IMPERIAL

EuroMLSys

Towards Low-Energy Adaptive Personalization for Resource-Constrained Devices

EuroMLSys' 2024

Yushan Huang, **Josh Millar**, Yuxuan Long, Yuchen Zhao, Hamed Haddadi

22/04/2024

Contents

Motivation

- On-device ML is vulnerable to distribution shift.
- Can group distribution shifts into 3 types: input-, feature-, and output-level [1].
- Currently, most fine-tuning (FT) approaches involve FT a full model or its last layer(s).

Input-Level Shift | Feature-Level Shift | Output-Level Shift

Source | Target | Source | Target | Source **"Yorkshire Terrier"** | Target **"Dog"**

Fig 1. Examples of the 3 types of distribution shift.

Contributions

Target Block Fine-Tuning (TBFT):

- fine-tunes model “blocks” corresponding to the type of shift present
    - e.g., last block(s) for output-level shift.
- achieves performance accuracy than full/last-layer FT at reduced/equivalent energy cost.

Experimental Setup

- **ResNet-26**: inherent “block” structure.
- TBFT on 3 datasets, one for each shift (i.e., input-, feature-, and output-level).
- Set training-sizes to 10%, 20%, and 30% of full dataset (validation- and test-sizes are both 10%).

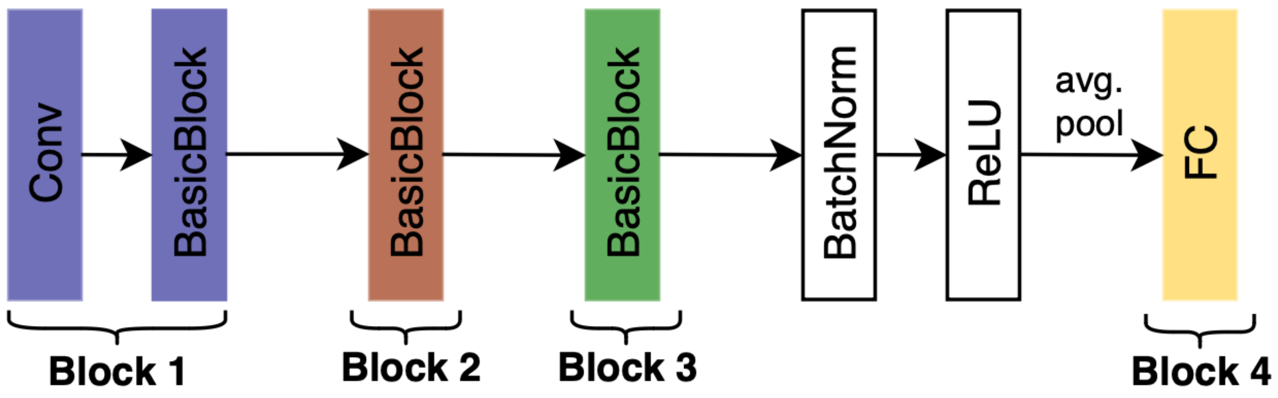


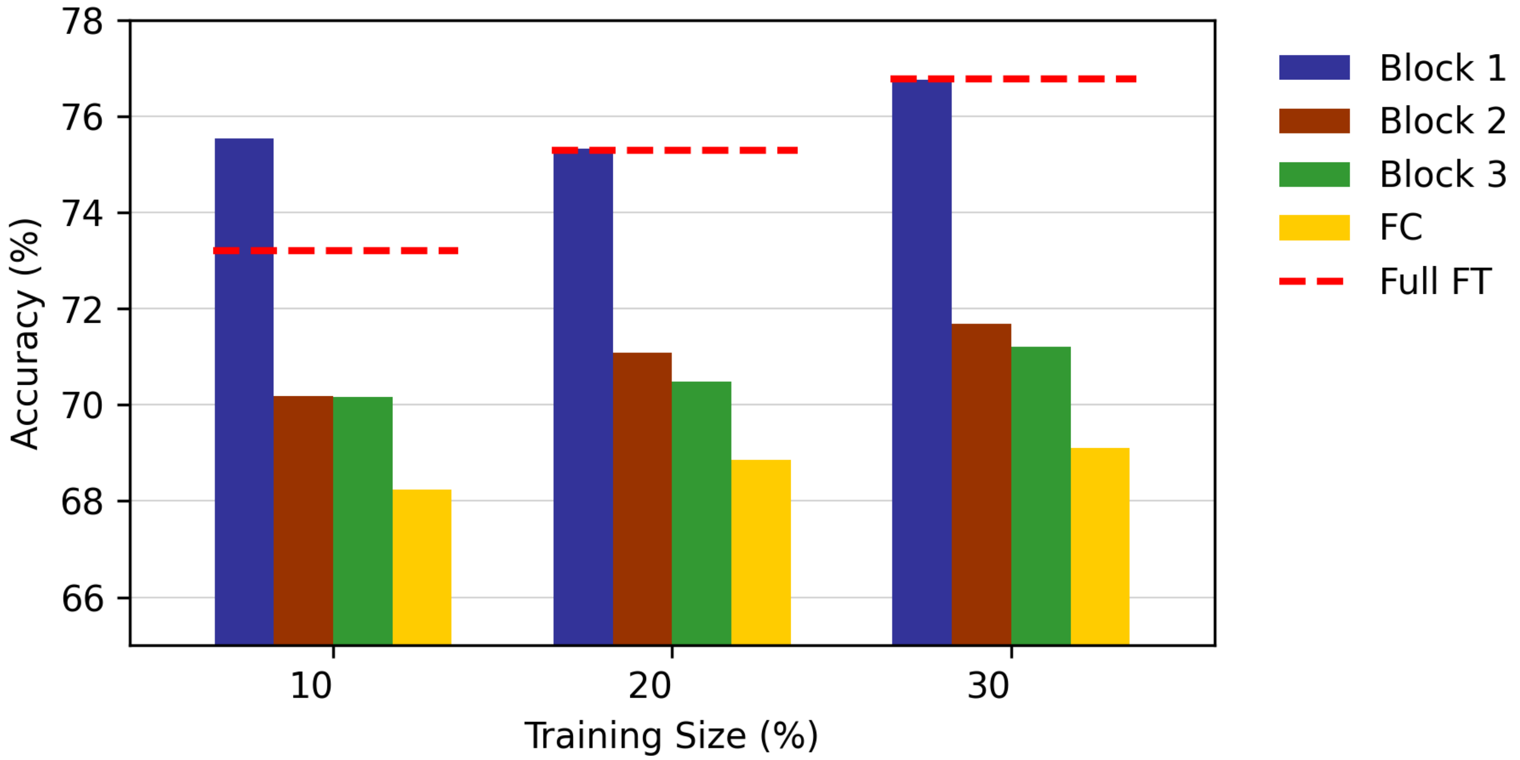
Fig 2. ResNet-26 and its “block” structure.

Results: Input-Level

- Source: CIFAR10
  - 32x32 images with 10 classes.
- Target: CIFAR10-C
  - corrupted images from same classes
  - 14 corruption types (e.g., Gaussian noise, motion blur, brightness).
- Performance averaged across corruption types.

Fig 3. Example corruptions from CIFAR10-C.

Results: Input-Level (cont.)



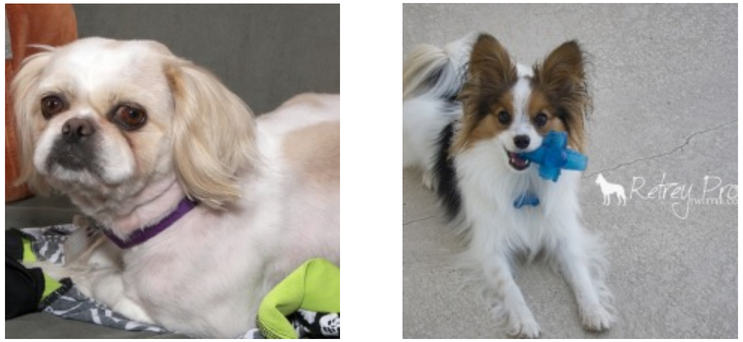
Training Size	Block 1	Block 2	Block 3	FC	Block Avg	Full
10%	**73.53±0.22**	70.18±0.02	70.17±0.06	68.24±0.08	70.53±0.10	73.21±0.37
20%	**75.32±0.10**	71.08±0.07	70.48±0.08	68.86±0.09	71.44±0.09	75.29±0.14
30%	**76.75±0.17**	71.68±0.34	71.20±0.05	69.11±0.06	72.19±0.16	76.78±0.31

Fig 4. Results of TBFT on CIFAR10-C.
The horizontal dashed line shows the accuracy of full FT.

Results: Feature-Level

Living17:

