MetaMix: Improved Meta-Learning with Interpolation-based Consistency Regularization

Yangbin Chen[1], Yun Ma[2], Tom Ko[3], Jianping Wang[1], Qing Li[2]

1. City University of Hong Kong 2. The Hong Kong Polytechnic University 3. Southern University of Science and Technology

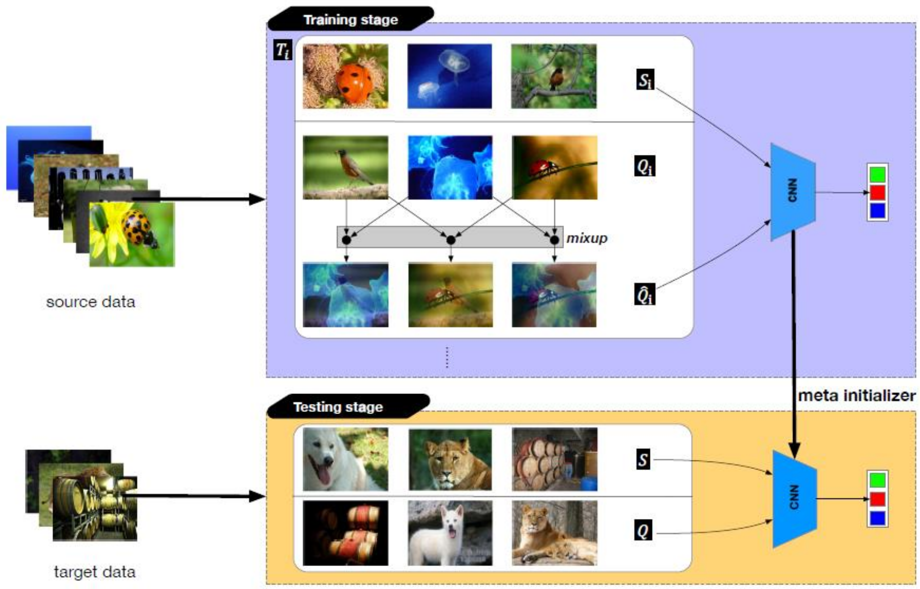
BACKGROUND

- Few-Shot Learning (FSL) problem is a machine learning problem that learns with limited labelled data of the target tasks by incorporating external source data with a different distribution.
- Few-Shot Classification is a few-shot learning task defined as N-way, K-shot, where N is the number of classes in the target task and K is the number of labelled examples per class.
- Model-Agnostic Meta-Learning (MAML) and its variants aim to train a model, which can adapt quickly to any new tasks using only a few examples.

MOTIVATION

- Conventional meta-learning algorithms face meta-overfitting problems, where the learned decision boundary stays too close to the limited labelled examples in few-shot classification tasks.
- The Empirical Risk Minimization (ERM) allows large neural networks to memorize (rather than generalize from) the training data.
- We aim to propose a regularization technique to solve the meta-overfitting problem.

METHODOLOGY



Algorithm 1 MetaMix with MAML

Require: $p(\mathcal{T})$: distribution over tasks
Require: $\mathcal{S}_i$: support set; $\mathcal{Q}_i$: query set
Require: α, β : learning rate
Require: $\breve{\alpha}$: Beta distribution parameter
Require: $mix_\lambda(a,b) = \lambda a + (1-\lambda)b, \lambda \sim \mathrm{B}(\breve{\alpha}, \breve{\alpha})$

```
Randomly initialize model parameters θ
while not done do
    Sample a batch of episodes T_i ~ p(T)
    for all T_i do
        Sample a support set S_i = {(x_j, y_j)}_{j=1}^J
        Evaluate ∇_θ L_{S_i}(f_θ) using S_i and L_{S_i}(f_θ)
        Compute adapted parameters with gradient descent: θ'_i = θ − α · ∇_θ L_{S_i}(f_θ)
        Sample a query set Q_i = {(x_z, y_z)}_{z=1}^Z
        Randomly select pairs of examples {(x_m, y_m)}_{m=1}^Z, {(x_n, y_n)}_{n=1}^Z from Q_i
        x̂_z = mix_λ(x_m, x_n), ŷ_z = mix_λ(y_m, y_n)
        Get new query set Q̂_i = {(x̂_z, ŷ_z)}_{z=1}^Z
    end for
    Update θ ← θ − β · ∇_θ Σ_i L_{Q̂_i}(f_{θ'_i})
end while
```

