Paper Review

[ICLR 2021] Free Lunch for Few-Shot Learning: Distribution Calibration

Dept. of Computer Science & Engineering
Artificial Intelligence and Data Mining Lab(AIDML)
202122029 Meeyun Kim

2022. 11. 14.

Contents

1. Introduction

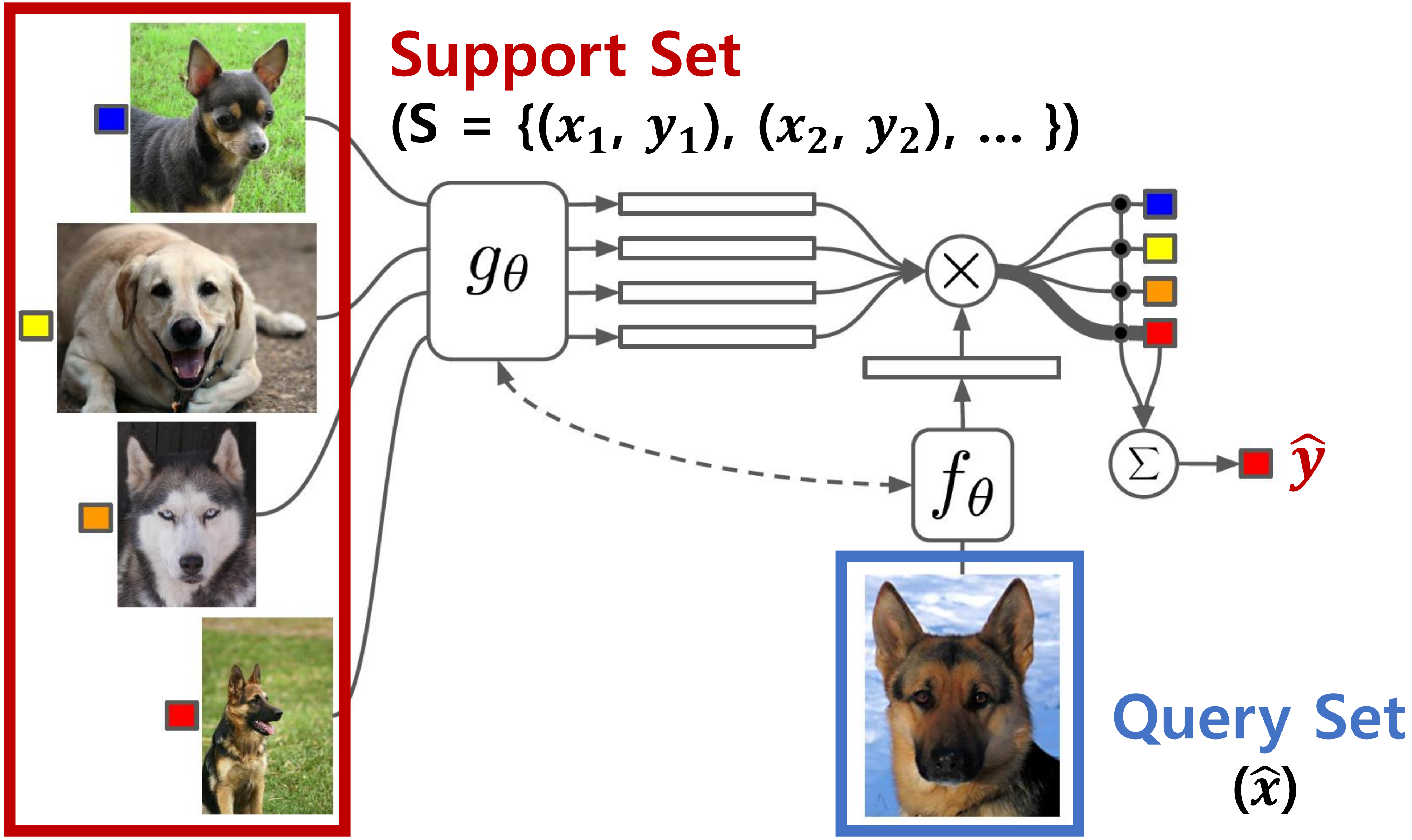
What is Meta Learning?

- ***Learning to Learn.***

- Train machine learning model well with **small amounts of data**.

- **Matching Net** (Vinyals et al. (2016)), **MAML** (Finn et al. (2017)), … etc.

