Watt for What: Rethinking Deep Learning's Energy-Performance Relationship

Shreyank N Gowda[1], Xinyue Hao[2], Gen Li[2], Shashank N Gowda[3], Xiaobo Jin[4], Laura Sevilla-Lara[2]
University of Oxford[1], University of Edinburgh[2], UCLA[3], Xian Jiatong-Liverpool University[4]

Introduction

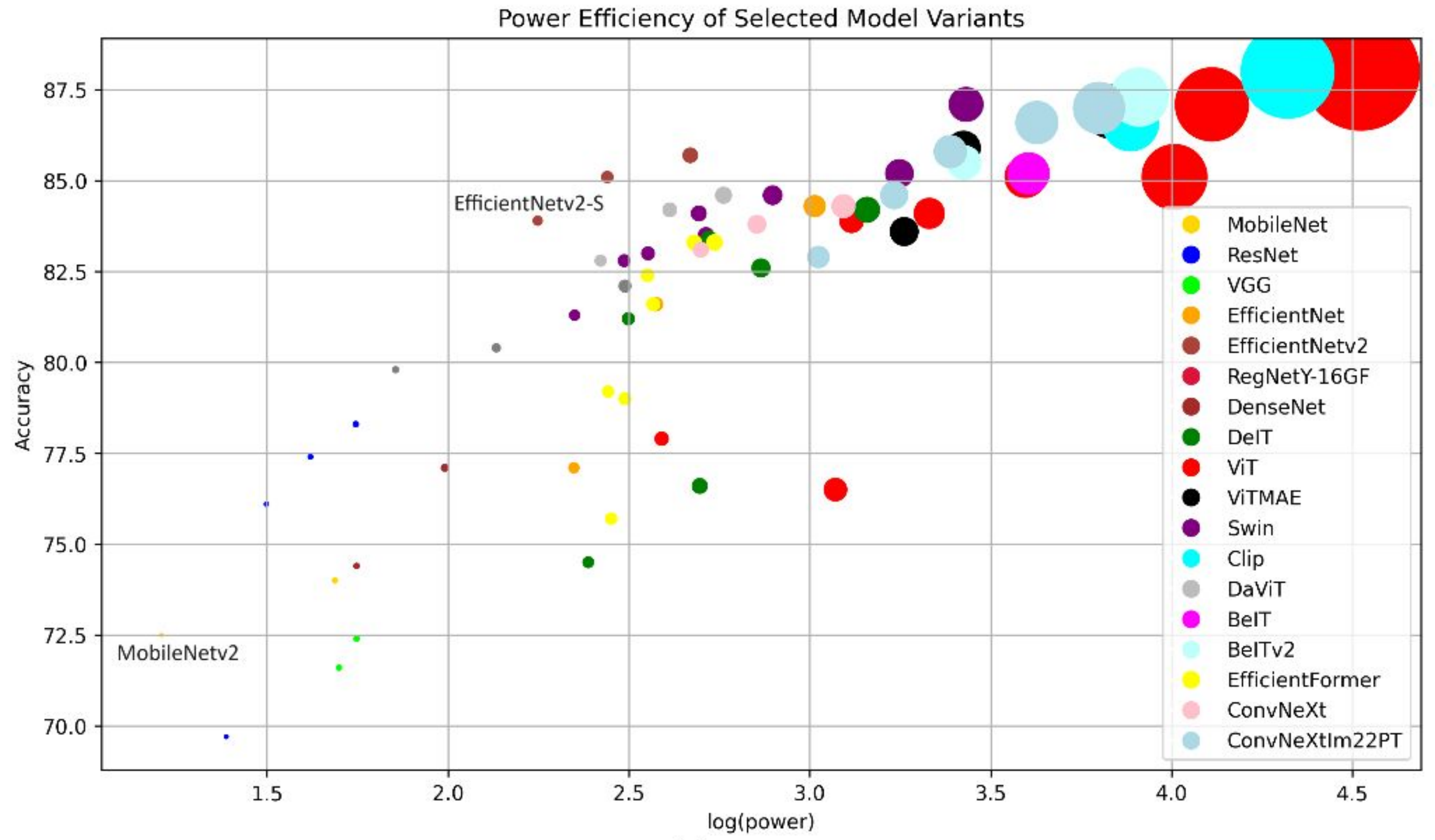
- **Energy Challenge:** Deep learning's increasing computational power demands are driving up electricity consumption, raising environmental and operational concerns.
- **Environmental Impact:** Training large AI models generates significant carbon emissions, which contributes to global climate change.
- **Resource Imbalance:** Smaller research entities are disadvantaged by limited access to computational resources, exacerbating inequity in the research landscape.
- **Proposed Solution:** This study introduces a metric that balances model accuracy with electricity consumption, promoting energy-efficient AI practices.
- **Goal:** To encourage sustainable deep learning and create a more equitable research environment while reducing environmental impact.