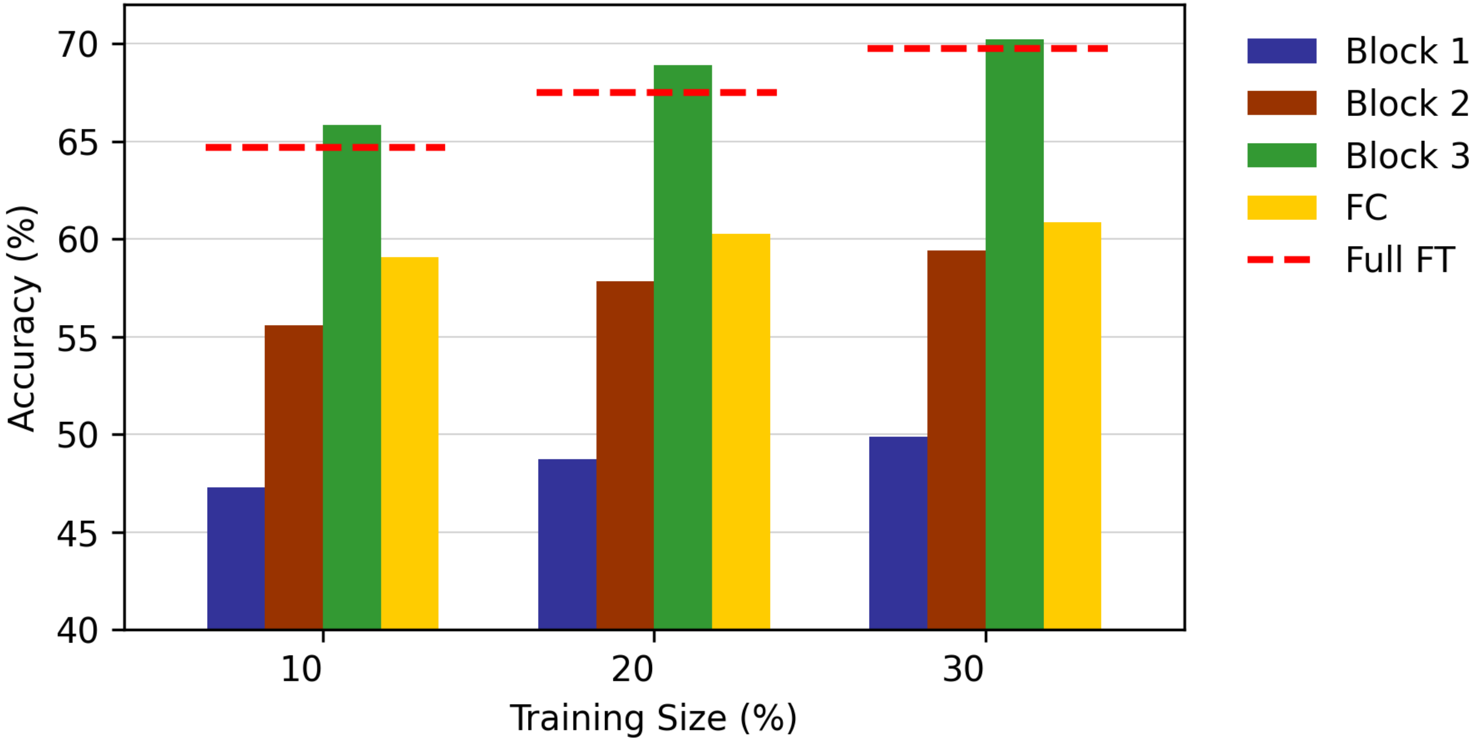
- ImageNet-based i.e., 128x128 images
- Created as part of BREEDS [2]
- Living things across 17 classes, each with 4 different subclasses
- Source and Target distributions consist of **different subclasses.**