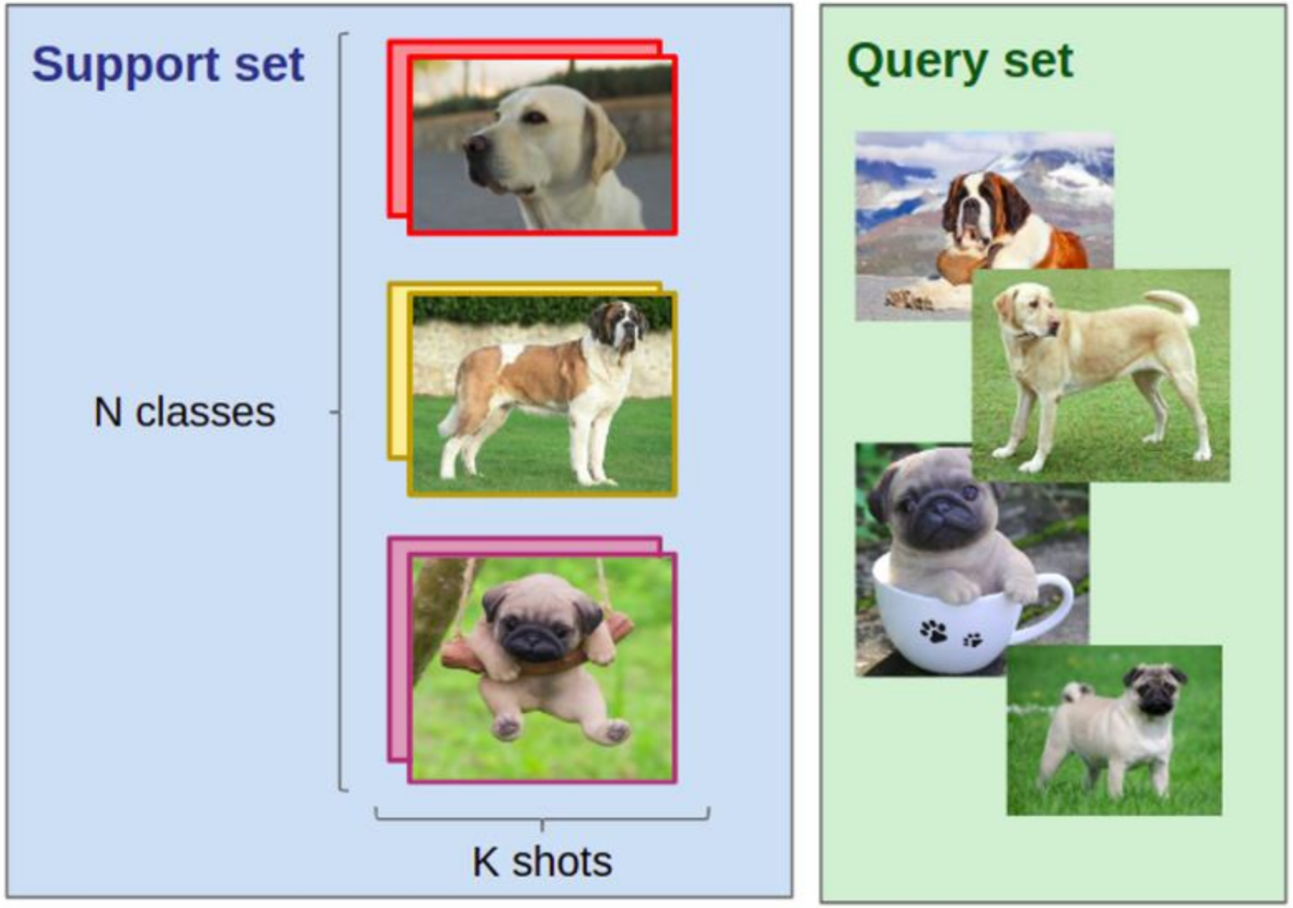
Example of Meta Learning (Matching Networks)



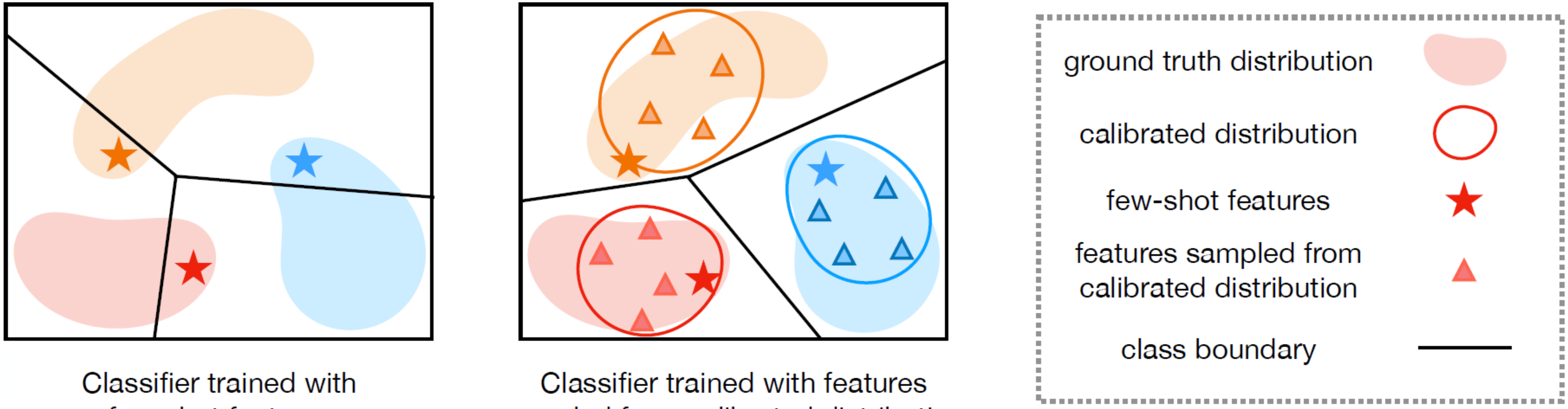
Oriol Vinyals et al., 2016. Matching networks for one shot learning. NeurIPS

Example of Meta Learning (Cont'd)

N-way K-shot task



Limitations of Few-show Learning?



Classifier trained with few-shot features

Classifier trained with features sampled from calibrated distribution

- Each features for few-shot learning is only a small fraction of the ground truth distribution.

