NEURAL INFORMATION PROCESSING SYSTEMS

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Revisit Multimodal Meta-Learning through the Lens of Multi-Task Learning

Milad Abdollahzadeh, Touba Malekzadeh, Ngai-Man Cheung

Singapore University of Technology and Design

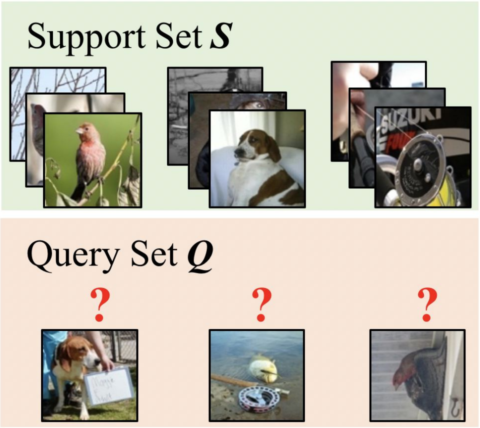
NeurIPS 2021

Meta-Learning

Meta-learning aims to improve the learning algorithm

Providing an opportunity to tackle many challenges of deep learning: data-efficient AI

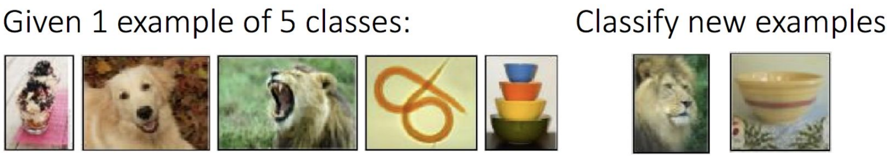
Meta-Learning as a solution for few-shot learning



Meta-Learning

Episodic Learning

Match the condition in which the model is trained (meta-train) and tested (meta-test)

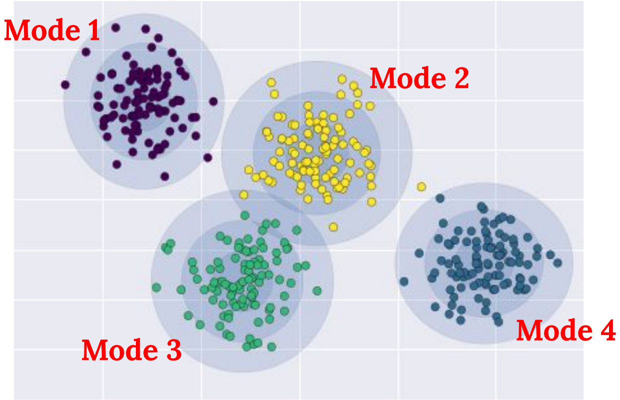


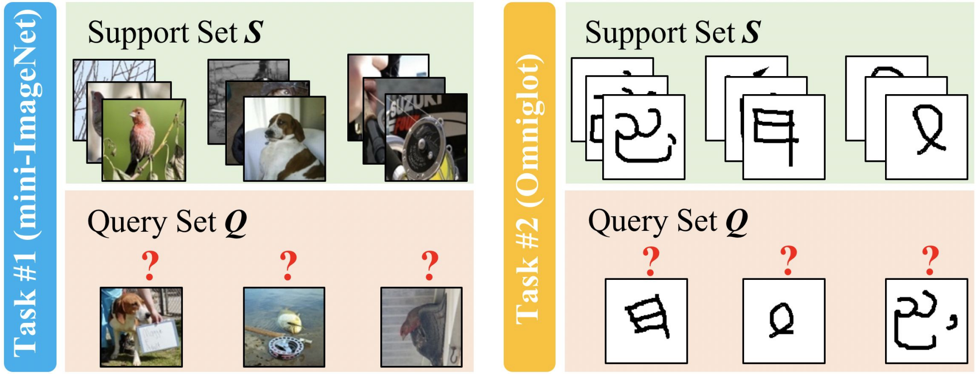
5-way, 1-shot image classification problem

Multimodal Meta-Learning

Generalizing the conventional setup for more diverse task distributions [1]

Multimodality in task distribution





More challenging to current meta-learners

[1] Vuorio, Risto, et al. "Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation." *Advances in Neural Information Processing Systems* 32 (2019): 1-12.

Research Gap

Main Claim of previous work:
Improving generalization by transferring knowledge between different modes of task distributions

However, there are two main questions that need to be addressed:

I) Not clear how task from one mode impacts the learning of task from another mode

II) Can we make a better use of knowledge transfer and have a better generalization?

