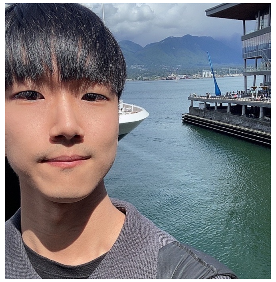
1 CARNEGIE MELLON UNIVERSITY
2 UNIVERSITY OF SEOUL
3 CHUNG-ANG UNIVERSITY
4 KAIST
5 SUNG KYUN KWAN UNIVERSITY (SKKU)
6 YONSEI UNIVERSITY

Towards Calibrated Robust Fine-Tuning of Vision-Language Models

Changdae Oh[1,2] Mijoo Kim[1,3] Hyesu Lim[1,4] Junhyeok Park[1,4] Euiseog Jeong[1,5]
Zhi-Qi Cheng[1] Kyungwoo Song[6]







NEURAL INFORMATION PROCESSING SYSTEMS

Workshop on Distribution Shifts: New Frontiers with Foundation Models
Fri 15 Dec, 10 a.m. EST (Room R06-R09)

Robust fine-tuning

- Adapting large-scale pre-trained models under distribution shifts
- Goal: good out-of-distribution **(OOD) generalization** as well as in-distribution **(ID) generalization** after fine-tuning

Standard fine-tuning

train & eval

ID Accuracy

eval

OOD Accuracy

Increasing ID adaptation trades off OOD generalization capability

Fig source: Finetune like you pretrain: Improved finetuning of zero-shot vision models, Goyal et al. 2023

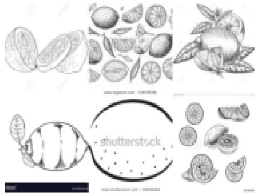
Robust fine-tuning

- Adapting large-scale pre-trained models under distribution shifts
- Goal: good out-of-distribution **(OOD) generalization** as well as in-distribution **(ID) generalization** after fine-tuning

Standard fine-tuning

train & eval

eval



ID Accuracy

OOD Accuracy

Increasing ID adaptation trades off OOD generalization capability

Robust fine-tuning

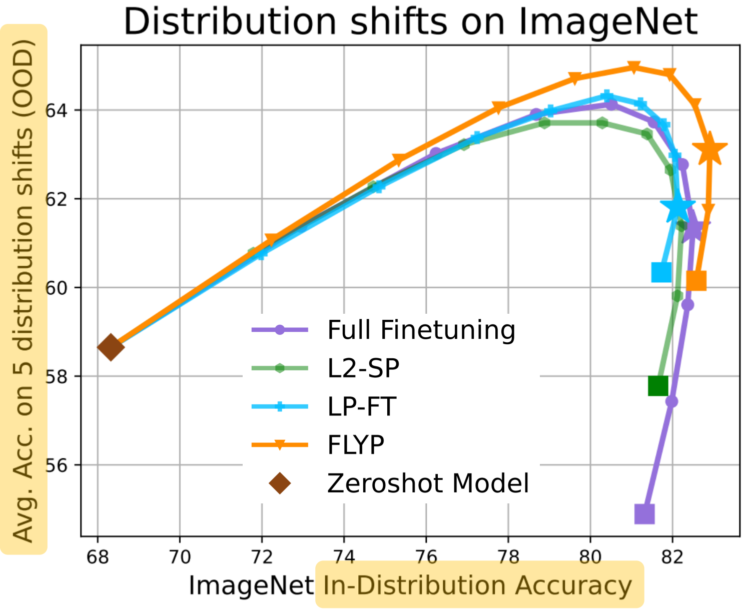
train & eval

eval

ID Accuracy

OOD Accuracy

Adapting on ID data while securing OOD generalization capability



Focus on achieving better trade-off between ID and OOD generalization

Fig source: Finetune like you pretrain: Improved finetuning of zero-shot vision models, Goyal et al. 2023

Research motivation

- There is another crucial aspect of model evaluation: ***confidence calibration***

