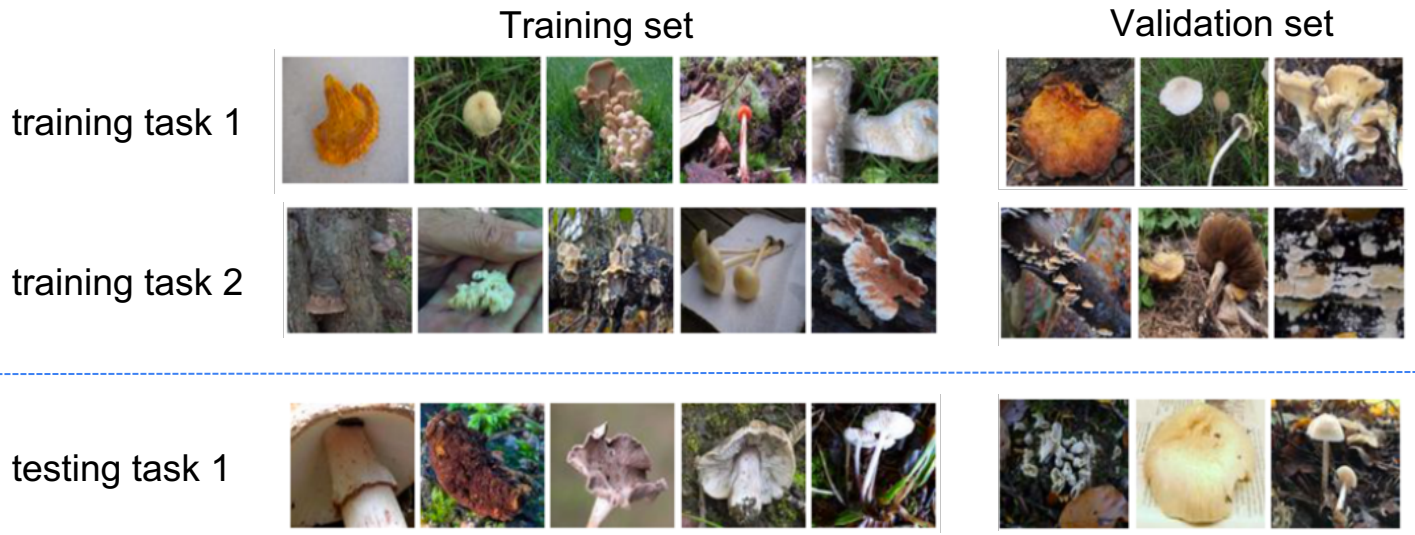
Subspace Learning for Effective Meta-Learning

Weisen Jiang[1 2] James T. Kwok[2] Yu Zhang[1 3]

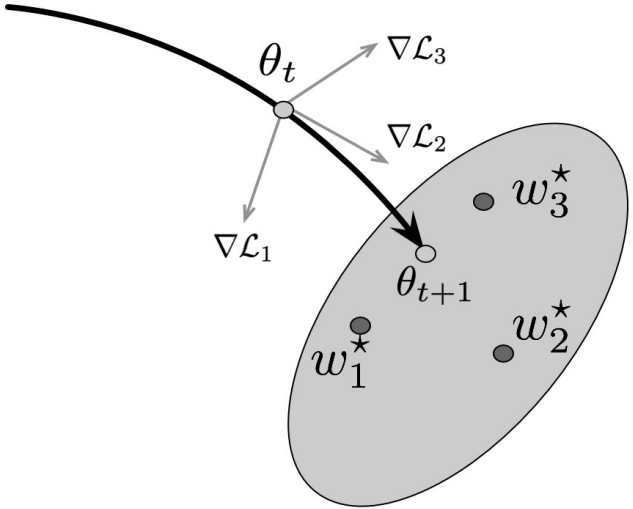
[1]Southern University of Science and Technology
[2]Hong Kong University of Science and Technology
[3]Peng Cheng Laboratory

Background: Meta-Learning



- Humans can learn new concepts with limited examples.
- However, deep networks are data-hungry.
- Meta-learning aims to extract meta-knowledge from seen tasks to accelerate learning unseen tasks.