Our Contributions

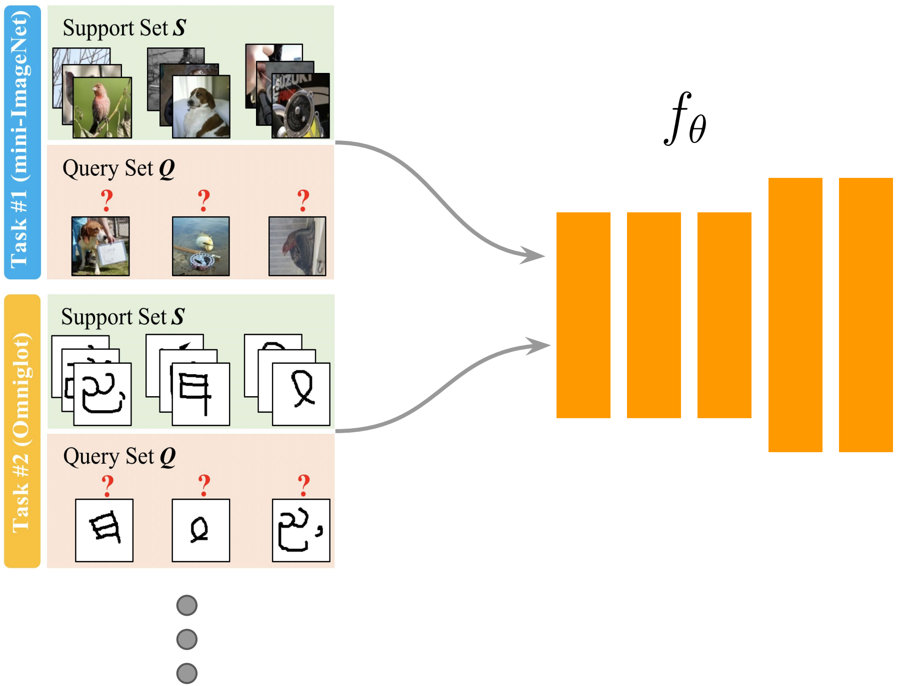
To **address** the discussed research gaps:

1) We extend the **transference idea** [1] from multi-task learning to episodic learning scenario of meta-learning to analyse information transfer between few-shot tasks
2) Propose **a new multimodal meta-learning** called **K**ernel **M**odu**L**ation (**KML**) which significantly advances state-of-the-art

[1] Fifty, Christopher, et al. "Measuring and Harnessing Transference in Multi-Task Learning." arXiv preprint arXiv:2010.15413 (2020).

Analysis of Information Transfer (Transference)

Episodic learning from the lens of Multi-Task learning



Information Transfer (Transference):
the effect of gradients update from one episode to the network parameters on the generalization performance on other episodes

Analysis of Information Transfer (Transference)

Measure transference from source (meta-train) task *i* to target (meta-test) task *j*

Calculate the loss on target task before and after updating parameters wrt source task

The ratio between the loss of task j after and before parameter update w.r.t task i

$$LR_{i \to j} = \frac{\mathcal{L}_{\mathcal{T}_j}(\mathcal{Q}_j; \theta_i^{t+1}, \mathcal{S}_j)}{\mathcal{L}_{\mathcal{T}_j}(\mathcal{Q}_j; \theta^t, \mathcal{S}_j)}$$

LR<1 : **Positive knowledge transfer** from source task i to target task j

LR>1 : **Negative knowledge transfer** from source task i to target task j

Analysis of Information Transfer (Transference)

Transference from source (meta-train) task i to target (meta-test) task j

#1. Calculate the loss of task j with initial model parameters

  a) adapt model using the support set of target task j

Initial model (θ^t)

Support Set

$\mathcal{S}_j$

Adapt the model to task j

Analysis of Information Transfer (Transference)

Transference from source (meta-train) task *i* to target (meta-test) task *j*

#1. Calculate the loss of task j with initial model parameters

  b) calculate the loss of adapted model using query set of target task *j*

Adapted model ($\theta^t, \mathcal{S}_j$)

Query Set

$\mathcal{Q}_j$

Loss on Query Set

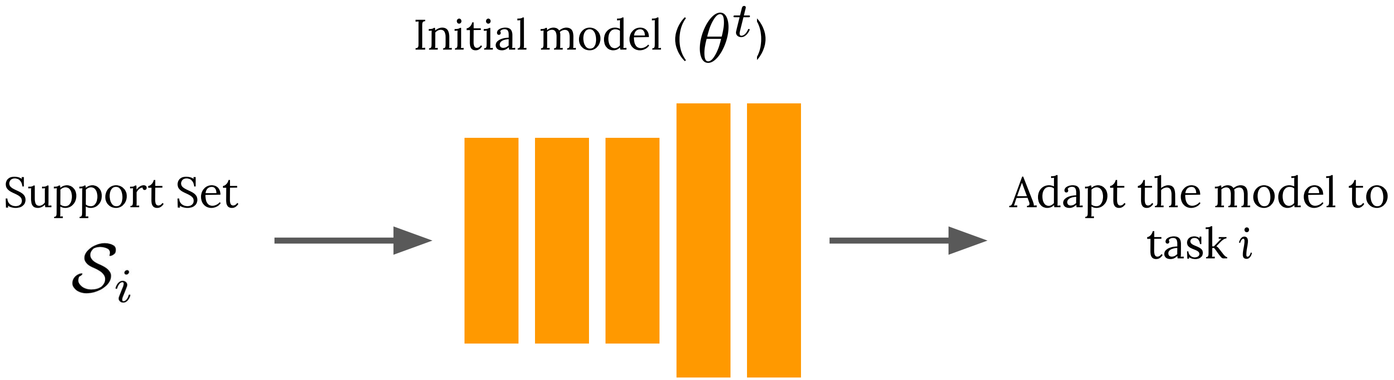
$\mathcal{L}_{\mathcal{T}_j}(\mathcal{Q}_j; \theta^t, \mathcal{S}_j)$