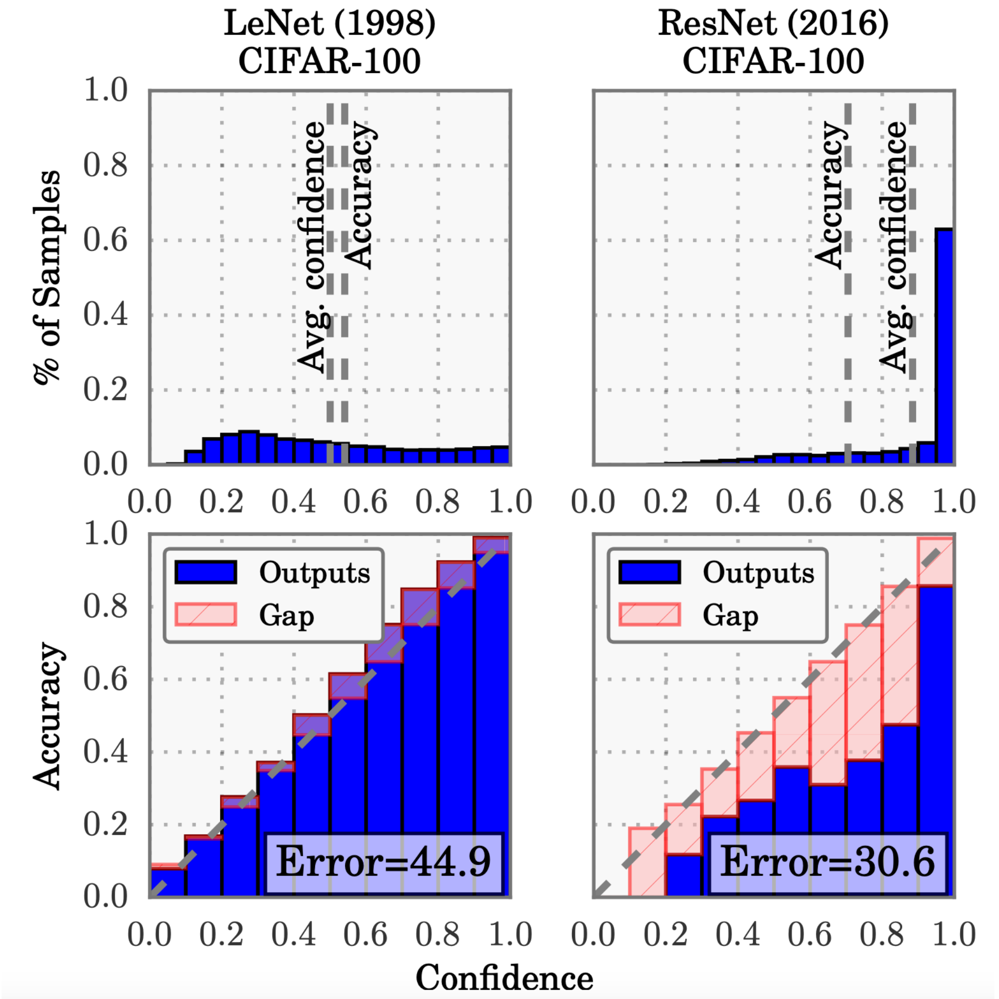
How well does the confidence output by our model match the accuracy?

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|$$

expected calibration error

Research motivation

- There is another crucial aspect of model evaluation: ***confidence calibration***

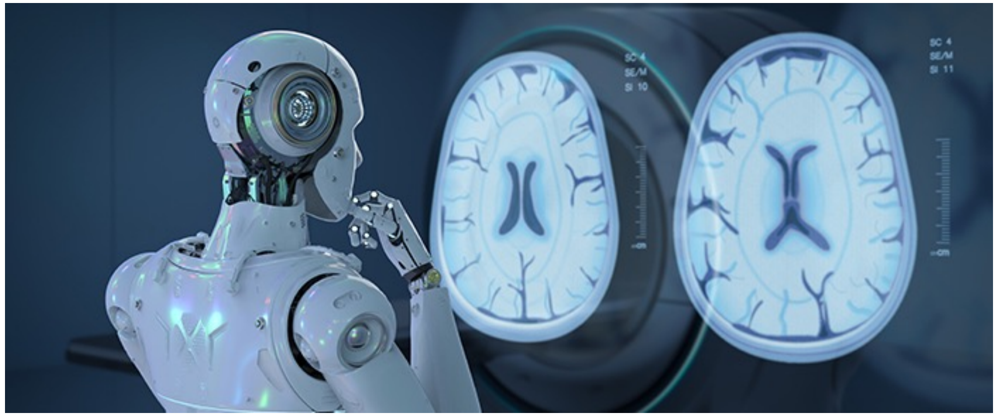


There have been many arguments that **modern neural networks exhibit poor calibration**!

Fig source: On Calibration of Modern Neural Networks, Guo et al. 2017

Research motivation

- There is another crucial aspect of model evaluation: ***confidence calibration***



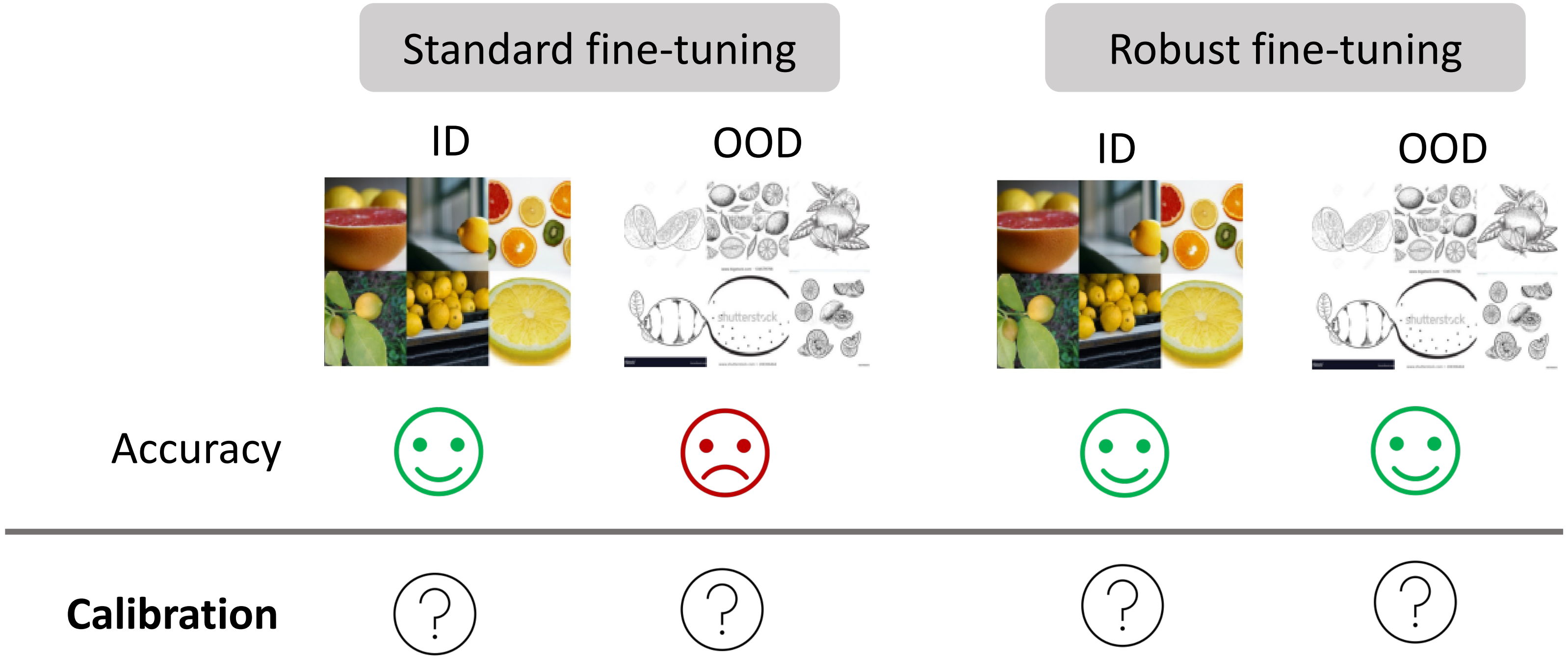
There have been many arguments that **modern neural networks exhibit poor calibration**!