- Thus, model tends to **overfit** on these few samples.

To Resolve Overfitting...

	Arctic fox	
	mean sim	var sim
white wolf	97%	97%
malamute	85%	78%
lion	81%	70%
meerkat	78%	70%
jellyfish	46%	26%
orange	40%	19%
beer bottle	34%	11%

- Obtain distribution from classes with sufficient data.

- Distribution is transferred to other classes based on the similarity.

→ **Distribution Calibration!**

2. Method

Problem Definition

Few Shot Learning

$$D = \{(\boldsymbol{x}_i, y_i)\}$$

$\boldsymbol{x}_i \in \mathbb{R}^d$

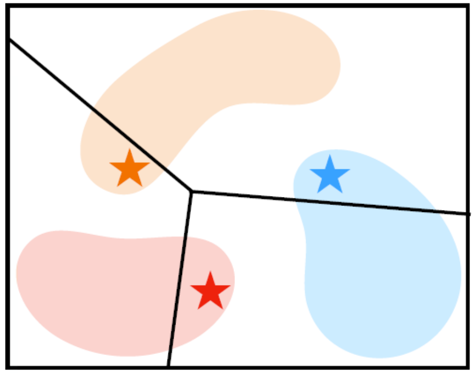
$y_i \in C, \qquad C_b \cap C_n = \emptyset, \qquad C_b \cup C_n = C$

Base classes C_b, Novel classes C_n

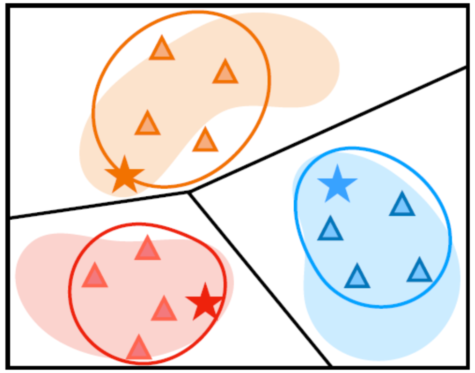
- Train a model on the data from the **base classes** so that the model can **generalize** well on tasks sampled from the **novel classes**.

Distribution Calibration

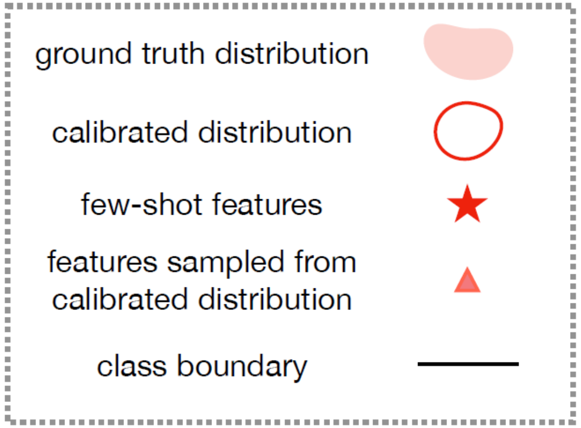
Introduction



Classifier trained with
few-shot features



Classifier trained with features
sampled from calibrated distribution



	Arctic fox	
	mean sim	var sim
white wolf	97%	97%
malamute	85%	78%
lion	81%	70%
meerkat	78%	70%
jellyfish	46%	26%
orange	40%	19%
beer bottle	34%	11%

- If the feature distribution is Gaussian, the **mean** and **variance** with respect to each class are correlated to the **semantic similarity** of each class.

- The statistics can be **transfer**red!

Distribution Calibration

Statistics of the Base Classes

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n_i} \boldsymbol{x}_j}{n_i} \qquad \boldsymbol{\Sigma}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^T$$

- The authors assume the feature distribution of base classes is **Gaussian**.

- **Mean vector** & **Covariance matrix** for the features from a base class i

Distribution Calibration (Cont'd)

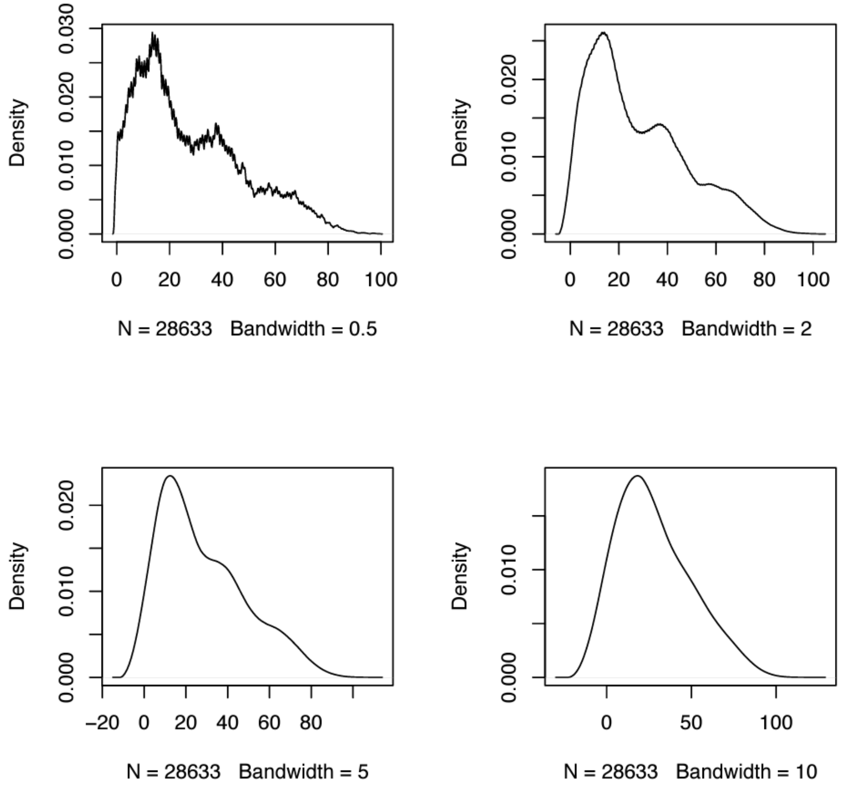
Calibrating Statistics of the Novel Classes

