Hierarchically Structured Meta-learning

Huaxiu Yao[1,2], Ying Wei[2], Junzhou Huang[1], Zhenhui Li[2]

[1]Pennsylvania State University
[2]Tencent AI Lab

Oral: Thu Jun 13th 09:35 -- 09:40 AM @ Room 103
Poster: Thu Jun 13th 06:30 -- 09:00 PM @ Pacific Ballroom #183

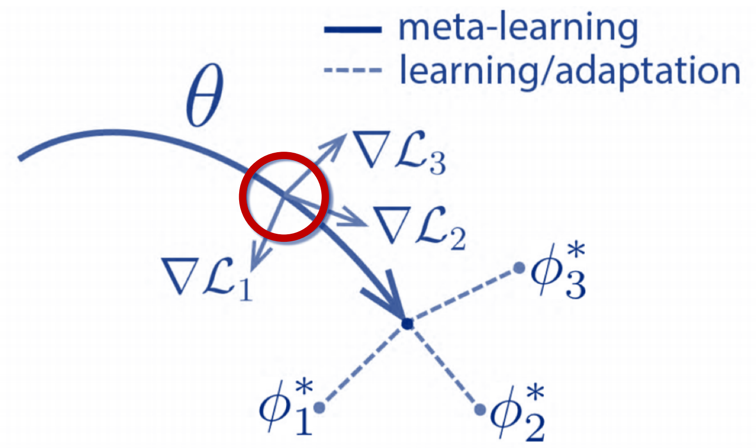
Gradient-based Meta-learning (MAML [1])

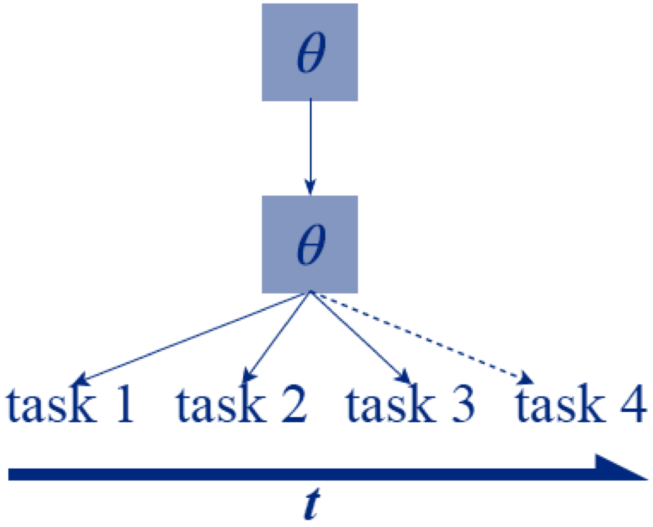
Is global initialization enough?

$$\min_{\theta} \sum_{\text{task } i} \mathcal{L}^{i}_{\text{test}}(\theta - \alpha \nabla_{\theta} \mathcal{L}^{i}_{\text{train}}(\theta))$$