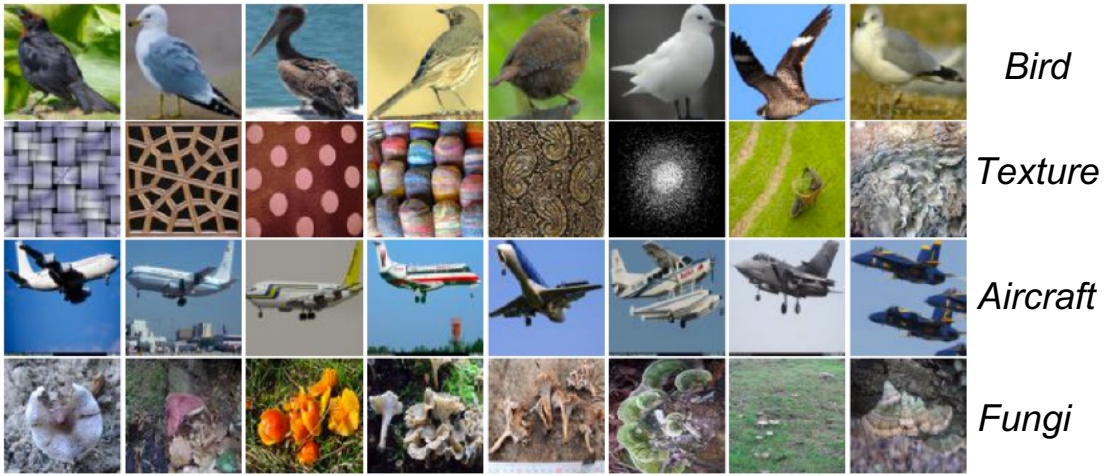
Background: MAML



MAML learns a globally-shared initialization for all tasks.

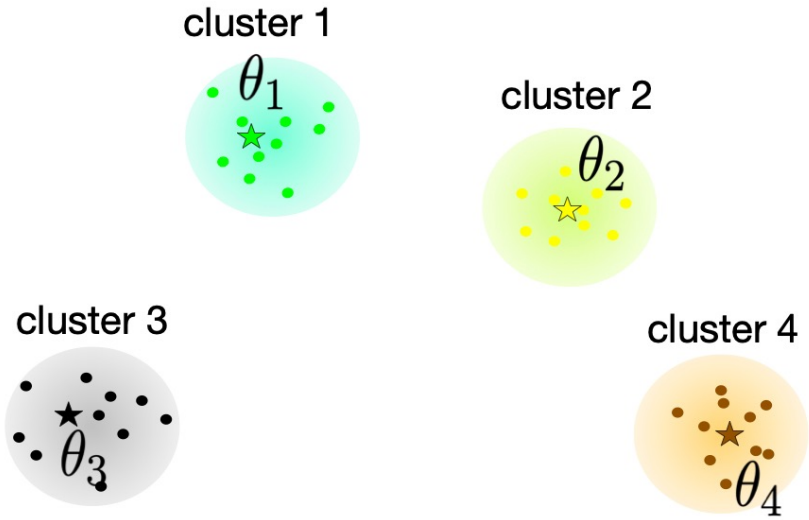
Model-agnostic meta-learning for fast adaptation of deep networks, Finn, C., Abbeel, P., and Levine, S., ICML 2017

Background: Complex Environment



A single meta-model may not be sufficient to capture all the meta-knowledge

Background: Structured Meta-Learning

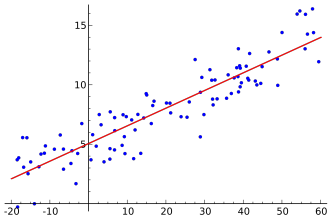


- Recent structured meta-learning methods (DPMM and TSA-MAML) cluster tasks into multiple groups .
- Cluster centroids form group-specific initializations.
- However, task model parameters may lie in a subspace mixture.

