Green TTA: Benchmarking the Energy Consumption of Test-Time Adaptation

Vivian White* and Evan Shelhamer
The University of British Columbia and Vector Institute for AI
*correspondence: vivwhite@cs.ubc.ca

Introduction

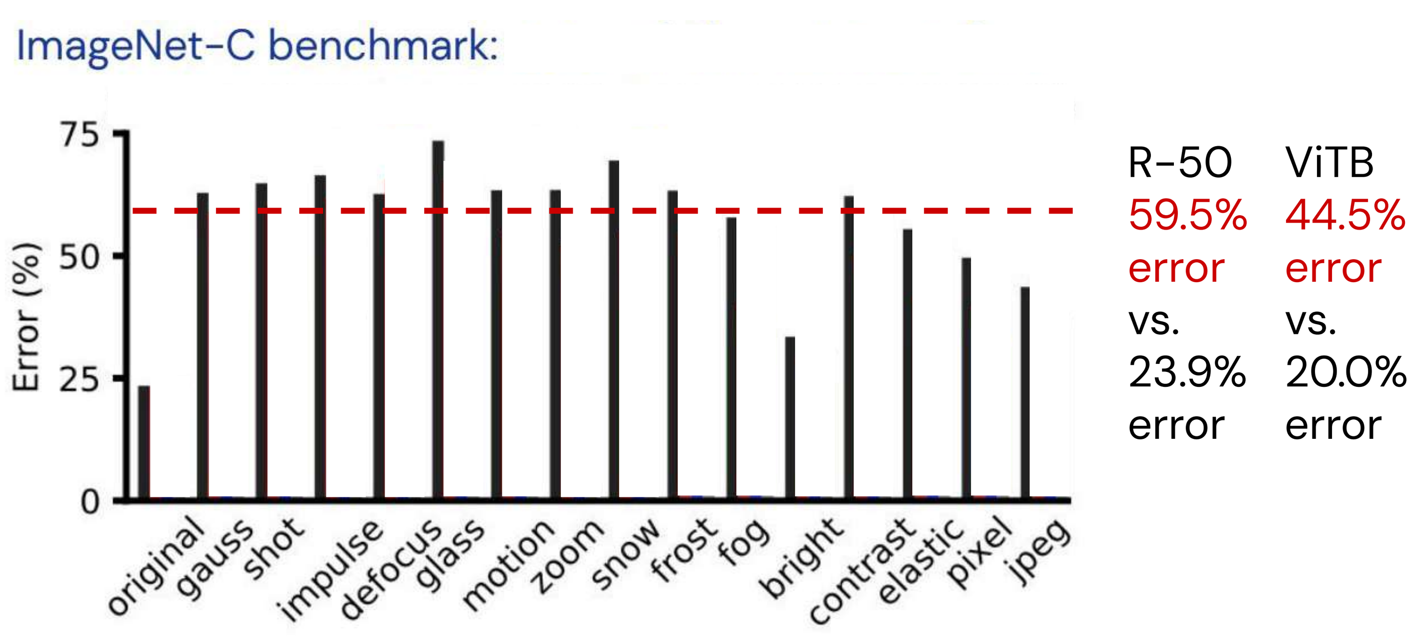
- Since “Green AI” was coined in 2019 by [1], many works have studied the carbon footprint of deep networks at the training and inference levels [1, 2, 3]
- Global AI footprint projected to reach 1,000 TWh by 2030
- Inference accounts for up to 90% of total ML emissions
- Test-time adaptation (TTA) adapts models to distribution shifts at inference time, without modifying model parameters through retraining

What is the energy footprint of test-time adaptation algorithms, and how does it relate to accuracy?

Test-Time Adaptation

- Adapt pre-trained models at inference time to new, shifted data
- ImageNet-C contains 15 types of corruptions that can double error compared to clean ImageNet

Original Image | *Gaussian Noise* | *Motion Blur* | *Snow* | *JPEG Compression*



Methods

- **Inference (Baseline):** Standard pre-trained ResNet-50 evaluated without any inference-time adaptation
- **Tent** [4]: minimizes Shannon entropy of model predictions by only updating affine Batch Normalization parameters
- **EATA** [5]: minimizes entropy with an added sample-filtering mechanism to exclude unreliable and redundant inputs and regularizes model parameters to prevent forgetting
- **LAME** [6]: parameter-free, forward-only adaptation that corrects the model’s final output probability distribution without gradients

CODE CARBON

- We use **CodeCarbon** [7] to track energy consumption (in kWh) of the GPU, CPU, and RAM during inference and adaptation