θ parameter vector being meta-learned

ϕ_i^* optimal parameter vector for task i

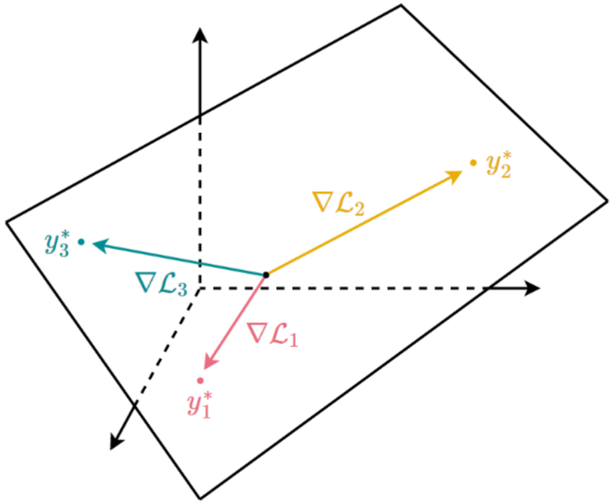


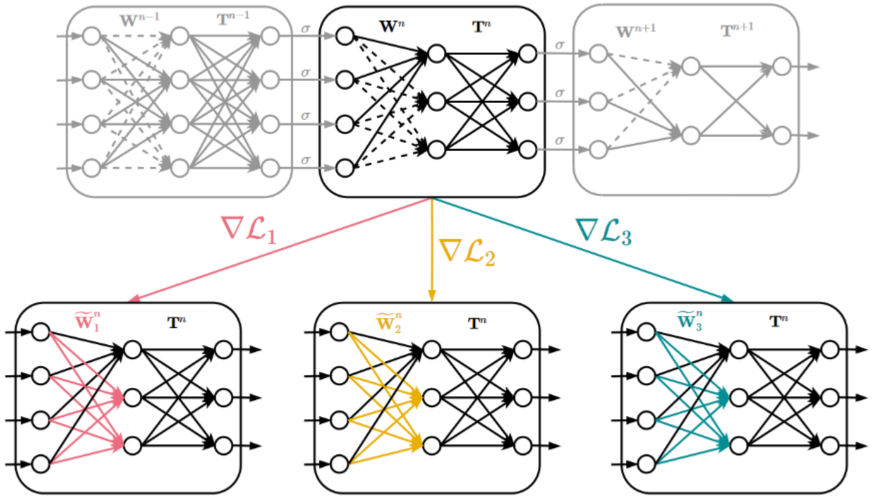


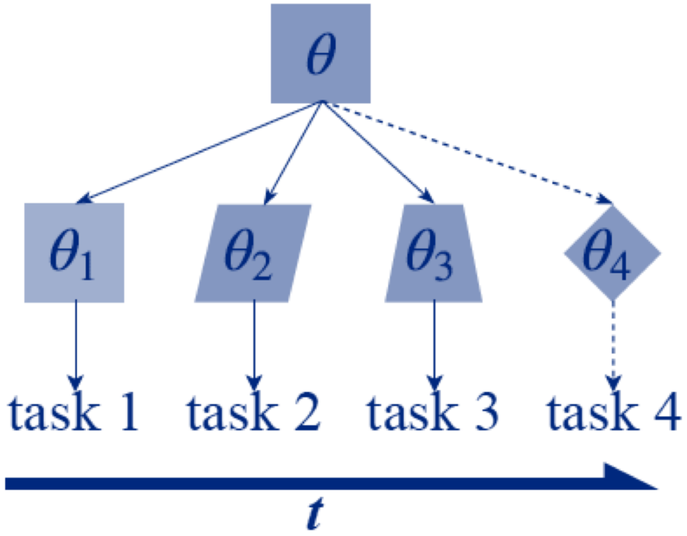
[1] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
http://people.eecs.berkeley.edu/~cbfinn/_files/metalearning_frontiers_2018_small.pdf

Task-specific Meta-learning (MT-Net [2])

Should the initialization be tailored to each task?







[2] Lee, Yoonho, and Seungjin Choi. "Gradient-Based Meta-Learning with Learned Layerwise Metric and Subspace." International Conference on Machine Learning. 2018.

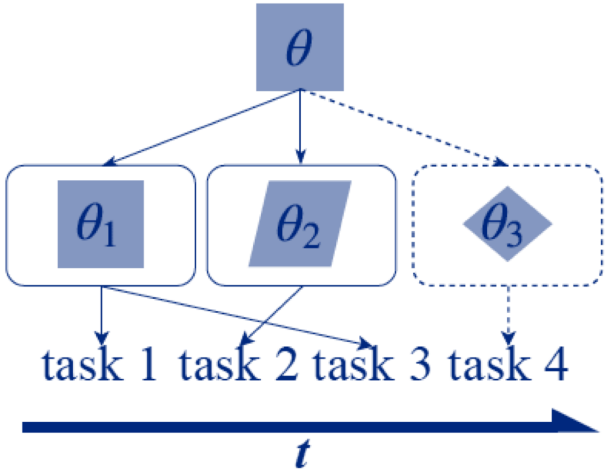
Human Beings: Knowledge Organization and Reuse

