Worst Case Matters for Few-Shot Recognition

Minghao Fu, Yun-Hao Cao, Jianxin Wu*
State Key Laboratory for Novel Software Technology, Nanjing University
fumh@lamda.nju.edu.cn, caoyh@lamda.nju.edu.cn, wujx2001@gmail.com

TEL AVIV 2022

1. Introduction & Motivation

This paper focuses on improving the worst-case accuracy of few-shot learning. To accomplish this goal, we propose two strategies, i.e., stability regularization (SR) and adaptability calibration (AC) from the perspective of bias-variance tradeoff.

Key idea

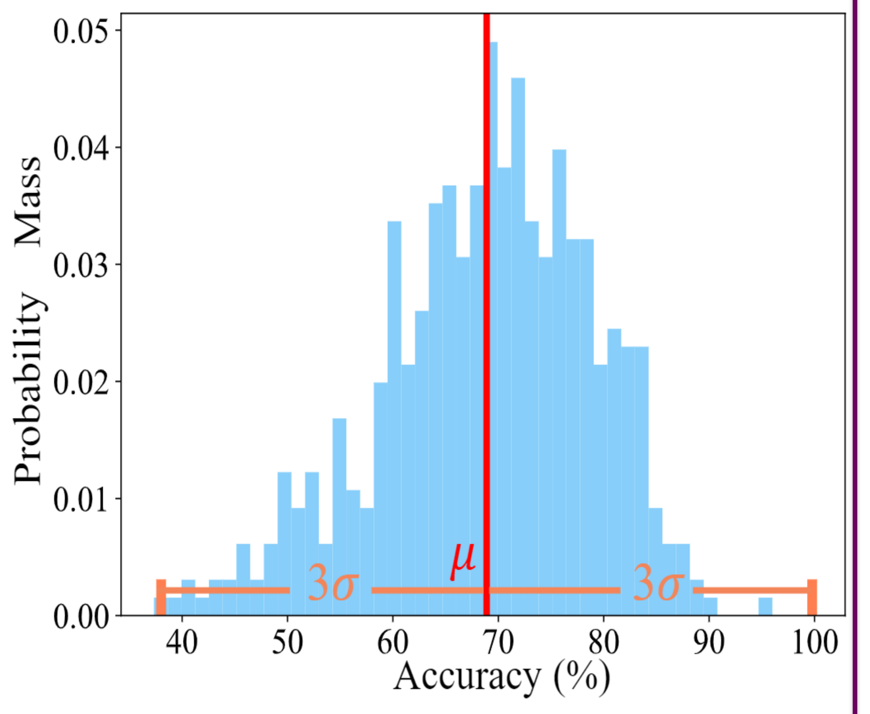
➢ We aim to boost the worst-case accuracy of few-shot learning.

Why worst-case accuracy?

- Current criterion: Average accuracy and 95% confidence interval over many episodes.
- A high average accuracy does not necessarily mean a high worst-case accuracy.

Method	ACC_m	$Z_{95\%}$	σ	ACC_1	ACC_{10}
Negative-Cosine [21]	61.72	0.81	10.12	24.27	36.13
MixtFSL [2]	64.31	0.79	9.87	30.67	35.07
S2M2$_R$ [23]	64.93	**0.18**	**9.18**	37.58	42.87
PT+NCM [16]	65.35	0.20	10.20	32.00	38.13
CGCS [11]	67.02	0.20	10.20	**38.70**	**44.00**
LR-DC [36]	**68.57**	0.55	10.28$^\flat$	37.33	42.72

- Few-shot learning is very unstable. The worst-case lags far behind the average.
- In real-world applications, worst-case accuracy is very important.



2. Surrogate of worst-case accuracy & Framework

- The accuracy distribution often fits well to a Gaussian. The worst-case accuracy is naturally estimated by the 3σ rule as μ − 3σ. Therefore we propose *maximizing mean accuracy μ* and *minimizing standard deviation σ* **simultaneously**.
- Since directly optimizing σ is not plausible, according to bias-variance decomposition theory, we propose **stability regularization (SR)** with model ensemble to reduce variance, and **adaptability calibration (AC)** to reduce bias.