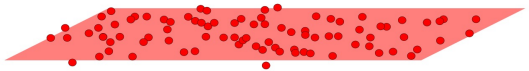
Reconciling meta-learning and continual learning with online mixtures of tasks. Jerfel, G., Grant, E., Griths, T., and Heller, K. A. NeurIPS 2019.
Task similarity aware meta learning: Theory-inspired improvement on MAML. Zhou, P., Zou, Y., Yuan, X., Feng, J., Xiong, C., and Hoi, S. UAI 2021.

Background: Research Gap

Prior work

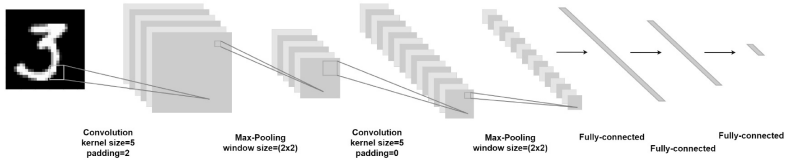


linear regression

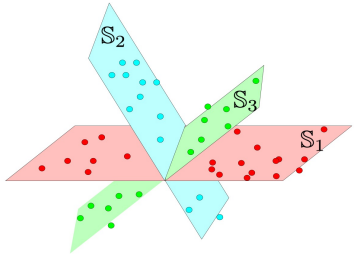


single subspace

This work



nonlinear models



a subspace mixture

Meta-learning for mixed linear regression. Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. ICML 2020.
Provable meta-learning of linear representations. Tripuraneni, N., Jin, C., and Jordan, M. ICML 2021.

MUSML: A Subspace Mixture

- task parameter w_τ's are assumed to lie in K subspaces $\{\mathbb{S}_1, \dots, \mathbb{S}_K\}$
- $S_k \in \mathbb{R}^{d \times m}$ is a basis of $\mathbb{S}_k$
- meta-parameters: $\{S_1, \dots, S_K\}$

MUSML: Base Learner

- in each $\mathbb{S}_k$, we search for a linear combination to form task model $w_\tau = S_k v^\star_{\tau,k}$:

$$v^\star_{\tau,k} = \arg\min_{v_\tau \in \mathbb{R}^m} \mathcal{L}(\mathcal{D}^{tr}_\tau; S_k v_\tau)$$

- use convex program when $\mathcal{L}(\mathcal{D}; w)$ is convex
- in nonconvex case, we seek **an approximate minimizer**: for $t' = 1, \ldots, T_{in}$,

$$v^{(t')}_{\tau,k} = v^{(t'-1)}_{\tau,k} - \alpha \nabla_{v^{(t'-1)}_{\tau,k}} \mathcal{L}\left(\mathcal{D}^{tr}_\tau; S_k v^{(t'-1)}_{\tau,k}\right)$$

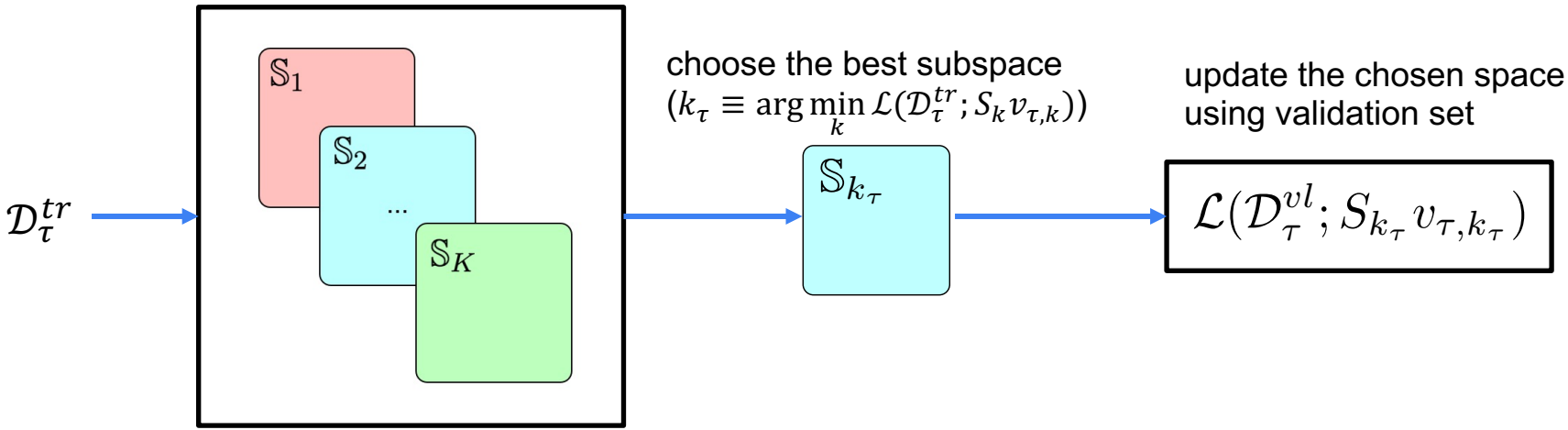
and then $v_{\tau,k} = v^{(T_{in})}_{\tau,k}$