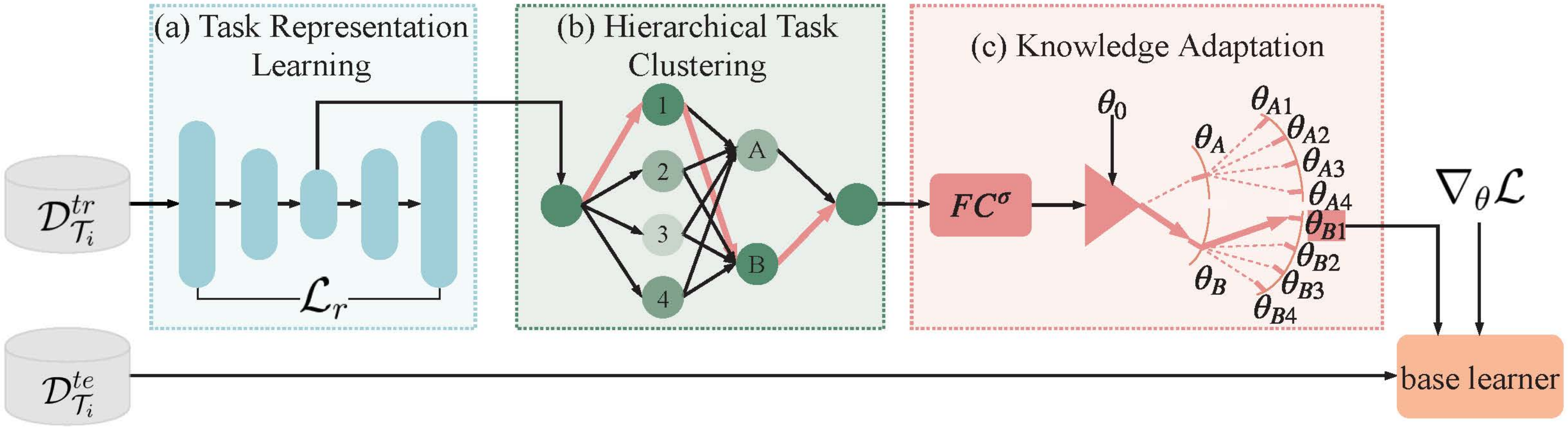
[3] Gershman, Samuel J., David M. Blei, and Yael Niv. "Context, learning, and extinction." Psychological review 117.1 (2010): 197.
[4] Gershman, Samuel J., et al. "Statistical computations underlying the dynamics of memory updating." *PLoS computational biology* 10.11 (2014): e1003939.

Our Solution: Hierarchically Structured Meta-learning

Balance between generalization and customization

- Organize tasks by hierarchical clustering
- Adapt the global initialization to each cluster of tasks





HSML: Optimization

Overall optimization problem

$$\min_{\Theta} \sum_{i=1}^{N_t} \mathcal{L}(f_{\theta_{0i} - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{tr}_{\mathcal{T}_i})}, \mathcal{D}^{te}_{\mathcal{T}_i}) + \xi \mathcal{L}_r(\mathcal{D}^{tr}_{\mathcal{T}_i}),$$

Extension to continual adaptation

- Incrementally increase the clusters as tasks sequentially arrive.
- Criterion for adding a cluster—evaluate the average loss over Q epochs

$$\bar{\mathcal{L}}_{new} > \mu \bar{\mathcal{L}}_{old}$$

Analysis

- For task $\mathcal{T}_i \sim \mathcal{E}$, training and testing samples are i.i.d. drawn from $\mathcal{S}_i$
- The initialization of HSML (K clusters) can be represented as $\theta_{0t} = \sum_{k=1}^{K} \widehat{\boldsymbol{B}}_k \theta_0$
- According to [5], the assumptions are $\mathcal{L} \in [0, 1]$ is η-smooth and has a ρ-Lipschitz Hessian, step size at the u-step $\alpha_u = c/u$ satisfying $c \leq \min\{\frac{1}{\eta}, \frac{1}{4(2\eta \ln U)^2}\}$ with total steps $U = n^{tr}$.
- The generalization of base learner $f_{\theta_{\mathcal{T}_i}}$ is bounded by $\epsilon(\mathcal{S}_i, \theta_0)$, where

$$\epsilon(\mathcal{S}_i, \theta_0) = \mathcal{O}\left(\left(1 + \frac{1}{c\hat{\gamma}^-}\right) \hat{R}_{\mathcal{D}_{\mathcal{T}_t}^{tr}}(\theta_{0t})^{\frac{c\hat{\gamma}^+}{1+c\hat{\gamma}^+}} \frac{1}{(n^{tr})^{\frac{1}{1+c\hat{\gamma}^+}}}\right)$$

- MAML can be regarded as a special case of HSML, i.e., $\forall k, \widehat{\boldsymbol{B}}_k = \boldsymbol{I}$
- After proving $\exists \{\hat{B}_k\}_{k=1}^{K}$, s.t., $\hat{R}_{\mathcal{D}_{\mathcal{T}_t^{tr}}}(\theta_{0t}) \leq \hat{R}_{\mathcal{D}_{\mathcal{T}_t^{tr}}}(\theta_0)$, we conclude that HSML achieves a tighter generalization bound than MAML

[5] Kuzborskij, Ilja, and Christoph Lampert. "Data-Dependent Stability of Stochastic Gradient Descent." International Conference on Machine Learning. 2018.