These raise concerns about developing AI-driven decision-making systems on high-stakes tasks

Fig source: [1] Autonomous Vehicles: Legal and Regulatory Developments in the United States, Jones Day 2021
[2] The Top Applications for AI in Medical Diagnostics, George 2021

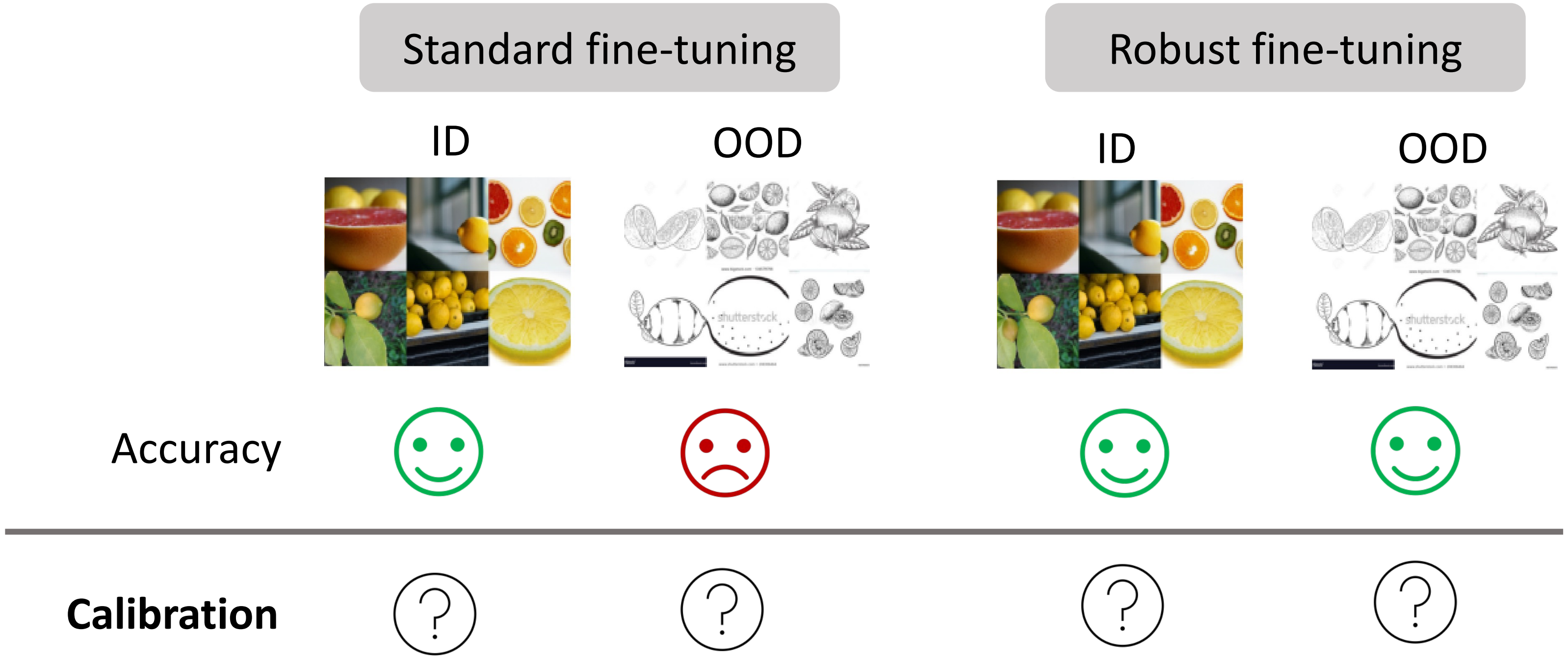
Research motivation

- Existing works on fine-tuning have overlooked confidence calibration!



Research motivation

- Existing works on fine-tuning have overlooked confidence calibration!



We initiate the discussion on calibration of fine-tuned foundation models under distribution shifts!