Analysis of Information Transfer (Transference)

Transference from source (meta-train) task i to target (meta-test) task j

#2. Update the initial model with respect to task i

  a) adapt the model using support set of source task i



Analysis of Information Transfer (Transference)

Transference from source (meta-train) task i to target (meta-test) task j

#2. Update the initial model with respect to task i

  b) calculate the loss of adapted model using query set of source task i

Adapted model ($\theta^t, \mathcal{S}_i$)

Support Set

$\mathcal{Q}_i$

Adapt the model to task i

$\mathcal{L}_{\mathcal{T}_i}(\mathcal{Q}_i; \theta^t, \mathcal{S}_i)$

Analysis of Information Transfer (Transference)

Transference from source (meta-train) task *i* to target (meta-test) task *j*

#2. Update the initial model with respect to task i

  c) use the gradient of the loss to update model parameters

Updated model (θ_i^{t+1})

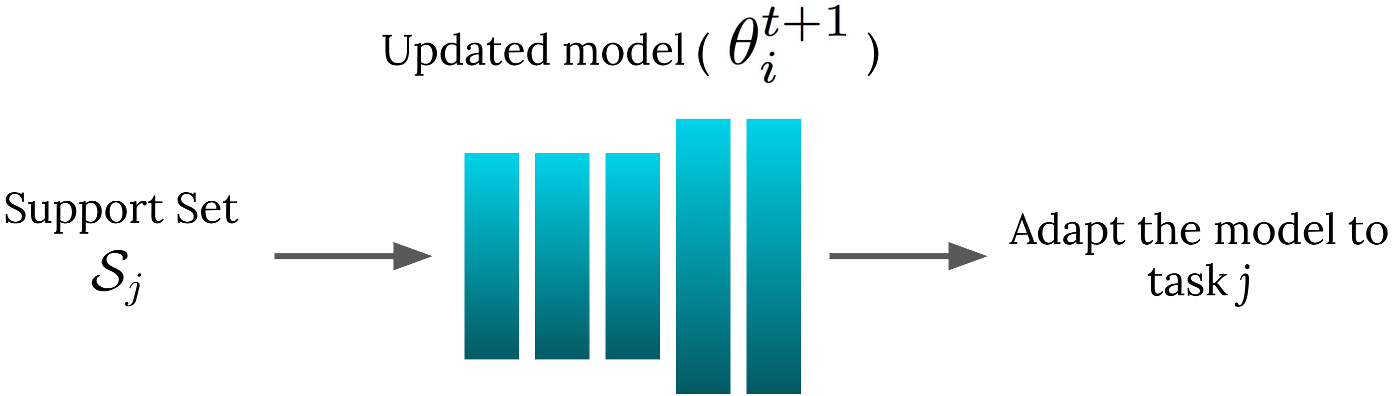
$$\theta_i^{t+1} = \theta^t - \alpha \nabla_{\theta^t} \mathcal{L}_{\mathcal{T}_i}(\mathcal{Q}_i; \theta^t, \mathcal{S}_i)$$

Analysis of Information Transfer (Transference)