Experiments: Toy Regression

Data

- 4 sync family functions—Sin, Line, Cubic, Quadratic
- K-shot: K samples are used as training (each task)

Base model

- 2 layers FC with 40 neurons each

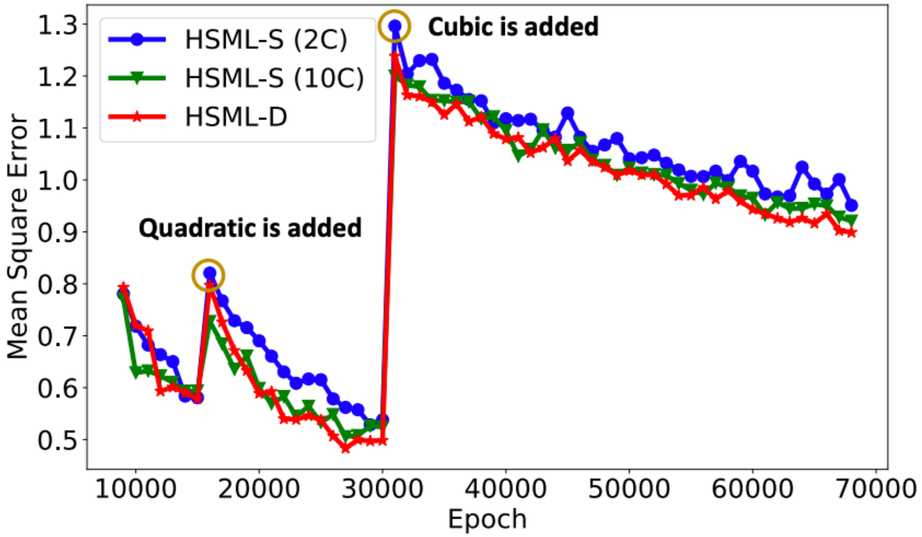
PennState Tencent AI Lab

Quantitative results

- Comparison on regression MSEs

Method	5-shot	10-shot
Global shared (MAML)	2.205±0.121	0.761±0.068
Task-specific (MUMOMAML[6])	1.096±0.085	0.256±0.028
Our method (HSML)	**0.856±0.073**	**0.161±0.021**

- Comparison in the continual adaptation scenario



Model	HSML-S (2C)	HSML-S (10C)	HSML-D
MSE ± 95% CI	0.933 ± 0.074	0.889 ± 0.071	**0.869 ± 0.072**

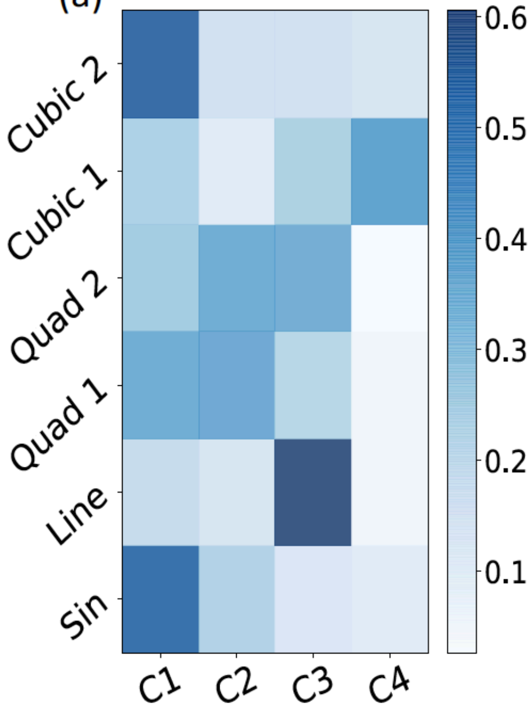
[6] Vuorio, Risto, Shao-Hua Sun, Hexiang Hu, and Joseph J. Lim. "Toward Multimodal Model-Agnostic Meta-Learning." arXiv preprint arXiv:1812.07172 (2018).