RQ1) *How the calibration of a pre-trained model will be affected by fine-tuning it on a specific dataset?*

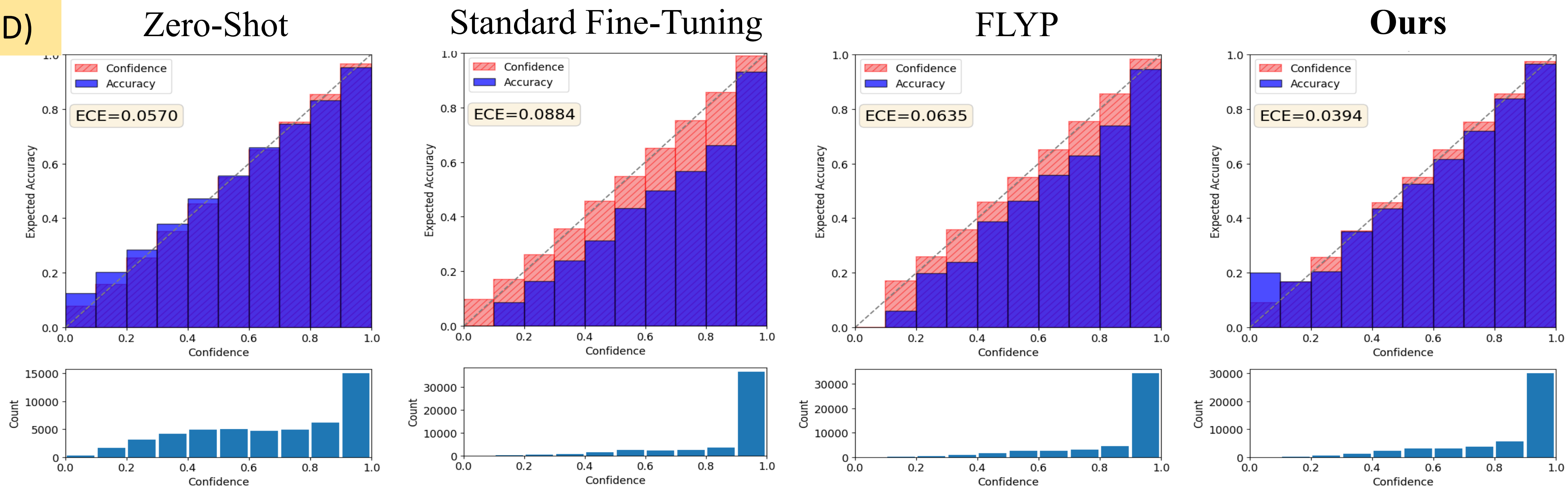
RQ2) *Will robust fine-tunings ensure calibration of the model as well as generalization both on ID and OOD?*

Research motivation

- Our findings
  - **Standard fine-tuning hurts** the calibration of zero-shot vision models in terms of ID and especially OOD expected calibration error (ECE).
  - While **SOTA robust fine-tuning method FLYP** [1] maintains ID calibration somewhat, it also **degenerates** OOD calibration.

ImageNet-1K (ID)

CLIP ViT-B/16
OpenAI ckpt



[1] Finetune like you pretrain: Improved finetuning of zero-shot vision models, Goyal et al. 2023

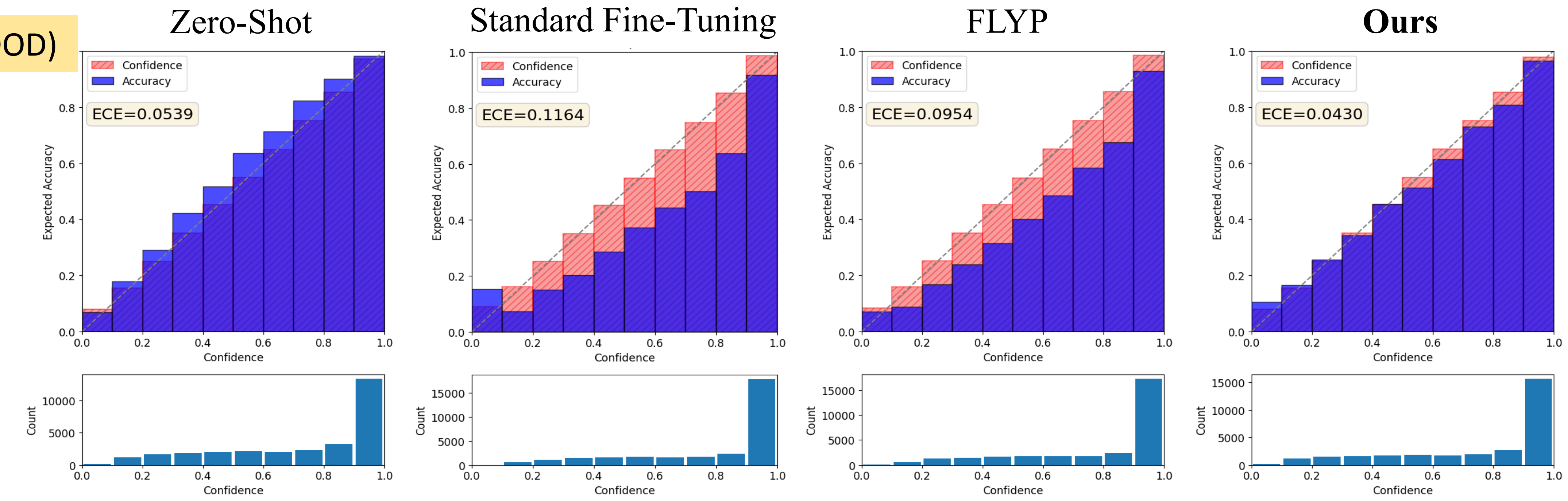
Research motivation

○ Our findings

- **Standard fine-tuning hurts** the calibration of zero-shot vision models in terms of ID and especially OOD expected calibration error (ECE).
- While **SOTA robust fine-tuning method FLYP** [1] maintains ID calibration somewhat, it also **degenerates** OOD calibration.