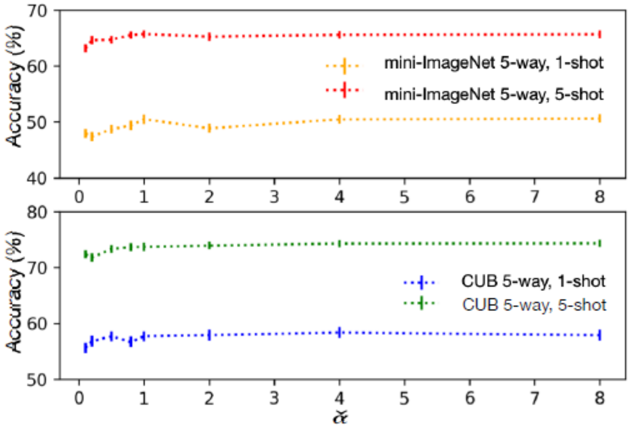
EXPERIMENT

- **Performance comparison of MetaMix and baseline approaches on 5-way classification tasks over three datasets**

Models	*mini*-ImageNet 1-shot	*mini*-ImageNet 5-shot	CUB 1-shot	CUB 5-shot	FC100 1-shot	FC100 5-shot
Matching Network	50.47 ± 0.80	64.83 ± 0.67	57.70 ± 0.87	71.42 ± 0.71	36.97 ± 0.67	49.44 ± 0.71
Prototypical Network	49.33 ± 0.82	65.71 ± 0.67	51.34 ± 0.86	67.56 ± 0.76	36.83 ± 0.69	51.21 ± 0.74
Relation Network	50.48 ± 0.80	65.39 ± 0.72	59.47 ± 0.96	73.88 ± 0.74	36.40 ± 0.69	51.35 ± 0.69
MAML	48.18 ± 0.78	63.05 ± 0.71	54.32 ± 0.91	71.37 ± 0.76	35.96 ± 0.71	48.06 ± 0.73
MetaMix+MAML	**50.51 ± 0.86**	**65.73 ± 0.72**	**57.70 ± 0.92**	**73.66 ± 0.74**	**37.09 ± 0.74**	**49.31 ± 0.72**
FOMAML	45.22 ± 0.77	60.97 ± 0.70	53.12 ± 0.93	70.90 ± 0.75	34.97 ± 0.70	47.41 ± 0.73
MetaMix+FOMAML	**47.78 ± 0.77**	**63.55 ± 0.70**	**54.81 ± 0.97**	**72.90 ± 0.74**	**36.48 ± 0.67**	**49.48 ± 0.71**
MetaSGD	49.93 ± 1.73	64.01 ± 0.90	56.19 ± 0.92	69.14 ± 0.75	36.36 ± 0.66	49.96 ± 0.72
MetaMix+MetaSGD	**50.60 ± 1.80**	**64.47 ± 0.88**	**57.64 ± 0.88**	**70.50 ± 0.70**	**37.44 ± 0.71**	**51.41 ± 0.69**
MTL	61.37 ± 0.82	78.37 ± 0.60	71.90 ± 0.86	84.68 ± 0.53	42.17 ± 0.79	56.84 ± 0.75
MetaMix+MTL	**62.74 ± 0.82**	**79.11 ± 0.58**	**73.04 ± 0.86**	**86.10 ± 0.50**	**43.58 ± 0.73**	**58.27 ± 0.73**