Transference from source (meta-train) task *i* to target (meta-test) task *j*

#3. Calculate the loss of task j with updated model parameters

  a) adapt model using the support set of target task *j*

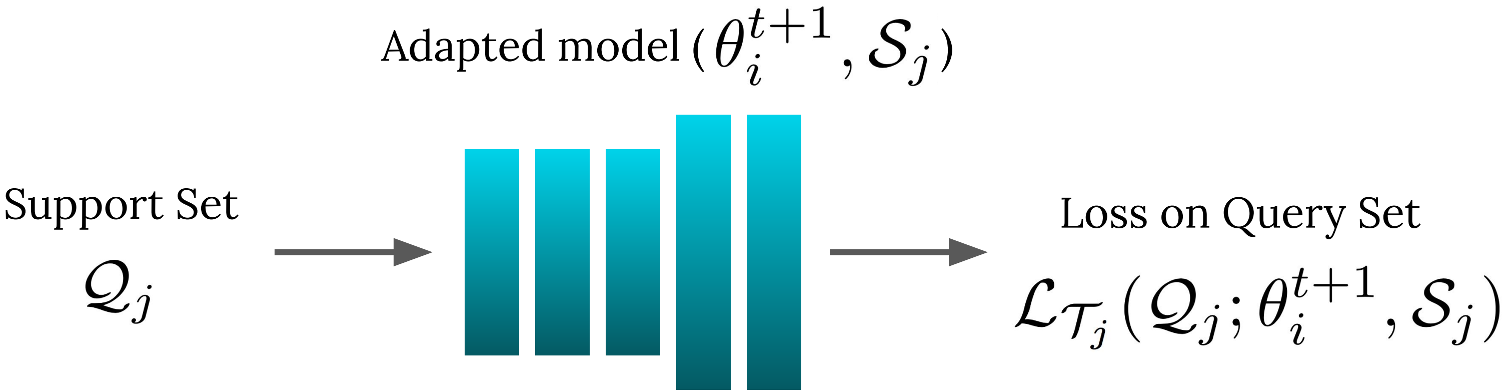


Analysis of Information Transfer (Transference)

Transference from source (meta-train) task *i* to target (meta-test) task *j*

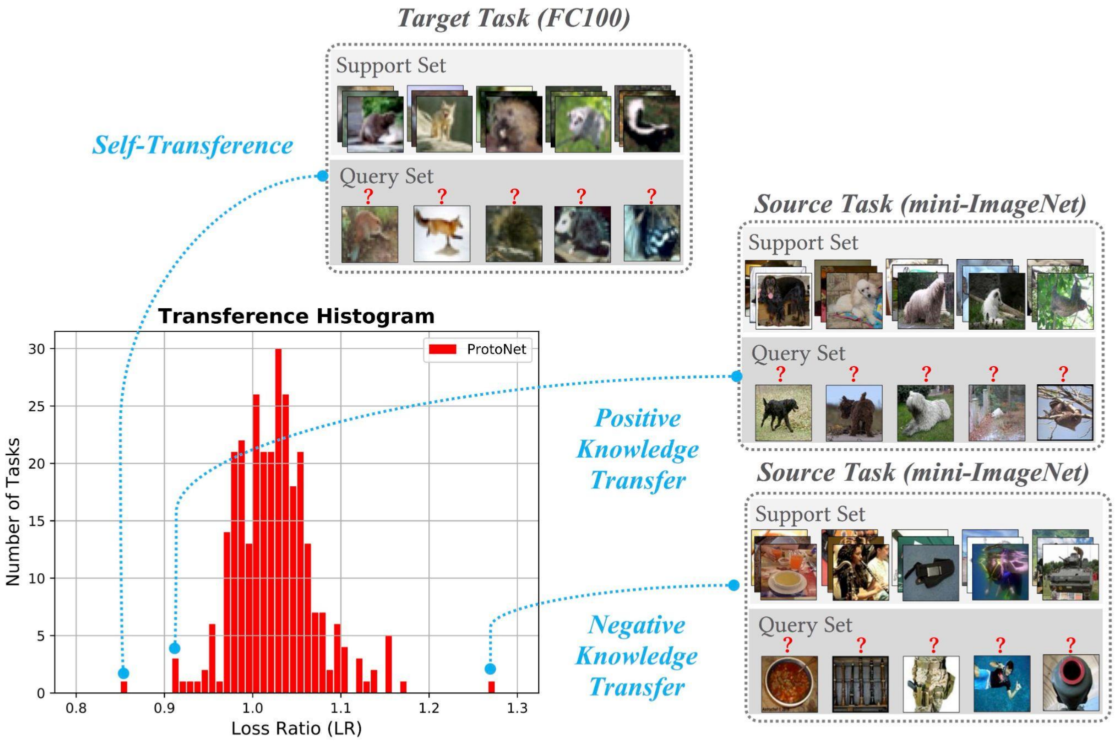
#3. Calculate the loss of task *j* with updated model parameters

  b) calculate the loss of adapted model using query set of target task *j*



Analysis of Information Transfer (Transference)

Transference histogram from 300 meta-train mini-ImageNet tasks to a meta-test FC100 target task