ImageNet-R (OOD)

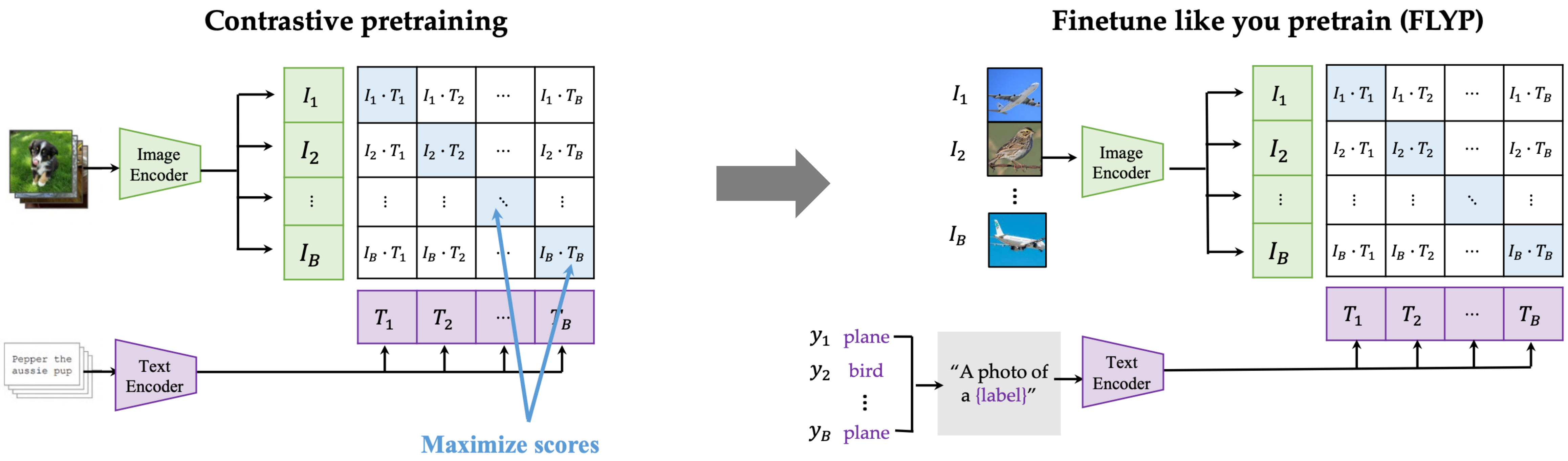
CLIP ViT-B/16
OpenAI ckpt



[1] Finetune like you pretrain: Improved finetuning of zero-shot vision models, Goyal et al. 2023

Method: Calibrated Robust Fine-tuning (CaRot)

- Following FLYP [1], we adopt a contrastive loss as our basic learning objective
  - Goyal et al. empirically showed that fine-tuning vision-language models (VLM) with contrastive loss brings huge benefits in terms of ID adaptation and OOD generalization.



[1] Finetune like you pretrain: Improved finetuning of zero-shot vision models, Goyal et al. 2023

Method: Calibrated Robust Fine-tuning (CaRot)

- Taking inspiration from a finding that label smoothing [2] helps calibration as well as generalization [3], we first try **equipping label smoothing with contrastive loss** ($\mathcal{L}_{\mathrm{MCL}\ w/\ \mathrm{LS}}$ in Figure).
- We further propose a multimodal (self-)knowledge distillation loss ($\mathcal{L}_{\mathrm{MKD}}$) which can be regarded as a form of data-dependent label smoothing [4].