EXPERIMENT

- **Effect of Beta distribution**

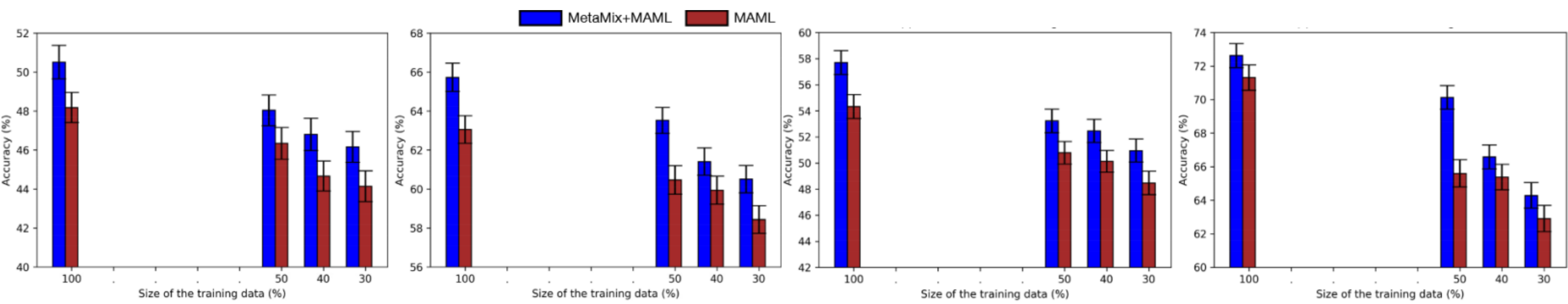


- **Effect of mixup on different sets**

Set(s)	*mini*-ImageNet 1-shot	*mini*-ImageNet 5-shot	CUB 1-shot	CUB 5-shot
Q	**50.51 ± 0.86**	**65.73 ± 0.72**	**57.70 ± 0.92**	**73.66 ± 0.74**
S	44.03 ± 0.79	53.74 ± 0.81	49.12 ± 0.96	63.27 ± 0.89
Q+S	48.36 ± 0.81	64.06 ± 0.72	54.32 ± 0.93	70.30 ± 0.75
w/o MetaMix	48.18 ± 0.78	63.05 ± 0.71	54.32 ± 0.91	71.37 ± 0.76

- **Effect of size of training data**

Models	*mini*-ImageNet 1-shot	*mini*-ImageNet 5-shot	CUB 1-shot	CUB 5-shot	FC100 1-shot	FC100 5-shot
MAML(100%)	48.18 ± 0.78	63.05 ± 0.71	54.32 ± 0.91	71.37 ± 0.76	35.96 ± 0.71	48.06 ± 0.73
MetaMix+MAML(100%)	**50.51 ± 0.86**	**65.73 ± 0.72**	**57.70 ± 0.92**	**73.66 ± 0.74**	**37.09 ± 0.74**	**49.31 ± 0.72**
MAML(50%)	46.34 ± 0.82	60.47 ± 0.73	50.78 ± 0.86	65.60 ± 0.81	35.38 ± 0.71	47.93 ± 0.78
MetaMix+MAML(50%)	**48.04 ± 0.79**	**63.52 ± 0.67**	**53.22 ± 0.91**	**70.13 ± 0.70**	**36.35 ± 0.74**	**48.11 ± 0.69**



(a) 1-shot task on *mini*-imagenet (b) 5-shot task on *mini*-imagenet (c) 1-shot task on *CUB* (d) 5-shot task on *CUB*

CONTRIBUTIONS

- We propose MetaMix as a regularization technique, which can be integrated with many meta-learning algorithms, including MAML and its variants, and improve their performance.
- MetaMix with MAML-based algorithms perform more robust with the reduction of training data, compared with original MAML-based algorithms.
- MetaMix with MTL achieves state-of-the-art performance.