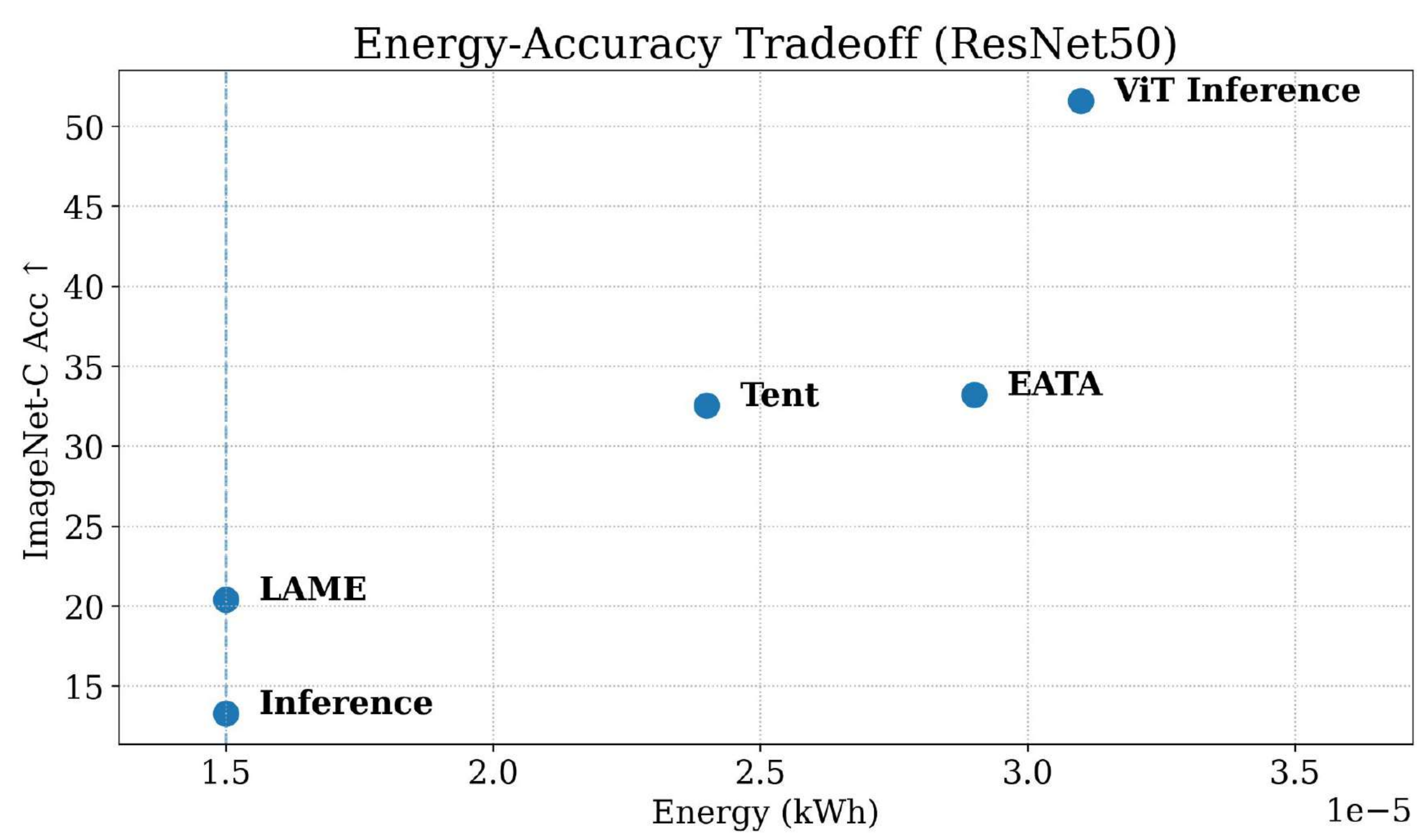
References

[1] Schwartz, Roy, et al. “Green AI.” *Commun. ACM,* vol. 63, no. 12, ACM, Nov 2020, pp. 54–63.
[2] Luccioni, Sasha, et al. “Quantifying the Carbon Emissions of Machine Learning.” *NeurIPS Workshop on Tackling Climate Change with Machine Learning*, 2019,
[3] Li, Baolin, et al. “Clover: Toward Sustainable AI with Carbon-Aware Machine Learning Inference Service.” *International Conference for High Performance Computing, Networking, Storage and Analysis*, ACM, 2023.
[4] Wang, Dequan, et al. “Tent: Fully Test-Time Adaptation by Entropy Minimization.” *ICLR*, 2021.
[5] Niu, Shuaicheng, et al. “Efficient Test-Time Model Adaptation without Forgetting.” *ICML*, 2022.
[6] Boudiaf, Malik, et al. “Parameter-Free Online Test-Time Adaptation.” *CVPR*, 2022.
[7] Courty, Benoit, et al. *Mlco2/Codecarbon: V2.4.1*, Zenodo, 2024.