Experiments: Toy Regression

Qualitive results

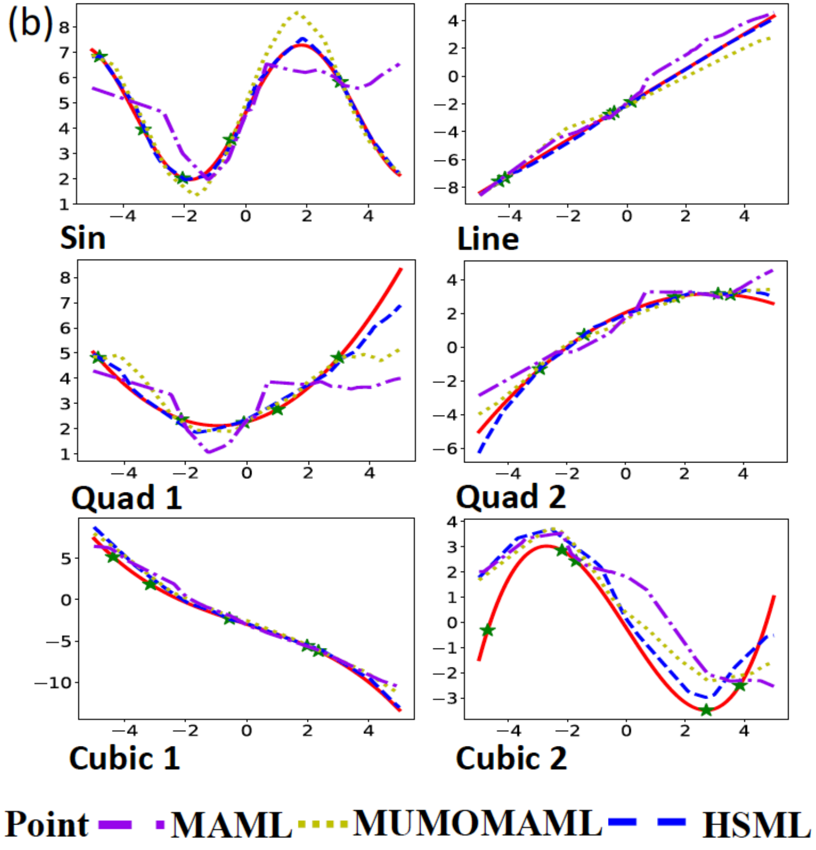
- Cluster assignment interpretation

(a)



- Regression results

(b)



Experiments: Few-shot Classification

Data

- 4 image classification datasets—Bird, Texture, Aircraft, Fungi
- 5-way, 1-shot

Base model

- a convolutional network with 4 convolution blocks

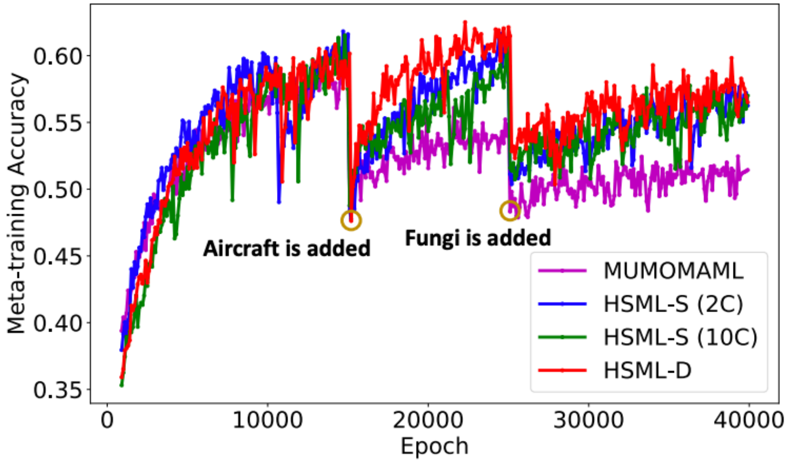
Experiments: Few-shot Classification

Quantitative results

- Comparison on accuracy

Method	Bird	Texture	Aircraft	Fungi
Global shared (MAML)	53.94%	31.66%	51.37%	42.12%
Task-specific (MUMOMAML[6])	56.82%	33.81%	53.14%	42.22%
Our method (HSML)	**60.98%**	**35.01%**	**57.38%**	**44.02%**