Algorithm 1 MUltiple Subspaces for Meta-Learning (MUSML).

Require: stepsize α, $\{\eta_t\}$; number of inner gradient steps T_{in}, number of subspaces K, subspace dimension m, temperature $\{\gamma_t\}$; initialization $\mathbf{v}^{(0)}$;

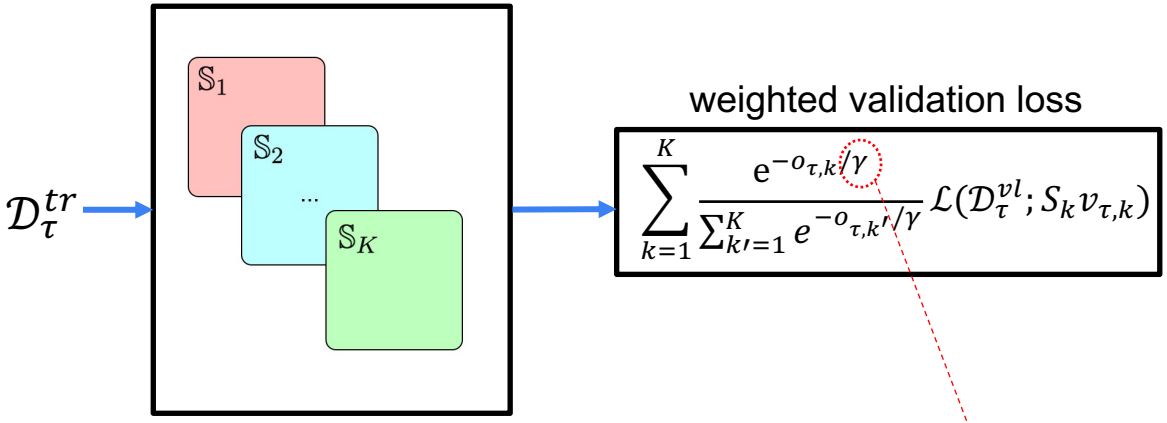
for $t = 0, 1, \ldots, T-1$ **do**
  sample a task τ with $\mathcal{D}^{tr}_\tau$ and $\mathcal{D}^{vl}_\tau$;
  base learner:
  for $k = 1, \ldots, K$ **do**
    initialize $\mathbf{v}^{(0)}_{\tau,k} = \mathbf{v}^{(0)}$;
    for $t' = 0, 1, \ldots, T_{in} - 1$ **do**
      $\mathbf{v}^{(t'+1)}_{\tau,k} = \mathbf{v}^{(t')}_{\tau,k} - \alpha \nabla_{\mathbf{v}^{(t')}_{\tau,k}} \mathcal{L}(\mathcal{D}^{tr}_\tau; \mathbf{S}_{k,t} \mathbf{v}^{(t')}_{\tau,k})$;
    end for
    $\mathbf{v}_{\tau,k} \equiv \mathbf{v}^{(T_{in})}_{\tau,k}$;
    $o_{\tau,k} = \mathcal{L}(\mathcal{D}^{tr}_\tau; \mathbf{S}_{k,t} \mathbf{v}_{\tau,k})$;
  end for
  meta-learner:
  $\mathcal{L}_{vl} = \sum_{k=1}^{K} \frac{\exp(-o_{\tau,k}/\gamma_t)}{\sum_{k'=1}^{K} \exp(-o_{\tau,k'}/\gamma_t)} \mathcal{L}(\mathcal{D}^{vl}_\tau; \mathbf{S}_{k,t} \mathbf{v}_{\tau,k})$;
  $\{\mathbf{S}_{1,t+1}, \ldots, \mathbf{S}_{k,t+1}\} = \{\mathbf{S}_{1,t}, \ldots, \mathbf{S}_{k,t}\} - \eta_t \nabla_{\{\mathbf{S}_{1,t}, \ldots, \mathbf{S}_{k,t}\}} \mathcal{L}_{vl}$;
end for
Return $\mathbf{S}_{1,T}, \ldots, \mathbf{S}_{K,T}$.