- $\boldsymbol{support\ set\ S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N\times K}$
- $\boldsymbol{query\ set\ Q} = \{(\boldsymbol{x}_i, y_i)\}_{i=N\times K+1}^{N\times K+N\times q}$

Distribution Calibration (Cont'd)

Calibrating Statistics of the Novel Classes (Cont'd)

$$\tilde{x} = \begin{cases} x^{\lambda}, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases}$$



- Use **Tukey's Ladder of Powers transformation**.

- It makes the feature distribution more Gaussian-like.

Distribution Calibration (Cont'd)

Calibration through Statistics Transfer

① $S_d = \{-\|\boldsymbol{\mu}_i - \tilde{x}\|^2 \mid i \in C_b\}$

② $S_N = \{i \mid -\|\boldsymbol{\mu}_i - \tilde{x}\|^2 \in topk(S_d)\}$

③ $\boldsymbol{\mu}' = \frac{\sum_{i \in S_N} \boldsymbol{\mu}_i + \tilde{x}}{k+1}$ $\quad \boldsymbol{\Sigma}' = \frac{\sum_{i \in S_N} \boldsymbol{\Sigma}_i}{k} + \alpha$

④ $S_y = \{(\boldsymbol{\mu}'_1, \boldsymbol{\Sigma}'_1), \dots, (\boldsymbol{\mu}'_K, \boldsymbol{\Sigma}'_K)\}$

① **Select the top *k* base classes with the closest distance** to the feature of sample x̃ from the support set.

② Stores the ***k* nearest base classes** with respect to feature vector x̃**.**

③ **Distribution is calibrated** by the statistics from the nearest base classes.

④ For few-shot learning, **calibration should be undertaken multiple times** with each feature vector.

Distribution Calibration (Cont'd)

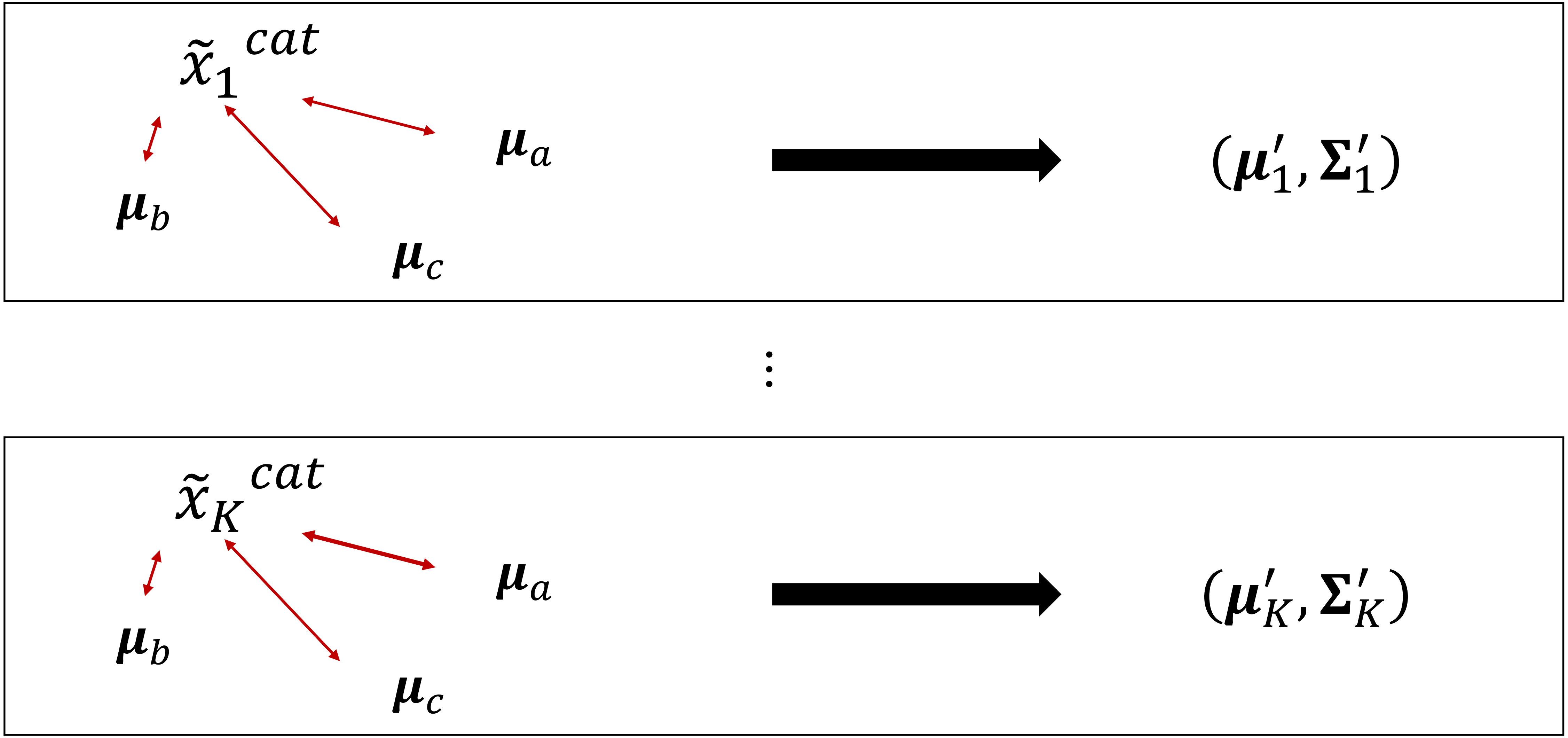
Sample Features from the Calibrated Distribution