Source Target

Fig 5. Example Source and Target images

Results: Feature-Level



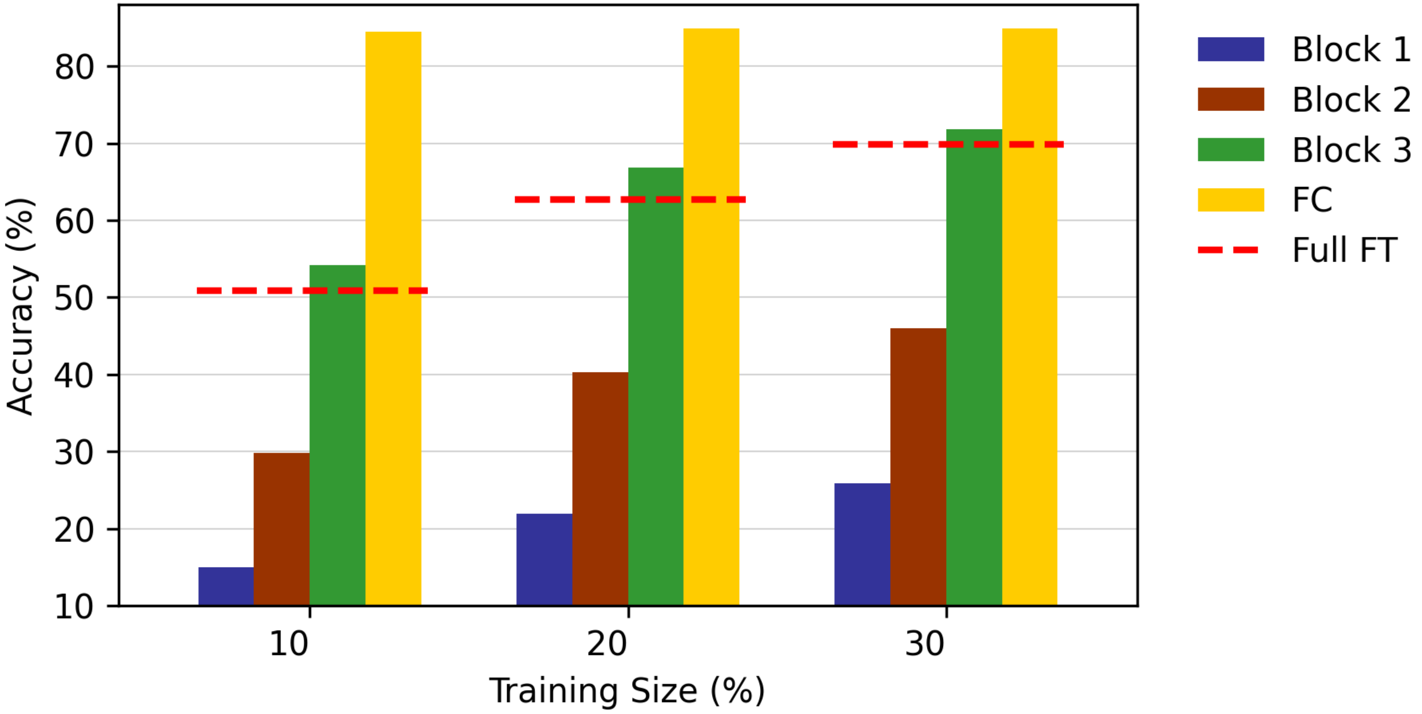
Training Size	Block 1	Block 2	Block 3	FC	Block Avg	Full
10%	47.30±1.05	55.57±0.46	**65.83±0.35**	59.06±0.41	56.94±0.57	64.69±0.29
20%	48.72±0.27	57.85±0.35	**68.88±0.42**	60.27±0.52	58.93±0.39	67.47±0.48
30%	49.88±0.82	59.40±0.46	**70.22±0.65**	60.86±0.78	60.09±0.68	69.76±0.58

Fig 6. Results of TBFT on Living17.

.

Results: Output-Level

- Source: CIFAR10
  - 32x32 images with 10 classes.
- Target: CIFAR10-flip
  - images from same classes with labels flipped.
  - i.e., label *y* becomes *9-y*



Training Size	Block 1	Block 2	Block 3	FC	Block Avg	Full
10%	15.94±0.98	29.78±0.34	54.20±0.28	**84.43±0.10**	46.08±0.24	50.86±0.35
20%	21.95±1.05	40.32±0.59	66.84±0.38	**84.87±0.17**	53.65±0.41	62.65±0.27
30%	25.87±0.68	45.98±0.40	71.78±0.43	**84.91±0.20**	57.14±0.16	69.81±0.45

Fig 7. Results of TBFT on CIFAR10-flip.

Results: Energy Cost

- Device: Raspberry Pi 4 Model B
- Metrics:
  - Energy Cost (J), derived from $E = Pt$ (where P is power and t is runtime)
  - P (W) recorded using Power Monitoring HAT [3]
  - Energy Saving (ES) vs full FT, calculated as $ES = (E_{\text{TBFT}} - E_{\text{FT}})/E_{\text{FT}}$

	Resolution 32, Output 10						Resolution 128, Output 17					
	Block 1	Block 2	Block 3	FC	Block Avg	Full	Block 1	Block 2	Block 3	FC	Block Avg	Full
Time (s)	0.45	0.33	0.29	0.17	0.31	0.53	6.82	5.40	4.28	2.46	4.74	8.12
Energy Cost (J)	0.090	0.066	0.058	0.034	0.062	0.106	1.705	1.350	1.070	0.615	1.185	2.030
Energy Saving (%)	15.09	37.74	45.28	67.92	41.51	-	16.01	33.50	47.29	69.70	41.63	-

Compared to full model FT, TBFT results in an average energy-saving of 41.57%.

Limitations + Future Work

- **Shift Type Discrimination**
  - TBFT relies on knowledge of shift type: unrealistic.
  - Parameter selection methods are gradient-based and expensive.
    - instead, select based on incoming data

- **Unsupervised Personalization Methods**: essential for realistic on-device learning.

- **Multidimensional Personalization**
  - complex and composite shifts in real applications (e.g., coinciding input- and feature-level shifts).

Summary

- We categorize distribution shift into 3 types: input-, feature-, and output-level.
- TBFT fine-tunes model blocks corresponding to distribution shift type.
- TBFT achieves better performance than full/last-layer FT with reduced/equivalent energy-consumption.
  - avg. **15.3% improvement** in **accuracy** over full FT.
  - avg. **41.6% improvement** in **energy-consumption** over full FT.

Thank you

Scan to read!