- Comparison in the continual adaptation scenario



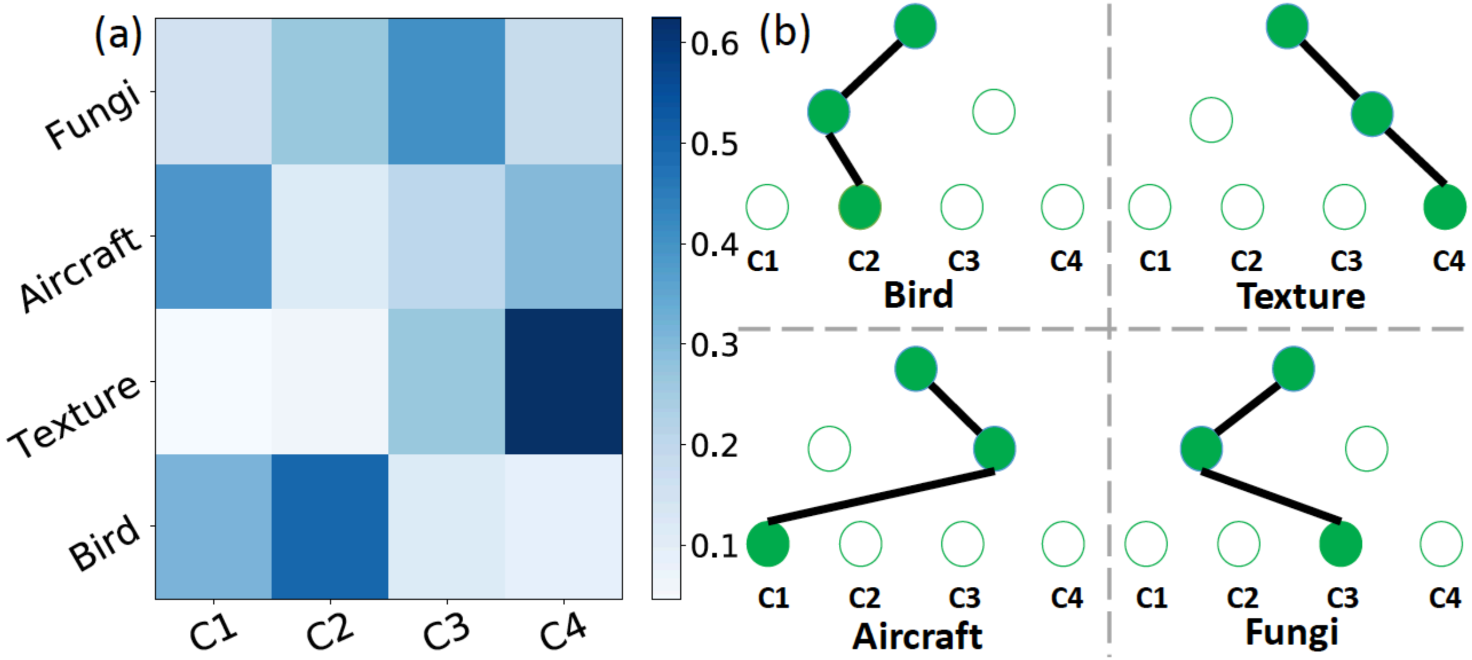
Model	Bird	Texture	Aircraft	Fungi
MUMOMAML	56.66%	33.68%	45.73%	40.38%
HSML-S (2C)	60.77%	33.41%	51.28%	40.78%
HSML-S (10C)	59.16%	34.48%	52.30%	40.56%
HSML-D	**61.16%**	**34.53%**	**54.50%**	**41.66%**

[6] Vuorio, Risto, Shao-Hua Sun, Hexiang Hu, and Joseph J. Lim. "Toward Multimodal Model-Agnostic Meta-Learning." arXiv preprint arXiv:1812.07172 (2018).

PennState Tencent AI Lab

Qualitive results

- Cluster assignment interpretation



Conclusions

- HSML simultaneously customizes task knowledge and preserves knowledge generalization via the hierarchical clustering structure.
- Experiments demonstrate the effectiveness and interpretability of HSML in both toy regression and few-shot classification problems.

THANK YOU

Oral: Thu Jun 13th 09:35 -- 09:40 AM @ Room 103
Poster: Thu Jun 13th 06:30 -- 09:00 PM @ Pacific Ballroom #183