- Generate a set of feature vectors with label y by sampling from calibrated Gaussian distributions:

$$D_y = \left\{(\boldsymbol{x}, y) | \boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \forall(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in S_y\right\}$$

- Generated per class is a hyperparameter and they are equally distributed for every calibrated distribution in S_y.

$$S_y = \{(\boldsymbol{\mu}_1', \boldsymbol{\Sigma}_1'), \ldots, (\boldsymbol{\mu}_K', \boldsymbol{\Sigma}_K')\}$$



3. Experiments & Results

Dataset for Evaluation

- ***mini*ImageNet**: **64 base** classes, **16 validation** classes, and **20 novel** classes.
  600 samples per class.

- **CUB**: 200 different classes of birds with a total of 11,788 images.
  100 base classes, **50 validation** classes, and **50 novel** classes.

- ***tiered*ImageNet**: 608 classes sampled from hierarchical category structure.
  In this paper, researchers used **351, 97,** and **160** classes for **training**, **validation**, and **test**, respectively.

Feature Extractor

- **WideResNet** trained by base classes and test performance using novel classes.

- Feature representation is extracted from penultimate layer with **ReLU**.

5way1shot and 5way5shot Classification Accuracy

Methods	*mini*ImageNet 5way1shot	*mini*ImageNet 5way5shot	*CUB* 5way1shot	*CUB* 5way5shot
Optimization-based				
MAML (Finn et al. (2017))	48.70 ± 1.84	63.10 ± 0.92	50.45 ± 0.97	59.60 ± 0.84
Meta-SGD (Li et al. (2017))	50.47 ± 1.87	64.03 ± 0.94	53.34 ± 0.97	67.59 ± 0.82
LEO (Rusu et al. (2019))	61.76 ± 0.08	77.59 ± 0.12	-	-
E3BM (Liu et al. (2020c))	63.80 ± 0.40	80.29 ± 0.25	-	-
Metric-based				
Matching Net (Vinyals et al. (2016))	43.56 ± 0.84	55.31 ± 0.73	56.53 ± 0.99	63.54 ± 0.85
Prototypical Net (Snell et al. (2017))	54.16 ± 0.82	73.68 ± 0.65	72.99 ± 0.88	86.64 ± 0.51
Baseline++ (Chen et al. (2019a))	51.87 ± 0.77	75.68 ± 0.63	67.02 ± 0.90	83.58 ± 0.54
Variational Few-shot(Zhang et al. (2019))	61.23 ± 0.26	77.69 ± 0.17	-	-
Negative-Cosine(Liu et al. (2020a))	62.33 ± 0.82	80.94 ± 0.59	72.66 ± 0.85	89.40 ± 0.43
Generation-based				
MetaGAN (Zhang et al. (2018))	52.71 ± 0.64	68.63 ± 0.67	-	-
Delta-Encoder (Schwartz et al. (2018))	59.9	69.7	69.8	82.6
TriNet (Chen et al. (2019b))	58.12 ± 1.37	76.92 ± 0.69	69.61 ± 0.46	84.10 ± 0.35
Meta Variance Transfer (Park et al. (2020))	-	67.67 ± 0.70	-	80.33 ± 0.61
Maximum Likelihood with DC (Ours)	66.91 ± 0.17	80.74 ± 0.48	77.22 ± 0.14	89.58 ± 0.27
SVM with DC (Ours)	$\mathbf{67.31 \pm 0.83}$	$\mathbf{82.30 \pm 0.34}$	$\mathbf{79.49 \pm 0.33}$	$\mathbf{90.26 \pm 0.98}$
Logistic Regression with DC (Ours)	$\mathbf{68.57 \pm 0.55}$	$\mathbf{82.88 \pm 0.42}$	$\mathbf{79.56 \pm 0.87}$	$\mathbf{90.67 \pm 0.35}$