MUSML: Meta-Learner (one-hot selection)



inefficient: only one subspace is updated at each step.

MUSML: Meta-Learner (soft selection)



All spaces are updated simultaneously.

Algorithm 1 MUltiple Subspaces for Meta-Learning (MUSML).

Require: stepsize α, $\{\eta_t\}$; number of inner gradient steps T_{in}, number of subspaces K, subspace dimension m, temperature $\{\gamma_t\}$; initialization $\mathbf{v}^{(0)}$;

for $t = 0, 1, \ldots, T-1$ **do**
  sample a task τ with $\mathcal{D}_\tau^{tr}$ and $\mathcal{D}_\tau^{vl}$;
  base learner:
  for $k = 1, \ldots, K$ **do**
    initialize $\mathbf{v}_{\tau,k}^{(0)} = \mathbf{v}^{(0)}$;
    for $t' = 0, 1, \ldots, T_{in} - 1$ **do**
      $\mathbf{v}_{\tau,k}^{(t'+1)} = \mathbf{v}_{\tau,k}^{(t')} - \alpha \nabla_{\mathbf{v}_{\tau,k}^{(t')}} \mathcal{L}(\mathcal{D}_\tau^{tr}; \mathbf{S}_{k,t} \mathbf{v}_{\tau,k}^{(t')})$;
    end for
    $\mathbf{v}_{\tau,k} \equiv \mathbf{v}_{\tau,k}^{(T_{in})}$;
    $o_{\tau,k} = \mathcal{L}(\mathcal{D}_\tau^{tr}; \mathbf{S}_{k,t} \mathbf{v}_{\tau,k})$;
  end for
  meta-learner:
  $\mathcal{L}_{vl} = \sum_{k=1}^{K} \frac{\exp(-o_{\tau,k}/\gamma_t)}{\sum_{k'=1}^{K} \exp(-o_{\tau,k'}/\gamma_t)} \mathcal{L}(\mathcal{D}_\tau^{vl}; \mathbf{S}_{k,t} \mathbf{v}_{\tau,k})$;
  $\{\mathbf{S}_{1,t+1}, \ldots, \mathbf{S}_{k,t+1}\} = \{\mathbf{S}_{1,t}, \ldots, \mathbf{S}_{k,t}\} - \eta_t \nabla_{\{\mathbf{S}_{1,t}, \ldots, \mathbf{S}_{k,t}\}} \mathcal{L}_{vl}$;
end for
Return $\mathbf{S}_{1,T}, \ldots, \mathbf{S}_{K,T}$.