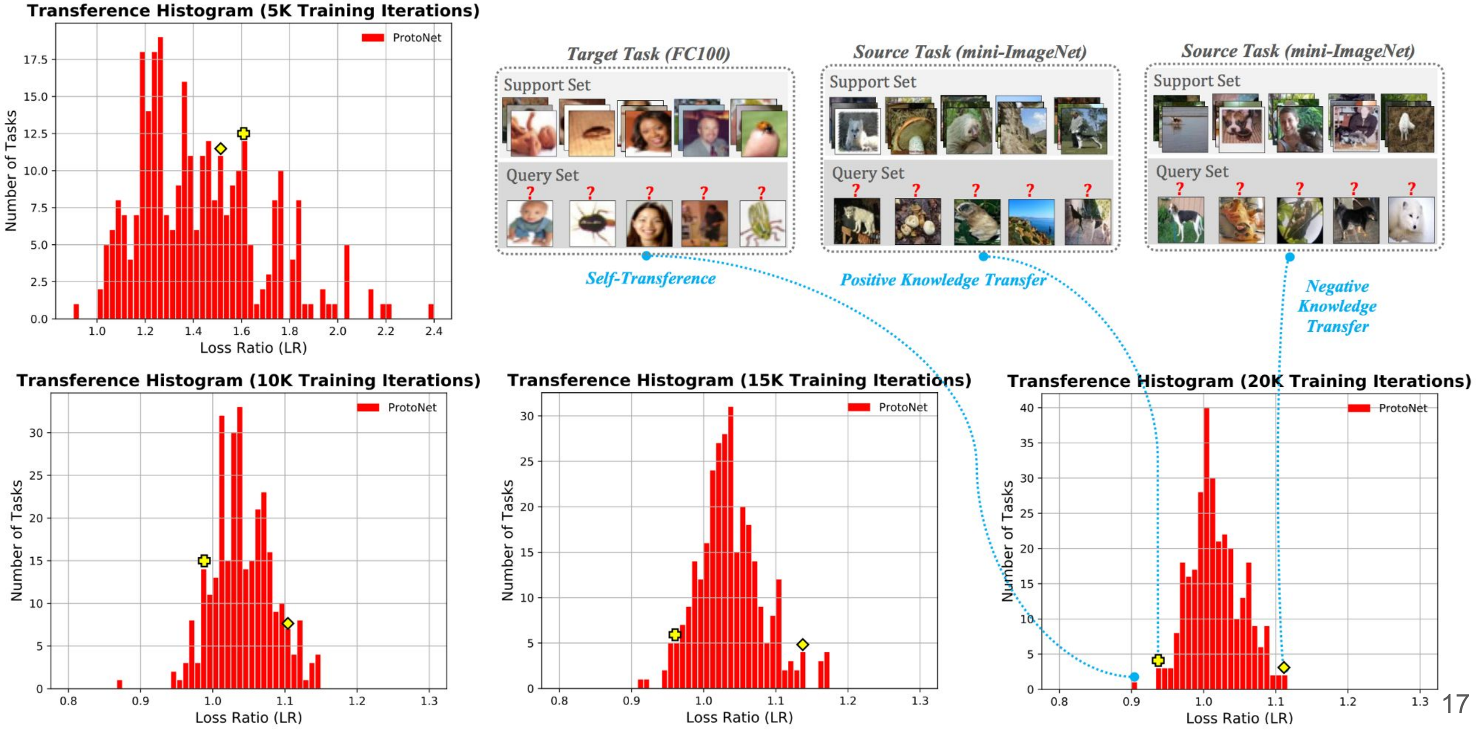


LR<1: Positive Knowledge Transfer
LR>1: Negative Knowledge Transfer

Analysis of Information Transfer (Transference)

Transference Histogram During Training

More positive transfer as training proceeds



Towards Reducing Negative Knowledge Transfer

Analysis results

both positive and negative transference

How does MTL reduce negative transfer?

hard parameter sharing and grouping tasks during training

Ideal learning episodes in our episodic training scenario

most compatible with meta-test task

Grouping and hard parameter sharing is not possible in episodic training

1. meta-test tasks are unseen and unknown during meta-training
2. episodic training involves tens of thousands of tasks

Solution: Task-Aware Layers

Proposed Multimodal Meta-Learner

MMAML General Framework [1]