Initial Results (ResNet50 on ImageNet-C)

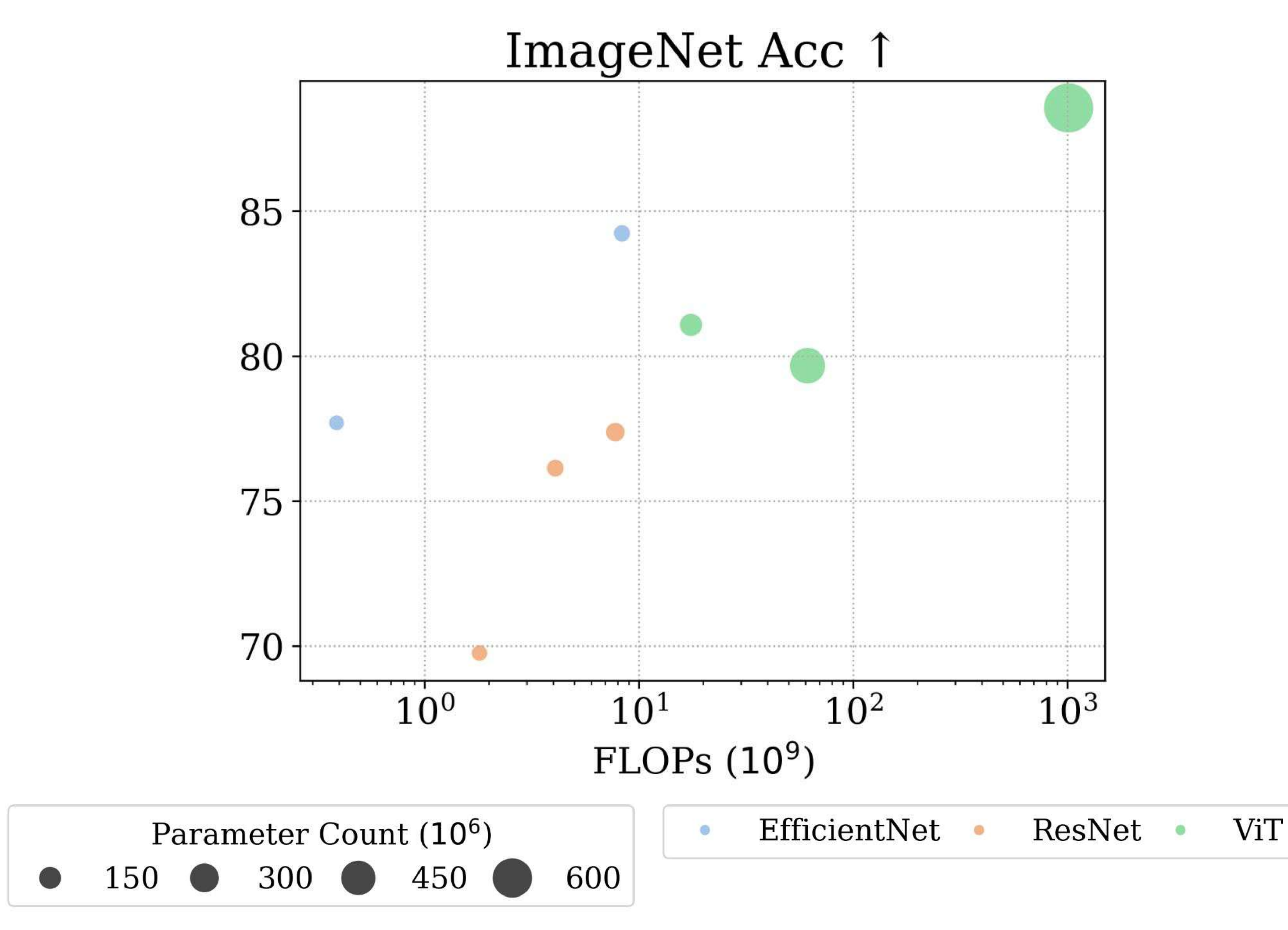
- Average accuracy, energy, and wall-time across 15 ImageNet-C corruptions

Method	Accuracy (%)	Energy (KwH)	Seconds	Adapted Params
Inference	13.29	0.000015	11.89	0
Tent	32.53	0.000024	16.81	106
EATA	33.21	0.000029	19.24	106
LAME	20.39	0.000015	15.35	0
ViT Inference	51.58	0.000031	18.54	0



Next Steps

- Explore different model sizes
  - ResNet18, ResNet50, ResNet101
  - ViT-B, ViT-L, ViT-H
  - EfficientNet, EfficientNetv2



Discussion

- Current TTA methods exhibit an energy-accuracy tradeoff
- Real-world deployment may be resource-constrained
- Next steps explore effect of model size
- Goal of developing more “Green-TTA” methods