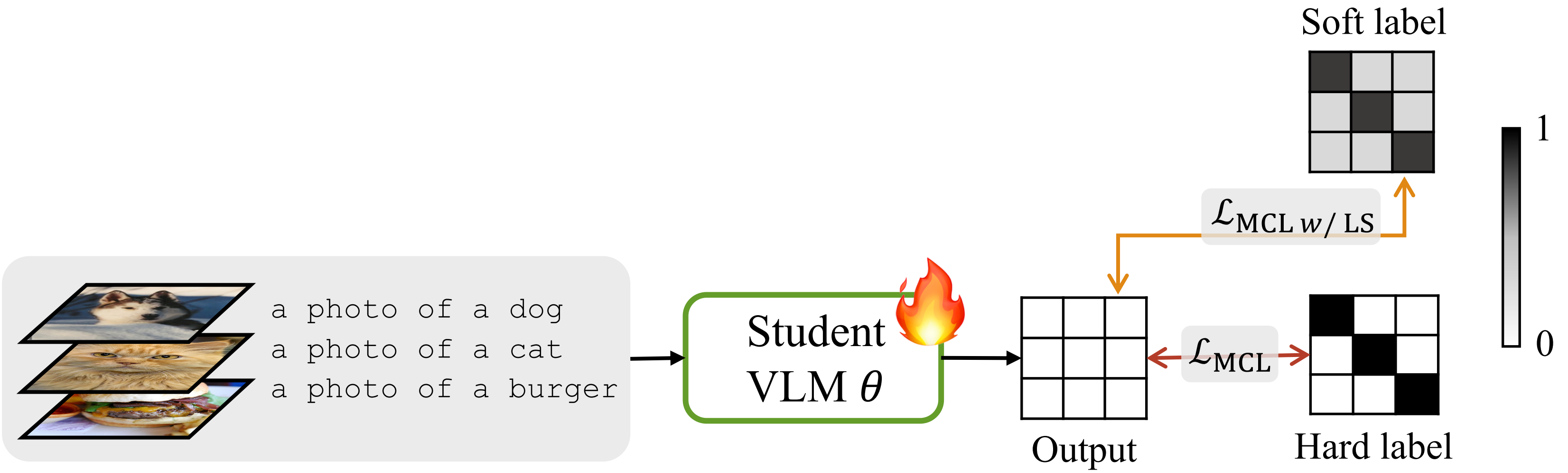


figure created by Hyesu Lim

[2] Rethinking the inception architecture for computer vision, Szegedy et al. 2016
[3] When Does Label Smoothing Help?, Muller et al. 2019
[4] Revisiting knowledge distillation via label smoothing regularization, Yuan et al. 2020

Method: Calibrated Robust Fine-tuning (CaRot)

- Taking inspiration from a finding that label smoothing [2] helps calibration as well as generalization [3], we first try equipping label smoothing with contrastive loss ($\mathcal{L}_{\mathrm{MCL}\ w/\ \mathrm{LS}}$ in Figure).
- We further propose a **multimodal (self-)knowledge distillation** loss ($\mathcal{L}_{\mathrm{MKD}}$) which can be worked as a form of data-dependent label smoothing [4].

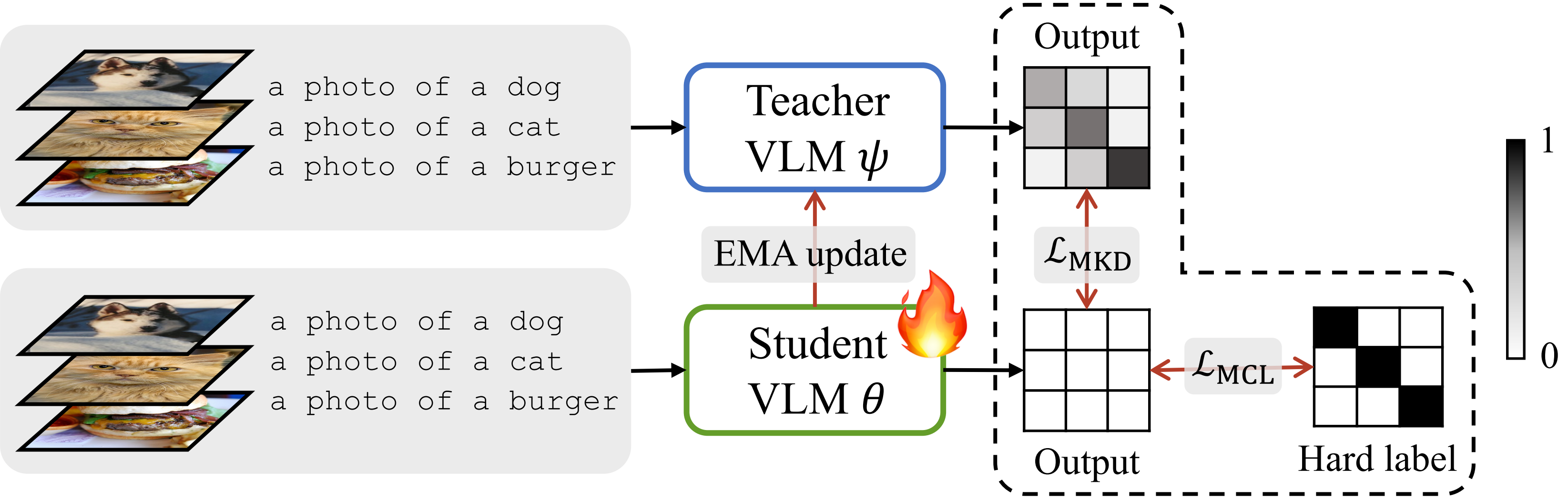


figure created by Hyesu Lim

[2] Rethinking the inception architecture for computer vision, Szegedy et al. 2016
[3] When Does Label Smoothing Help?, Muller et al. 2019
[4] Revisiting knowledge distillation via label smoothing regularization, Yuan et al. 2020

Method: Calibrated Robust Fine-tuning (CaRot)

- Understanding **multimodal knowledge distillation** loss
  1. Exponential moving average (EMA) of VLM's learning weights $\psi \leftarrow \alpha\psi + (1-\alpha)\theta$
     - it gradually blends a multi-domain calibrated one (pre-trained VLM) with an ID calibrated one (fine-tuned VLM)