✓ Stability Regularization:

- A loss function to prevent model from completely forgetting the representation learned from the base set.

$$\mathcal{L}_S(\mathbf{x}) = -\frac{f(\mathbf{x}) \cdot \hat{f}(\mathbf{x})}{\|f(\mathbf{x})\|\|\hat{f}(\mathbf{x})\|}$$

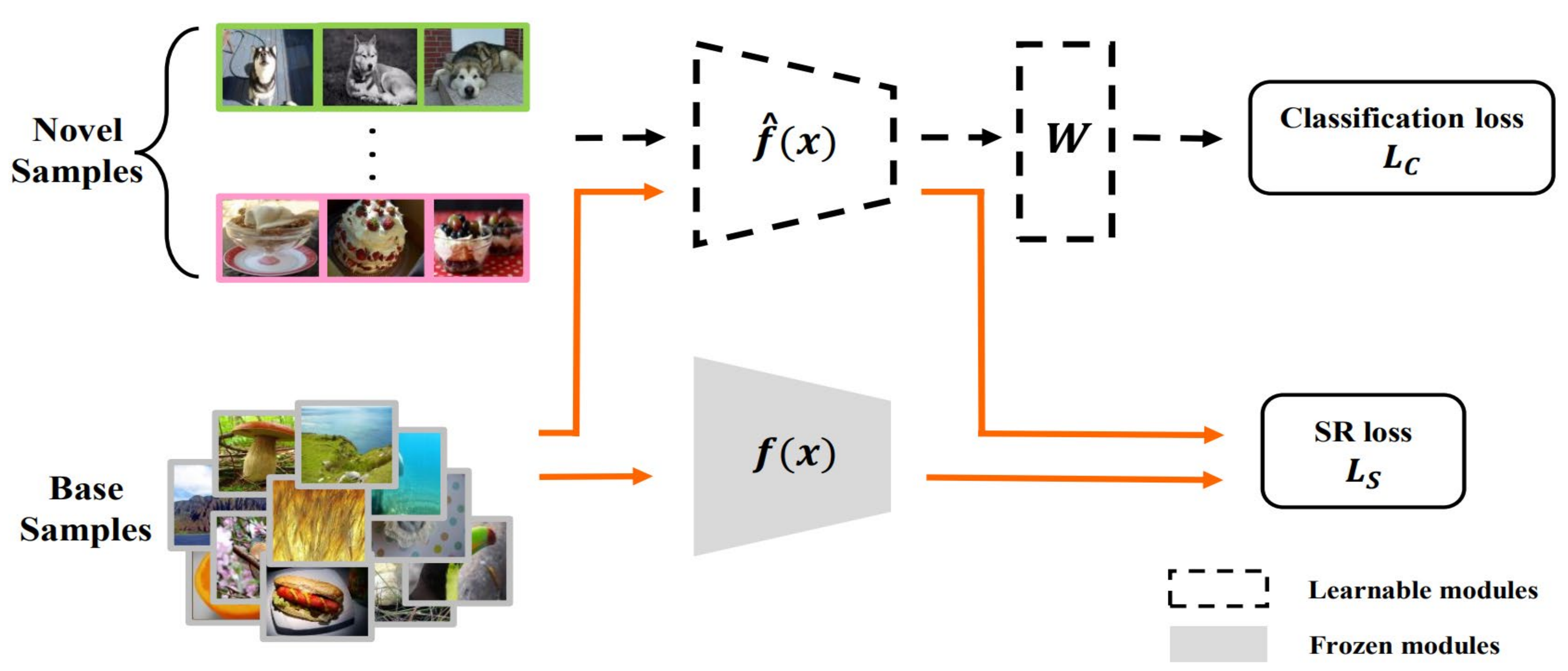
- Generalizable to other unlabeled images.
- Ensemble.

✓ Adaptability Calibration:

- A remedy to properly calibrated capacity increase.

Bias-variance Tradeoff of SR and AC

- SR: More stable classifier has a smaller variance (σ).
- AC: Larger model capacity has a smaller bias. Because of the contradiction of two terms, a balance is necessary.