Task Encoder

$$\boldsymbol{\upsilon}_{\mathcal{T}} = h_{\varphi}(\mathcal{S}_{\mathcal{T}})$$

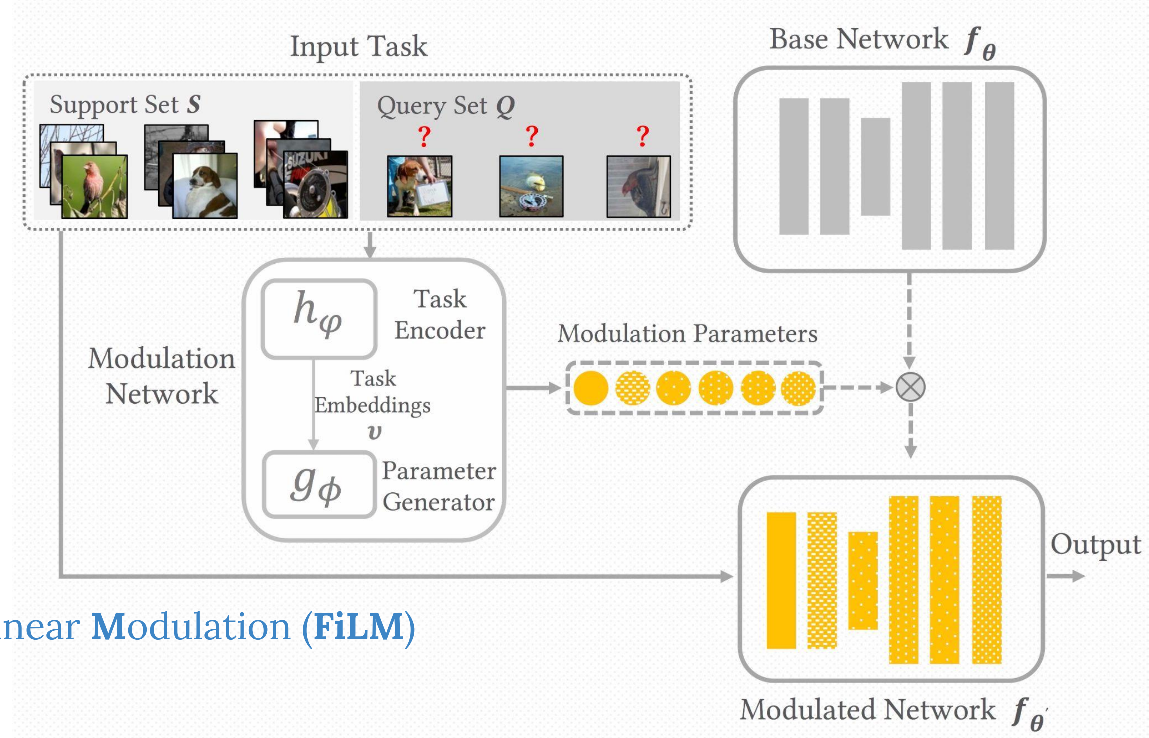
Parameter Generator

$$\boldsymbol{\omega}_{\mathcal{T}} = g_{\phi}(\boldsymbol{\upsilon}_{\mathcal{T}})$$

Modulation Scheme

$$\boldsymbol{\omega}_{\mathcal{T}} = \{\boldsymbol{\eta}_{\mathcal{T}}, \boldsymbol{\gamma}_{\mathcal{T}}\}$$

$$\hat{\mathbf{Y}}_i = \eta_i \mathbf{Y}_i + \gamma_i$$ Feature-wise Linear Modulation (**FiLM**)



[1] Vuorio, Risto, et al. "Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation." *Advances in Neural Information Processing Systems* 32 (2019): 1-12.

Proposed Multimodal Meta-Learner

Limitation of FiLM for Multimodal Meta-Learning

Convolution operator for each channel

$$\mathbf{Y}_i = \mathbf{W}_i * \mathbf{X} + b_i$$

Re-writing FiLM modulation

$$\hat{\mathbf{Y}}_i = (\eta_i \mathbf{W}_i) * \mathbf{X} + (\eta_i b_i + \gamma_i)$$

New interpretation of FiLM modulation:

Convolution with modulated parameters

$$\hat{\mathbf{W}}_i = \eta_i \mathbf{W}_i \qquad \hat{b}_i = \eta_i b_i + \gamma_i$$

modulated kernel **modulated bias**

Proposed Multimodal Meta-Learner

Proposed **K**ernel **Modu****L**ation (**KML**)

Modulate every parameter within network

$$\hat{\mathbf{W}}^l_{\mathcal{T}} = \mathbf{W}^l \odot (\mathbf{J} + \mathbf{M}^l(\boldsymbol{\upsilon}_{\mathcal{T}}, \phi))$$

$$\hat{\mathbf{b}}^l_{\mathcal{T}} = \mathbf{b}^l + \Delta\mathbf{b}^l(\boldsymbol{\upsilon}_{\mathcal{T}}, \phi)$$

Modulated parameters for whole network

$$\hat{\theta}_{\mathcal{T}} = \{\hat{\mathbf{W}}^1_{\mathcal{T}}, \ldots, \hat{\mathbf{W}}^L_{\mathcal{T}}, \hat{\mathbf{b}}^1_{\mathcal{T}}, \ldots, \hat{\mathbf{b}}^L_{\mathcal{T}}\}$$

Generator Design for KML

- use **MLP** to be computationally efficient
- Limitation: large number of required parameters in MLP

Proposed Multimodal Meta-Learner

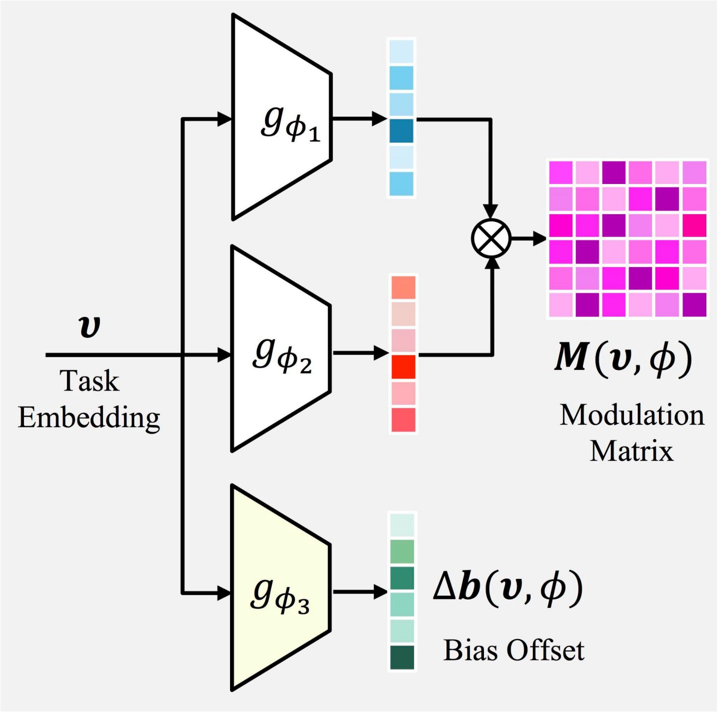
Proposed Simplified Structure for Parameter Generator