- $\gamma \to 0$, soft selection becomes one-hot
- $\gamma \to \infty$, soft selection becomes uniform
- In practice, a linear annealing schedule

Analysis

- Under some conditions, we obtain an upper bound for the population risk

$$\mathcal{R}(\mathcal{S}) \leq \mathcal{R}^{\star} + \rho\sqrt{m}\mathbb{E}_{\tau'}\mathbb{E}_{\mathcal{D}_{\tau'}^{tr}}\left\|v_{\tau',k_{\tau'}} - v_{\tau',k_{\tau'}}^{\star}\right\|^2 + \varrho\ \mathbb{E}_{\tau'}\mathbb{E}_{\mathcal{D}_{\tau'}^{tr}}\mathrm{dist}\left(\mathrm{w}_{\tau'}^{\star}, \mathbb{S}_{k_{\tau'}}\right) + K\sqrt{\frac{\nu^2+12\varrho\nu(1+m\alpha\delta)^{T_{in}}}{2N_{tr}}}$$

(1) minimum risk

(2) distance between approximate minimizer and exact minimizer

(3) approximation error using learned subspaces

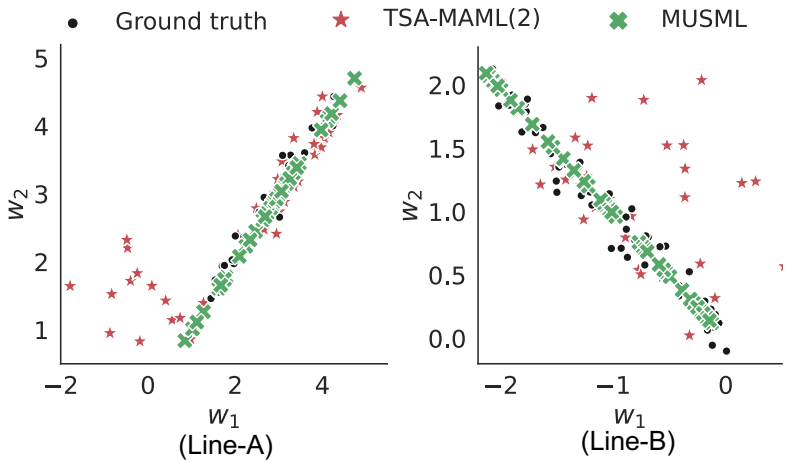
(4) complexity of subspaces (K and m)

Experiments: Few-Shot Regression (Synthetic Data : Setup)

- a nonlinear model: $f(x; w_\tau) = e^{0.1 w_{\tau,1} x} + w_{\tau,2} |\sin(x)|$;
- $w_\tau = [w_{\tau,1}; w_{\tau,2}]$ is sampled from one of the two subspaces:
  - *Line-A*: $w_\tau = S_1 a_\tau + 0.1\xi_\tau$, where $S_1 = [1; 1], a_\tau \sim \mathcal{U}(1,5), \xi_\tau \sim \mathcal{N}(0, I)$;
  - *Line-B*: $w_\tau = S_2 a_\tau + 0.1\xi_\tau$, where $S_2 = [-1; 1], a_\tau \sim \mathcal{U}(0,2), \xi_\tau \sim \mathcal{N}(0, I)$;
- samples: $y = f(x; w_\tau) + 0.05\xi$, where $x \sim \mathcal{U}(-5,5)$ and $\xi \sim \mathcal{N}(0,1)$;
- $K = 2, m = 1$.
- γ_t: a linear annealing schedule

Experiments: Few-Shot Regression (Synthetic Data : Results)

visualization of task model parameters



➢ MUSML discovers underlying subspaces

meta-testing MSE

MAML	0.74
BMG	0.67
DPMM	0.56
HSML	0.49
ARML	0.60
TSA-MAML	0.58
MUSML	**0.07**

➢ MUSML performs the best.