3. Experiments

State-of-the-art comparison

Dataset	Method	1-shot ACC_1	1-shot $\mu-3\sigma$	1-shot σ	1-shot ACC_m	5-shot ACC_1	5-shot $\mu-3\sigma$	5-shot σ	5-shot ACC_m
mini-ImageNet	ProtoNet† [29]	19.76	23.41	10.25	54.16	43.74	49.32	8.12	73.68
	Transductive-FT [7]	24.00	32.82	10.97	65.73	50.67	53.23	8.39	78.40
	Negative-Cosine [21]	24.27	31.36	10.12	61.72	53.30	61.18	6.87	81.79
	MixtFSL [2]	30.67	34.70	9.87	64.31	46.67	59.16	7.50	81.66
	PT+NCM [16]	32.00	34.75	10.20	65.35	56.00	63.98	6.63	83.87
	LR-DC [36]	37.33	37.73	10.28	68.57	60.52	63.02	6.62	82.88
	S2M2$_R$ [23]	37.58	37.39	**9.18**	64.93	58.66	66.35	**5.61**	83.18
	CGCS [11]	38.70	36.42	10.20	67.02	49.30	60.90	7.14	82.32
	AC+SR (ours)	40.52	40.25	9.71	69.38	63.20	66.46	6.47	85.87
	AC+EnSR (ours)	**40.52**	**40.67**	9.64	**69.59**	**63.48**	**66.71**	6.42	**85.97**
CUB	ProtoNet† [29]	28.00	39.99	11.00	72.99	53.33	67.53	6.37	86.64
	Negative-Cosine [21]	36.00	40.80	10.62	72.66	70.70	73.29	5.37	89.40
	PT+NCM [16]	40.00	49.97	10.20	80.57	69.33	75.85	5.10	91.15
	MixtFSL [2]	40.00	32.69	13.75	73.94	57.33	67.26	6.25	86.01
	LR-DC [36]	44.00	49.74	9.94	79.56	68.80	75.49	5.06	90.67
	CGCS [11]	50.67	42.53	10.71	74.66	57.33	70.01	6.12	88.37
	S2M2$_R$ [23]	52.00	50.32	10.12	80.68	73.86	74.35	5.50	90.85
	AC+SR (ours)	52.78	58.44	8.90	85.14	76.00	81.85	4.19	94.42
	AC+EnSR (ours)	**53.04**	**58.84**	**8.86**	**85.42**	**77.58**	**82.14**	**4.13**	**94.53**

- Our method performs best on worst-case and average accuracy.
- μ − 3σ correlates better with ACC1 than σ or μ.

Results for different degree of AC

W	AC 5	AC 4	AC 3	AC 2+1	1-shot ACC_m	1-shot σ	1-shot ACC_1	1-shot ACC_{10}	1-shot ACC_{100}	5-shot ACC_m	5-shot σ	5-shot ACC_1	5-shot ACC_{10}	5-shot ACC_{100}
✓					59.53	**9.99**	30.42	35.09	45.05	73.61	8.16	44.54	50.16	61.36
✓	✓				62.76	10.16	**32.78**	**37.55**	**47.84**	79.23	7.62	46.14	56.96	67.91
✓	✓	✓			62.77	10.38	30.12	35.44	47.46	79.91	**7.54**	45.60	**57.84**	68.77
✓	✓	✓	✓		**62.80**	10.50	22.96	34.93	47.27	**80.04**	7.57	45.58	57.74	**68.85**
✓	✓	✓	✓	✓	62.19	10.50	25.60	35.29	46.68	79.87	7.58	**48.00**	57.41	68.67

- Fine-tuning W+ ‘res5’ is the AC strategy used in our method.

Generalize SR to other data

D_{sr}	ACC_m	σ	ACC_1	ACC_{10}	ACC_{100}
-	60.09	10.13	25.60	34.55	45.24
mini-ImageNet	**62.33**	10.27	32.00	**37.03**	**47.36**
CUB [34]	60.66	9.99	30.14	36.43	46.29
Cars [18]	61.05	10.29	30.94	36.64	46.20
DTD [6]	61.99	10.23	31.20	36.75	47.23
Pets [25]	61.57	**9.95**	32.00	36.62	47.27
Flower [24]	61.03	10.07	**32.26**	36.51	46.37
CIFAR-100 [19]	60.47	10.04	30.68	36.69	46.02

- Don't require images used in SR to be visually similar or semantically correlated to support samples.
- SR is consistently useful.

4. Contributions & Conclusions

✓ We are the first to emphasize the importance and to advocate the adoption of **worst case accuracy** in few-shot learning.

✓ We argue that in addition to maximizing the average accuracy μ, we must also **simultaneously** reduce the standard deviation σ.

✓ We propose stability regularization (SR) with model ensemble and adaptability calibration (AC) strategies from the perspective of bias-variance tradeoff.

✓ Our method shows superior results compared to current state-of-the-art methods in terms of not only average, but also worst-case accuracy.