Use multiple smaller MLPs instead of a large one

$$\mathbf{M}^l(\boldsymbol{v}_\mathcal{T}, \phi) = \mathbf{g}^l_{\phi_1}(\boldsymbol{v}_\mathcal{T}) \otimes \mathbf{g}^l_{\phi_2}(\boldsymbol{v}_\mathcal{T})$$

$$\Delta\mathbf{b}^l(\boldsymbol{v}_\mathcal{T}, \phi) = \mathbf{g}^l_{\phi_3}(\boldsymbol{v}_\mathcal{T})$$

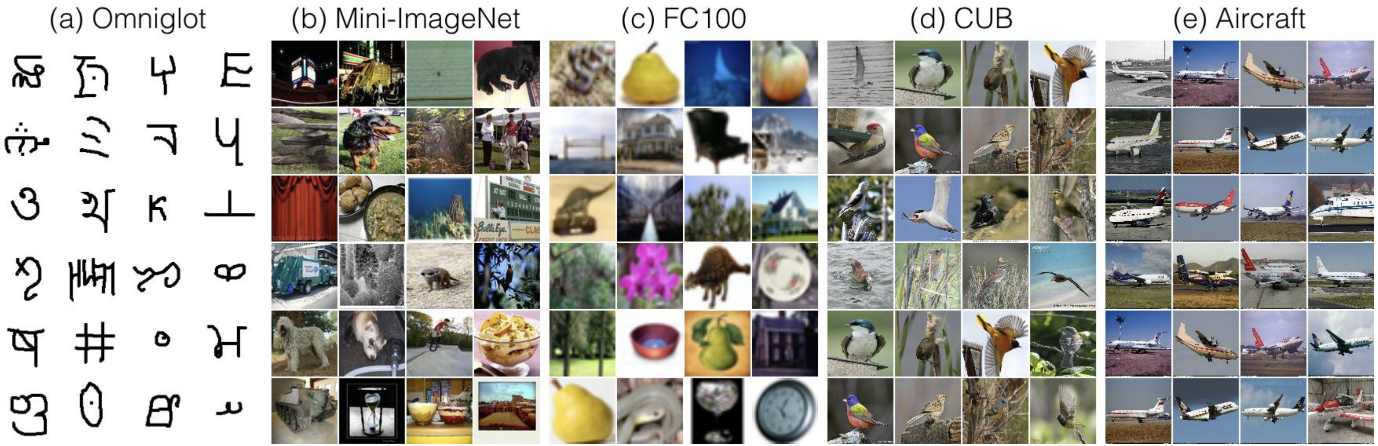
- Reduces the required parameters by a factor of **152**
- Improves generalization performance



Experimental Results

Datasets

mini-ImageNet:
natural objects
FC100:
natural objects with lower resolution
Omniglot:
handwritten characters
CUB:
fine-grained bird classification
Aircraft:
fine-grained aircraft classification



Datasets used to generate meta-dataset for multimodal meta-learning following MMAML

2Mode†: mini-ImageNet and FC100; **2Mode**: mini-ImageNet and Omniglot
3Mode: mini-ImageNet, Omniglot and FC100; **5Mode**: mini-ImageNet, Omniglot, FC100, AIRCRAFT and CUB

Experimental Results

Multimodal Few-Shot Classification Results

Setup		Method			
		MAML [19]*	Multi-MAML	MMAML [1]*	MMAML+KML (ours)
2Mode†	1-shot	40.53±68%	39.27±0.76%	39.11±0.62%	**40.73±0.66%**
	5-shot	**54.11±0.63%**	53.51±0.72%	52.02±0.63%	53.72±0.60%
2Mode	1-shot	65.18±0.61%	66.77±0.68%	67.67±0.63%	**68.01±0.59%**
	5-shot	74.18±0.57%	73.07±0.61%	73.52±0.71%	**77.02±0.66%**
3 Mode	1-shot	54.40±0.56%	56.01±0.66%	57.35±0.61%	**57.68±0.59%**
	5-shot	66.51±0.54%	65.92±0.62%	64.21±0.57%	**67.12±0.55%**
5Mode	1-shot	47.19±0.49%	48.33±0.58%	49.53±0.50%	**50.31±0.49%**
	5-shot	58.13±0.48%	59.20±0.52%	58.89±0.47%	**60.51±0.47%**