Experiments: Few-shot Classification (Setup)

- meta-datasets:
  1. *Meta-Dataset-BTAF*
  2. *Meta-Dataset-ABF*
  3. *Meta-Dataset-CIO*
- $f(\cdot; w)$: *Conv4* backbone + Prototype classifier
- K and m are chosen from 1 to 5

	#classes (meta-train/meta-valid/meta-test)
Bird	64/16/20
Texture	30/7/10
Aircraft	64/16/20
Fungi	64/16/20
CIFAR-FS	64/16/20
Mini-ImageNet	64/16/20
Omniglot	71/15/16

Experiments: Few-shot Classification (Results)

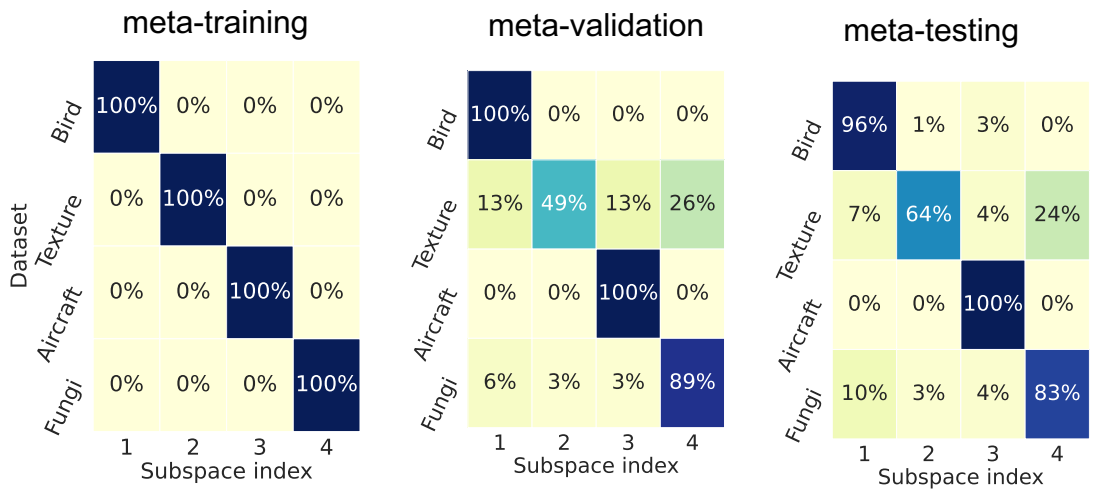
Accuracy of 5-way 5-shot classification

	Meta-Dataset-BTAF	*Meta-Dataset-ABF*	*Meta-Dataset-CIO*
MAML	57.78	63.86	74.46
ProtoNet	62.29	65.62	76.51
ANIL	58.57	64.43	74.61
BMG	60.10	65.80	77.46
DPMM	63.00	66.26	76.63
TSA-MAML	63.20	68.17	76.89
HSML	62.39	64.17	75.54
ARML	63.95	64.52	76.12
TSA-ProtoNet	63.57	68.77	77.27
MUSML	**66.18**	**71.10**	**77.83**

- MUSML is more accurate than both structured and unstructured meta-learning methods.

Experiments: Few-shot Classification (Results)

task assignment to the learned subspaces
(5-way 5-shot setting on *Meta-Dataset-BTAF*)



- MUSML discovers task structure.

Experiments: Cross-Domain Few-shot Classification

accuracy of cross-domain 5-way 5-shot setting
(*Meta-Dataset-BTAF -> Meta-Dataset-CIO*)

MAML	64.25
ProtoNet	66.13
ANIL	65.19
BMG	66.98
DPMM	66.73
TSA-MAML	66.85
HSML	65.18
ARML	65.37
TSA-ProtoNet	66.92
MUSML	**67.41**

- MUSML is also effective on unseen domains.

Experiments: Improving Existing Meta-learning Algorithms

accuracy of 5-way 5-shot classification

	Meta-Dataset-BTAF	***Meta-Dataset-ABF***	***Meta-Dataset-CIO***
Meta-SGD	58.93	64.19	75.95
MUSML-SGD	65.72	69.15	77.48
Meta-Curvature	60.02	64.51	76.13
MUSML-Curvature	**66.10**	**69.23**	**77.96**

- A subspace mixture is beneficial for both Meta-SGD and Meta-Curvature.
- MUSML can be used for any meta-learning algorithms.

Summary

- Problem: meta-learning in complex environment (task models are diverse)
- MUSML: learning a subspace mixture for building task models
- Each subspace can be viewed as a type of meta-knowledge
- Experimental results verify the effectiveness of MUSML