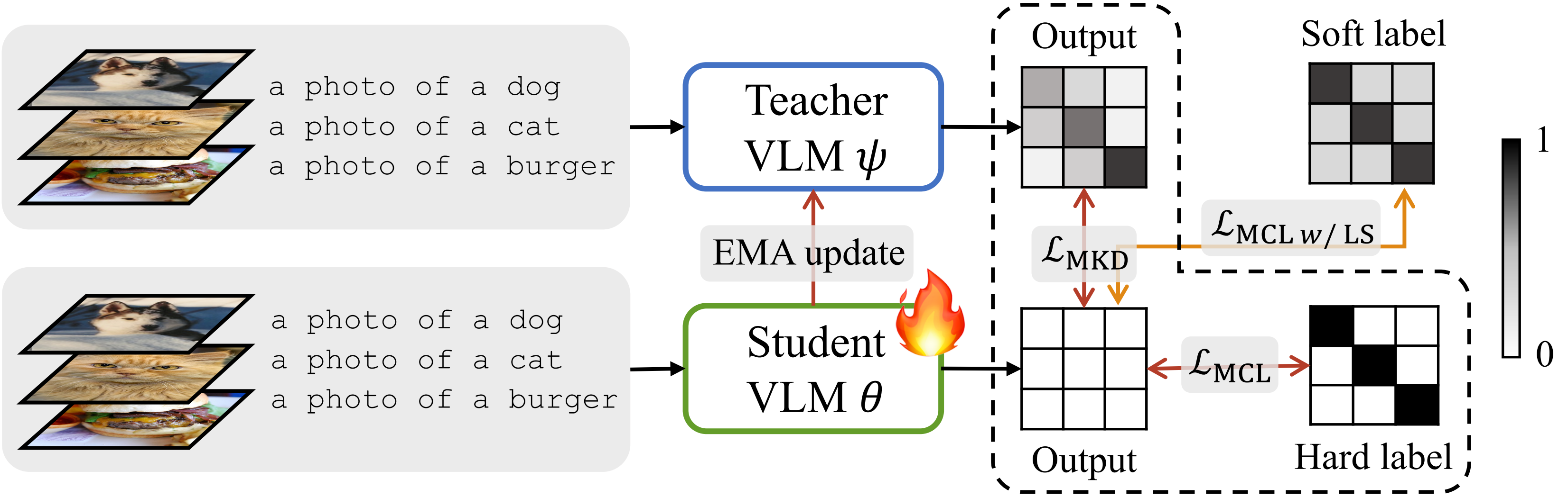


figure created by Hyesu Lim

[5] Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift, Kumar et al. 2022
[6] Robust fine-tuning of zero-shot models, Wortsman et al. 2022

Method: Calibrated Robust Fine-tuning (CaRot)

- Understanding **multimodal knowledge distillation** loss
  2. Output similarity map of the EMA teacher model
     - Holds rich multimodal relation structure for each instance
     - Produce data-dependent soft label that supports and regularizes the learning of student model

$$\mathcal{L}_{\text{MKD}}(\mathcal{B},\theta) := \sum_{i=1}^{B}[KL(\tilde{q}_i^I||q_i^I))) + KL(\tilde{q}_i^T||q_i^T)))]$$

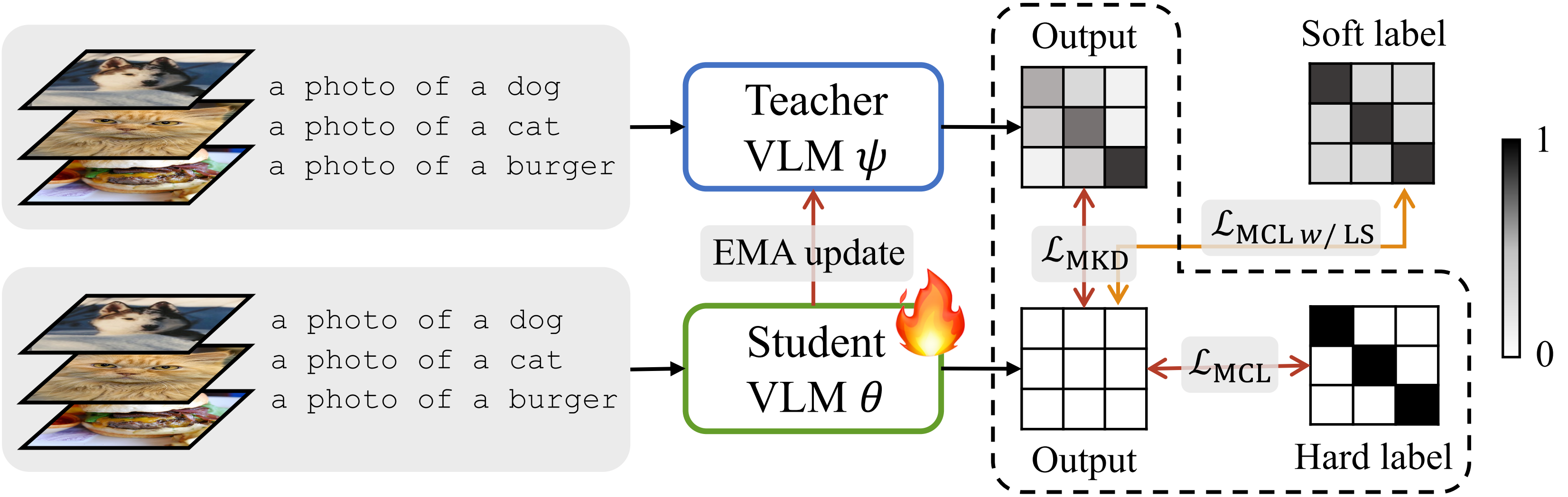


figure created by Hyesu Lim

Results

- Findings
  1. During adaptation on the ID dataset, **FT** sacrifices the OOD generalization capability of pre-trained model (zero-shot CLIP) as well as ID/OOD calibration
  2. While **WiSE-FT** [6] showcases strong OOD Acc., it significantly degenerates the calibration of the pre-trained model on ID and OOD datasets