Experimental Results

Multimodal Few-Shot Classification Results

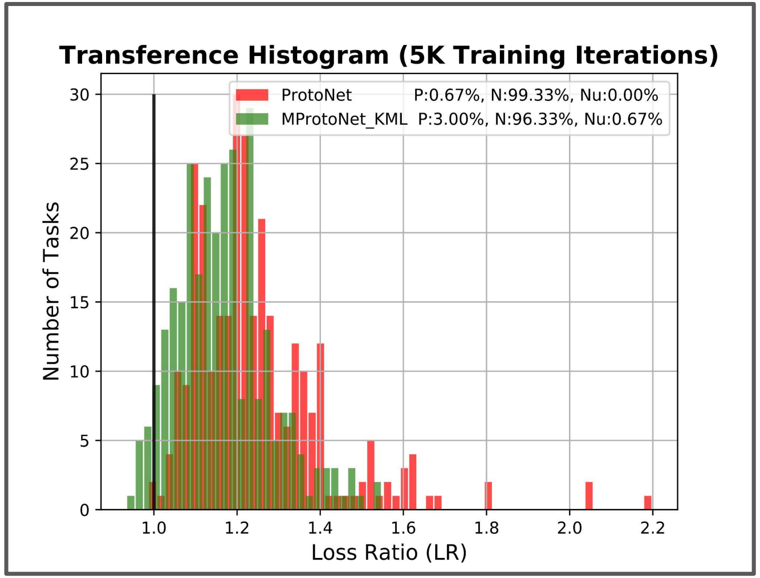
Setup		Method			
		ProtoNet [11]**	Multi-ProtoNet	MProtoNet [1]**	MProtoNet+KML (ours)
2Mode†	1-shot	43.05±0.58%	43.42±0.56%	43.57±0.59%	**44.40±0.65%**
	5-shot	57.70±0.59%	56.73±0.64%	56.03±0.64%	**59.31±0.62%**
2Mode	1-shot	69.55±0.54%	70.17±0.61%	70.60±0.56%	**73.69±0.52%**
	5-shot	75.12±0.41%	75.33±0.46%	75.72±0.47%	**79.82±0.40%**
3 Mode	1-shot	58.14±0.49%	59.89±0.50%	59.62±0.54%	**62.08±0.54%**
	5-shot	66.84±0.44%	67.03±0.44%	67.51±0.47%	**70.03±0.43%**
5Mode	1-shot	49.31±0.53%	50.69±0.57%	51.75±0.52%	**56.72±0.46%**
	5-shot	58.91±0.51%	59.88±0.54%	59.95±0.42%	**64.91±0.38%**

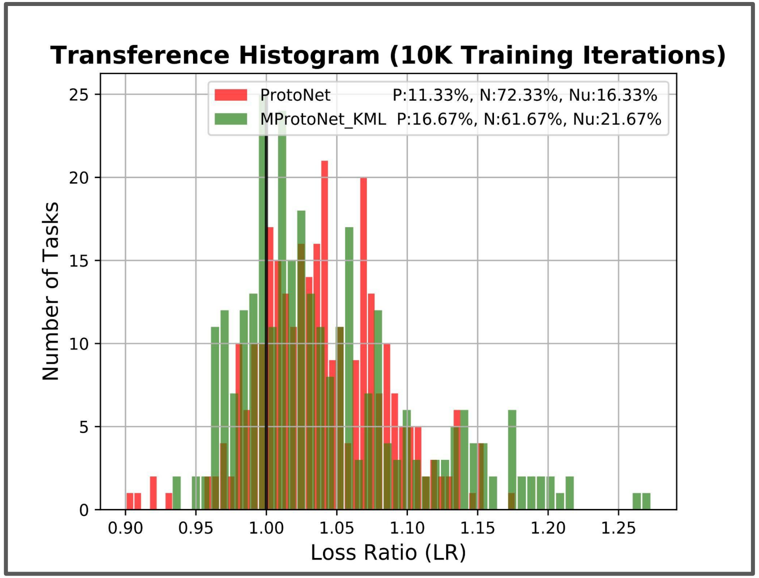
Experimental Results

Transference Results (mini-ImageNet ⇒ FC100)

$$LR_{i \to j} = \frac{\mathcal{L}_{\mathcal{T}_j}(\mathcal{Q}_j; \theta_i^{t+1}, \mathcal{S}_j)}{\mathcal{L}_{\mathcal{T}_j}(\mathcal{Q}_j; \theta^{t}, \mathcal{S}_j)}$$

LR<1: Positive Knowledge Transfer





Summary

Research gaps in multimodal meta-learning

How can we measure interaction between few-shot tasks?

How can we improve the generalization performance?

Proposed work

Adapt transference idea from MTL to quantify interaction between few-shot tasks

A new interpretation of FiLM scheme

Kernel modulation to improve generalization

Experimental Results

Significant improvement over previous state-of-the art in both micro and macro-level

Transference analysis and proposed KML can be extended to **conventional meta-learning** (Supplementary Material).