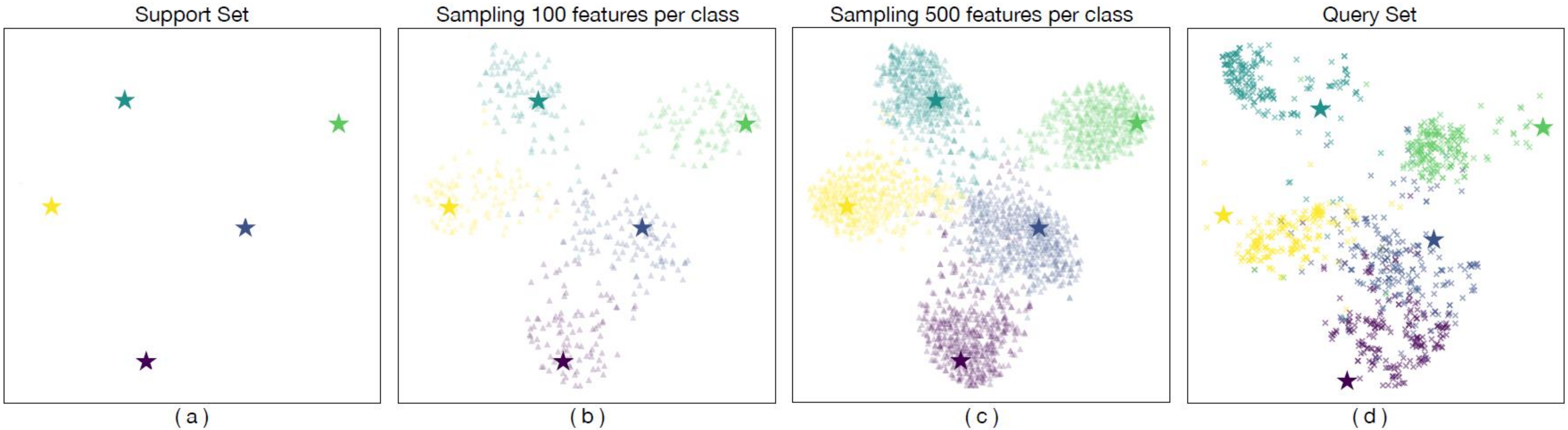
SOTA!

5way5shot Classification Accuracy

Methods	*tiered*ImageNet 5way1shot	*tiered*ImageNet 5way5shot
Matching Net (Vinyals et al. (2016))	68.50 ± 0.92	80.60 ± 0.71
Prototypical Net (Snell et al. (2017))	65.65 ± 0.92	83.40 ± 0.65
LEO (Rusu et al. (2019))	66.33 ± 0.05	82.06 ± 0.08
E3BM (Liu et al. (2020c))	71.20 ± 0.40	85.30 ± 0.30
DeepEMD (Zhang et al., 2020)	71.16 ± 0.87	86.03 ± 0.58
Maximum Likelihood with DC (Ours)	75.92 ± 0.60	87.84 ± 0.65
SVM with DC (Ours)	$\mathbf{77.93 \pm 0.12}$	$\mathbf{89.72 \pm 0.37}$
Logistic Regression with DC (Ours)	$\mathbf{78.19 \pm 0.25}$	$\mathbf{89.90 \pm 0.41}$

SOTA!

