered-ImageNet	
		1-shot	5-shot	1-shot	5-shot
MAML [10]*	1.31	58.84±0.25	74.62±0.38	63.02±0.30	67.26±0.32
MAML [10] + Lu et al. [32] - 5%*	5.04	59.14±0.33	75.77±0.29	64.52±0.30	68.39±0.34
MAML [10] + Lu et al. [32] - 10%*	8.32	60.06±0.35	76.34±0.42	**65.23**±0.45	70.02±0.33
MAML [10] + SiMT	1.64	**62.05**±0.39	**78.77**±0.45	63.91±0.32	**77.43**±0.47

[1] Bootstrapped Meta-Learning, Flennerhag et al., ICLR 2022
[2] Towards enabling meta-learning from target models, Lu et al., NeurIPS 2021