Method	ID Acc. (↑)	OOD Acc. (↑)	w/o *TS* ID ECE (↓)	w/o *TS* OOD ECE (↓)	w/ *TS* ID ECE (↓)	w/ *TS* OOD ECE (↓)
ZS	0.6832	0.5840	0.0571	0.0836	0.0561	0.0748
FT	0.8153	0.5750	0.0884	0.2186	0.0629	0.1629
FT w/ LS	0.8223	0.5833	0.0460	0.1147	0.0481	0.1282
WiSE-FT	0.8043	0.6350	0.2129	0.1764	0.0872	0.1533
WiSE-FT w/ LS	0.8068	**0.6405**	0.5231	0.3601	0.3382	0.2425
FLYP	0.8258	0.5946	0.0643	0.1831	0.0392	0.1217
FLYP w/ LS	0.8271	0.5975	0.0459	0.1295	0.0427	0.1145
CaRot	**0.8319**	0.6197	**0.0395**	**0.1093**	**0.0380**	**0.0980**

[6] Robust fine-tuning of zero-shot models, Wortsman et al. 2022

Results

- Findings
  3. **FLYP** [1] achieves strong generalization on ID and OOD, and relatively good ID calibration, but still greatly degrades the OOD calibration.
  4. Temperature scaling (**TS**) helps calibration somewhat, but the gap between ZS OOD and fine-tuned ones still non-negligible.

Method	ID Acc. (↑)	OOD Acc. (↑)	w/o *TS* ID ECE (↓)	w/o *TS* OOD ECE (↓)	w/ *TS* ID ECE (↓)	w/ *TS* OOD ECE (↓)
ZS	0.6832	0.5840	0.0571	0.0836	0.0561	0.0748
FT	0.8153	0.5750	0.0884	0.2186	0.0629	0.1629
FT w/ LS	0.8223	0.5833	0.0460	0.1147	0.0481	0.1282
WiSE-FT	0.8043	0.6350	0.2129	0.1764	0.0872	0.1533
WiSE-FT w/ LS	0.8068	**0.6405**	0.5231	0.3601	0.3382	0.2425
FLYP	0.8258	0.5946	0.0643	0.1831	0.0392	0.1217
FLYP w/ LS	0.8271	0.5975	0.0459	0.1295	0.0427	0.1145
CaRot	**0.8319**	0.6197	**0.0395**	**0.1093**	**0.0380**	**0.0980**

[1] Finetune like you pretrain: Improved finetuning of zero-shot vision models, Goyal et al. 2023

Results

- Findings
  3. **FLYP** [1] achieves strong generalization on ID and OOD, and relatively good ID calibration, but still greatly degrades the OOD calibration.
  4. Temperature scaling (**TS**) helps calibration somewhat, but the gap between ZS OOD and fine-tuned ones still non-negligible.

Method	ID Acc. (↑)	OOD Acc. (↑)	w/o *TS* ID ECE (↓)	w/o *TS* OOD ECE (↓)	w/ *TS* ID ECE (↓)	w/ *TS* OOD ECE (↓)
ZS	0.6832	0.5840	0.0571	0.0836	0.0561	0.0748
FT	0.8153	0.5750	0.0884	0.2186	0.0629	0.1629
FT w/ LS	0.8223	0.5833	0.0460	0.1147	0.0481	0.1282
WiSE-FT	0.8043	0.6350	0.2129	0.1764	0.0872	0.1533
WiSE-FT w/ LS	0.8068	**0.6405**	0.5231	0.3601	0.3382	0.2425
FLYP	0.8258	0.5946	0.0643	0.1831	0.0392	0.1217
FLYP w/ LS	0.8271	0.5975	0.0459	0.1295	0.0427	0.1145
CaRot	**0.8319**	0.6197	**0.0395**	**0.1093**	**0.0380**	**0.0980**

[1] Finetune like you pretrain: Improved finetuning of zero-shot vision models, Goyal et al. 2023

Results

- Findings
  5. Label smoothing (**LS**) remarkably improves the calibration as well as generalization for both contrastive learning and cross-entropy-based learning.
  6. **CaRot** gets superior results overall metrics ID/OOD generalization and calibration which verify the effectiveness of data-dependent LS coupled with contrastive loss.

Method	ID Acc. (↑)	OOD Acc. (↑)	w/o *TS* ID ECE (↓)	w/o *TS* OOD ECE (↓)	w/ *TS* ID ECE (↓)	w/ *TS* OOD ECE (↓)
ZS	0.6832	0.5840	0.0571	0.0836	0.0561	0.0748
FT	0.8153	0.5750	0.0884	0.2186	0.0629	0.1629
FT w/ LS	0.8223	0.5833	0.0460	0.1147	0.0481	0.1282
WiSE-FT	0.8043	0.6350	0.2129	0.1764	0.0872	0.1533
WiSE-FT w/ LS	0.8068	**0.6405**	0.5231	0.3601	0.3382	0.2425
FLYP	0.8258	0.5946	0.0643	0.1831	0.0392	0.1217
FLYP w/ LS	0.8271	0.5975	0.0459	0.1295	0.0427	0.1145
CaRot	**0.8319**	0.6197	**0.0395**	**0.1093**	**0.0380**	**0.0980**

Thank you!

paper

https://changdaeoh.github.io/

https://www.linkedin.com/in/changdae-oh-440587215/

https://twitter.com/Changdae_Oh