Visualization of Generated Samples

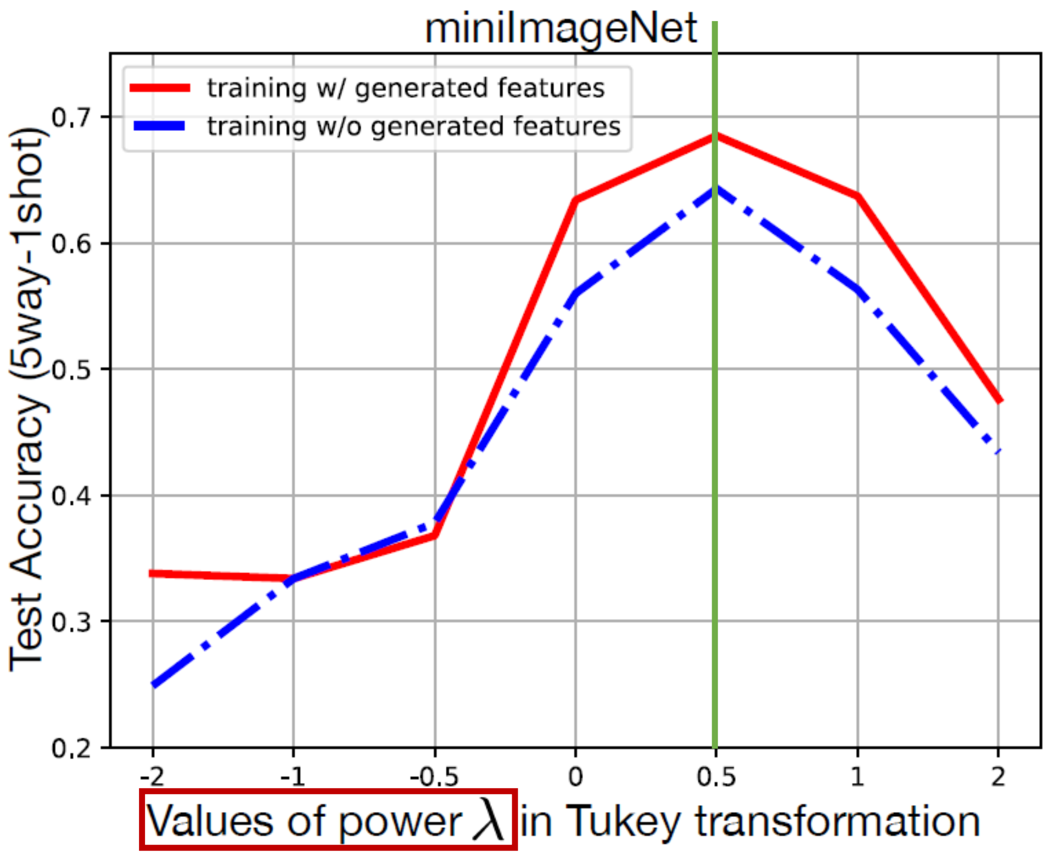


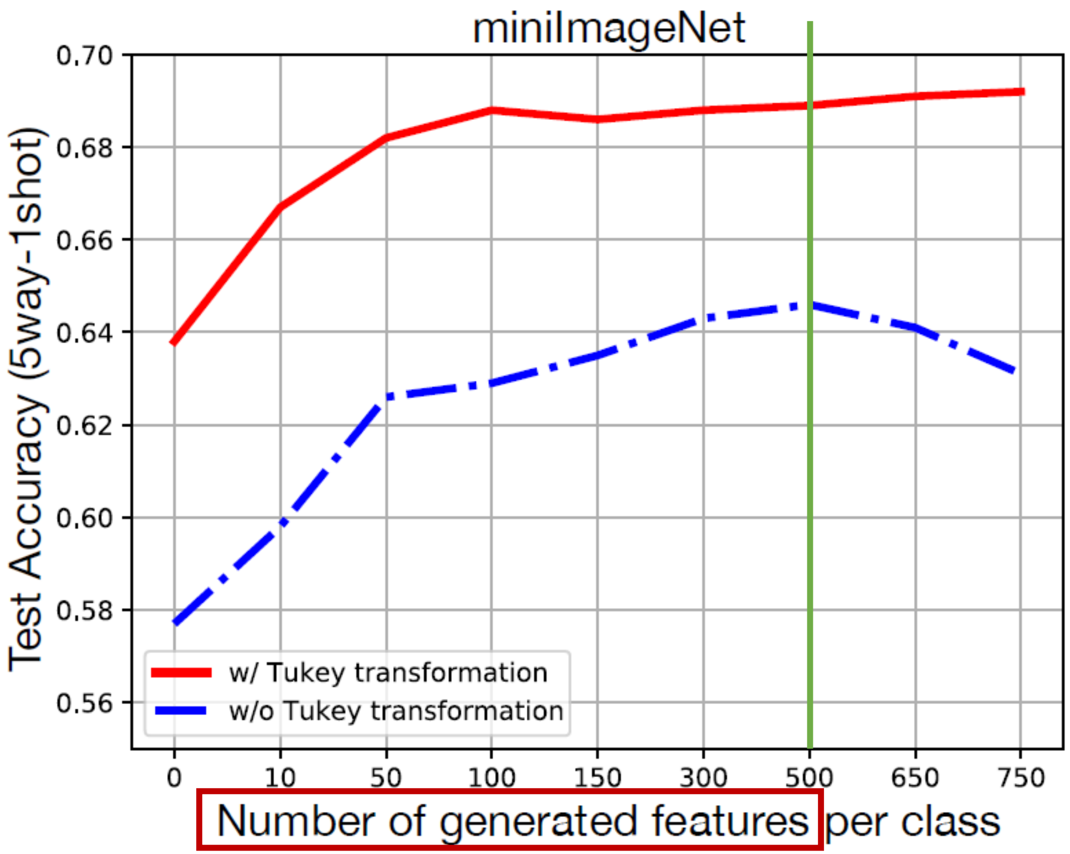
'★': support set features, 'x' in figure (d): query set features, '▲' in figure (b)(c): generated features.

- **t-SNE analysis** is conducted.

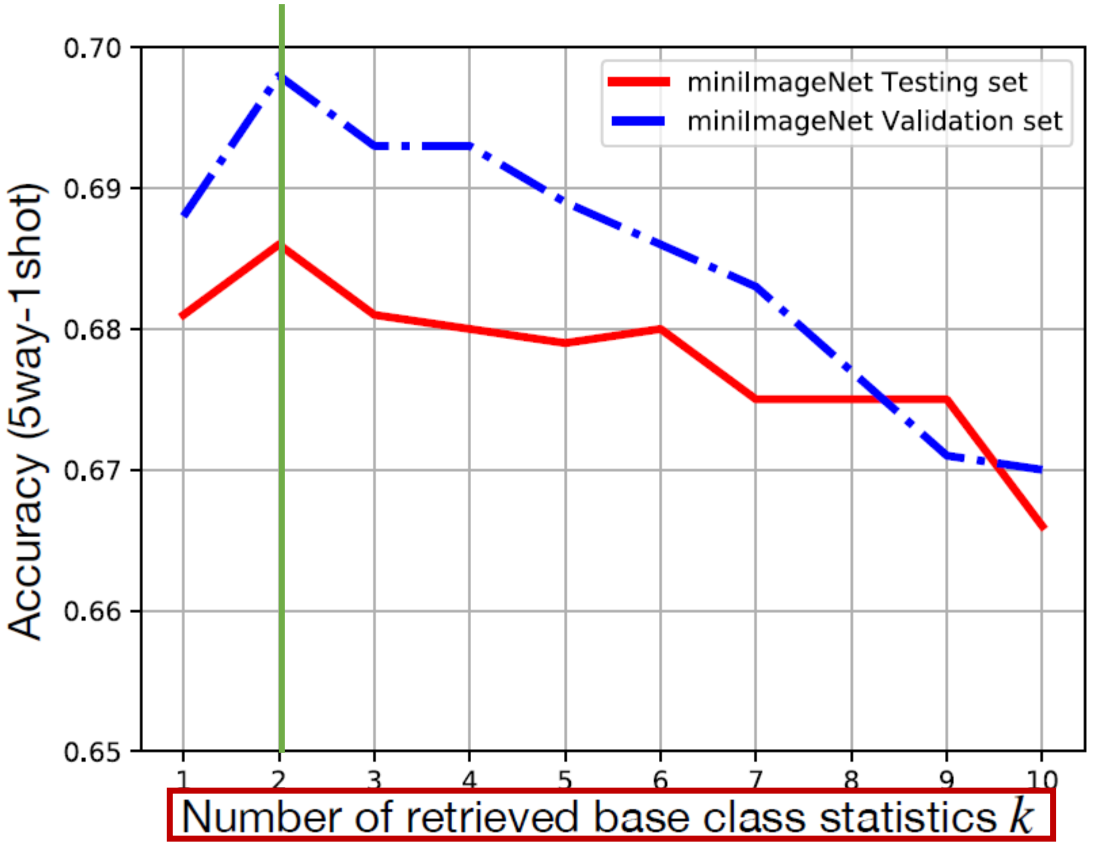
- Training with these generated features can alleviate the mismatch.

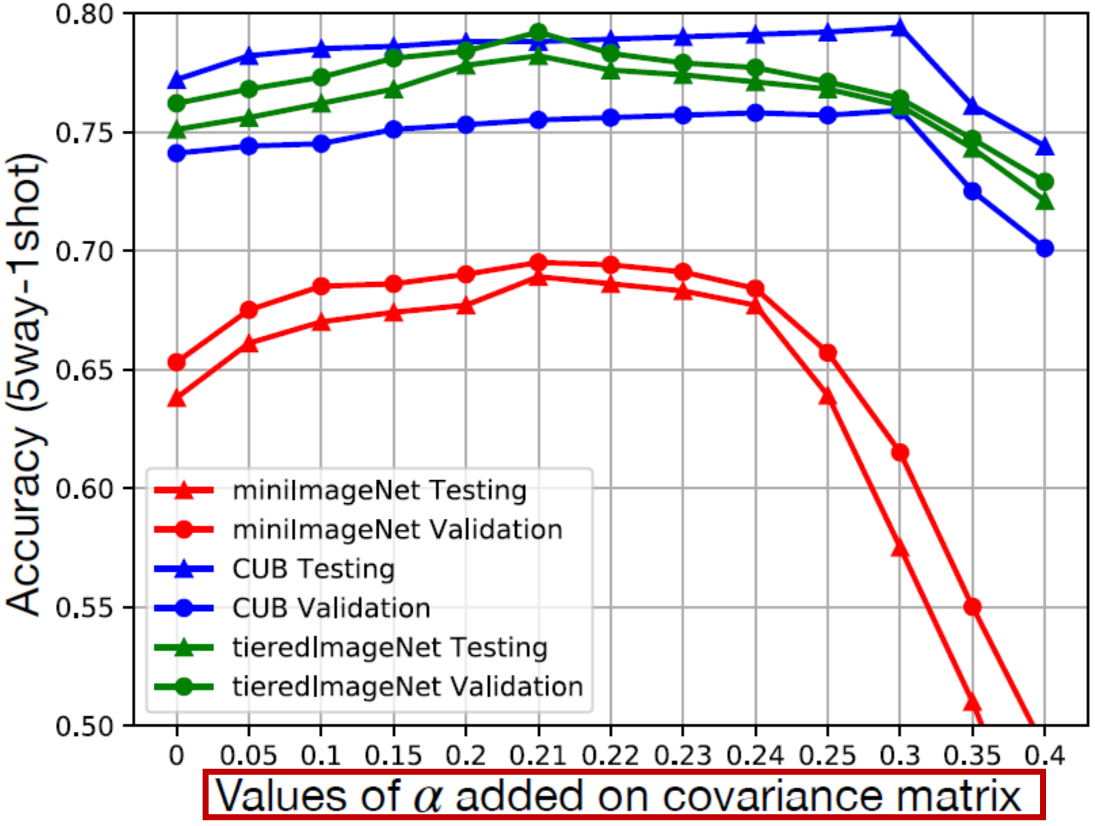
Effects of Hyperparameters





Effects of Hyperparameters (Cont'd)





4. Conclusion

- Researchers proposed simple and effective **distribution calibration strategy**.
- Achieve **better performance** than other meta learning models.
- Distribution calibration in a variety of problem environments will be studied.

References

- O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., “Matching networks for one shot learning,” Advances in neural information processing systems, vol. 29, 2016.

- https://en.wikipedia.org/wiki/Arctic_fox

- https://en.wikipedia.org/wiki/Arctic_wolf

- https://en.wikipedia.org/wiki/Jellyfish

- https://en.wikipedia.org/wiki/Beer_bottle

- https://blog.si-analytics.ai/3

- J. W. Tukey et al., Exploratory data analysis, vol. 2. Reading, MA, 1977.