Proposed Metric

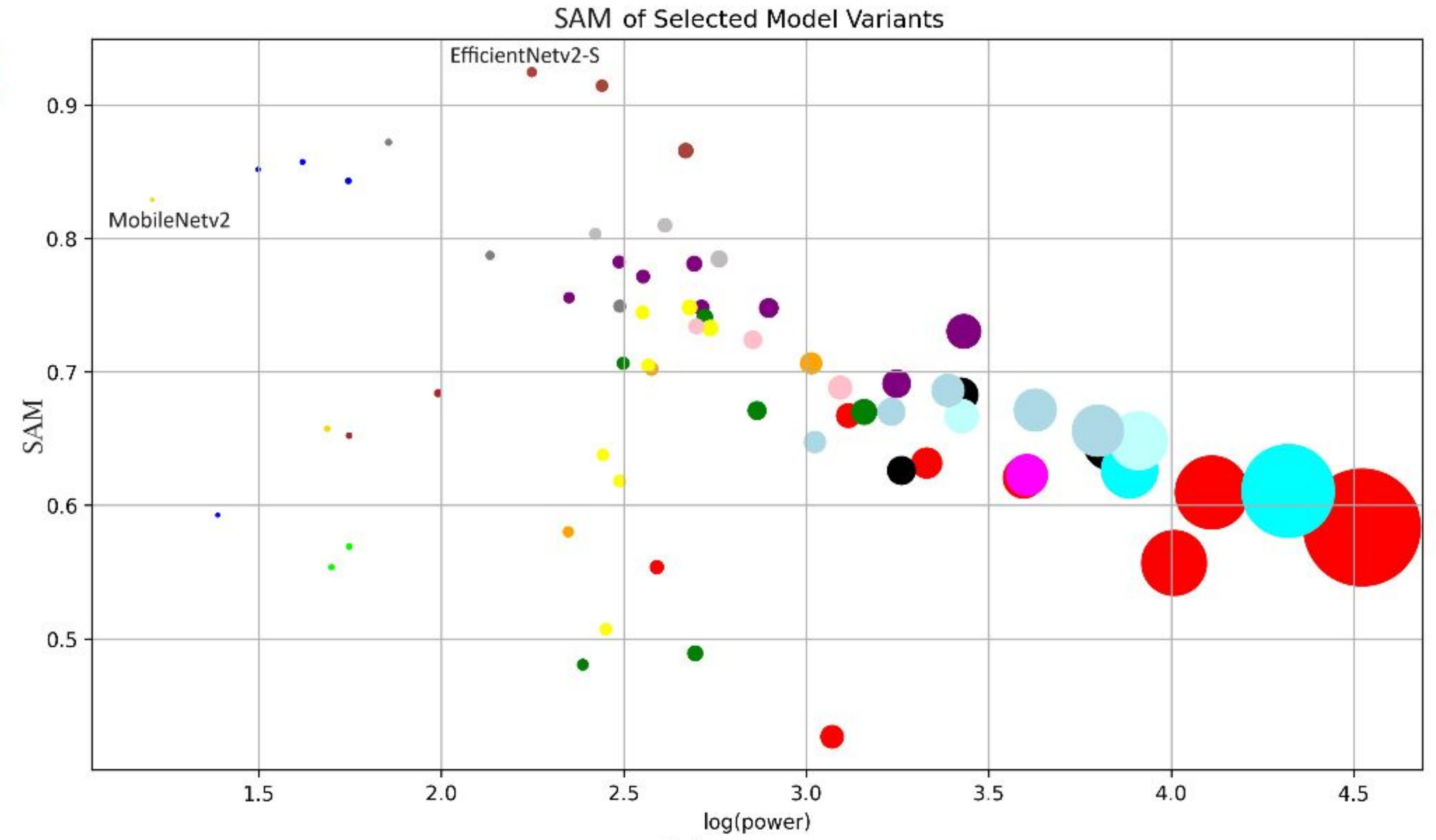
$$SAM = \beta \times \frac{accuracy^{\alpha}}{log_{10}(electricity)}$$

- **Energy Efficiency Focus:** The metric emphasizes the trade-off between model accuracy and electricity consumption, penalizing high energy use.
- **Logarithmic Scaling:** It uses a logarithmic scale to account for differences in power consumption across models, ensuring fair comparisons.
- **Accuracy-Power Balance:** The metric rewards models that achieve high accuracy with minimal energy usage, encouraging sustainable model development.
- **Equitable Comparison:** It allows smaller entities with fewer computational resources to compete fairly by prioritizing efficiency over raw computational power.
- **Scalability:** The metric can be applied across various deep learning tasks, from image classification to video action recognition, ensuring broad utility.

Results



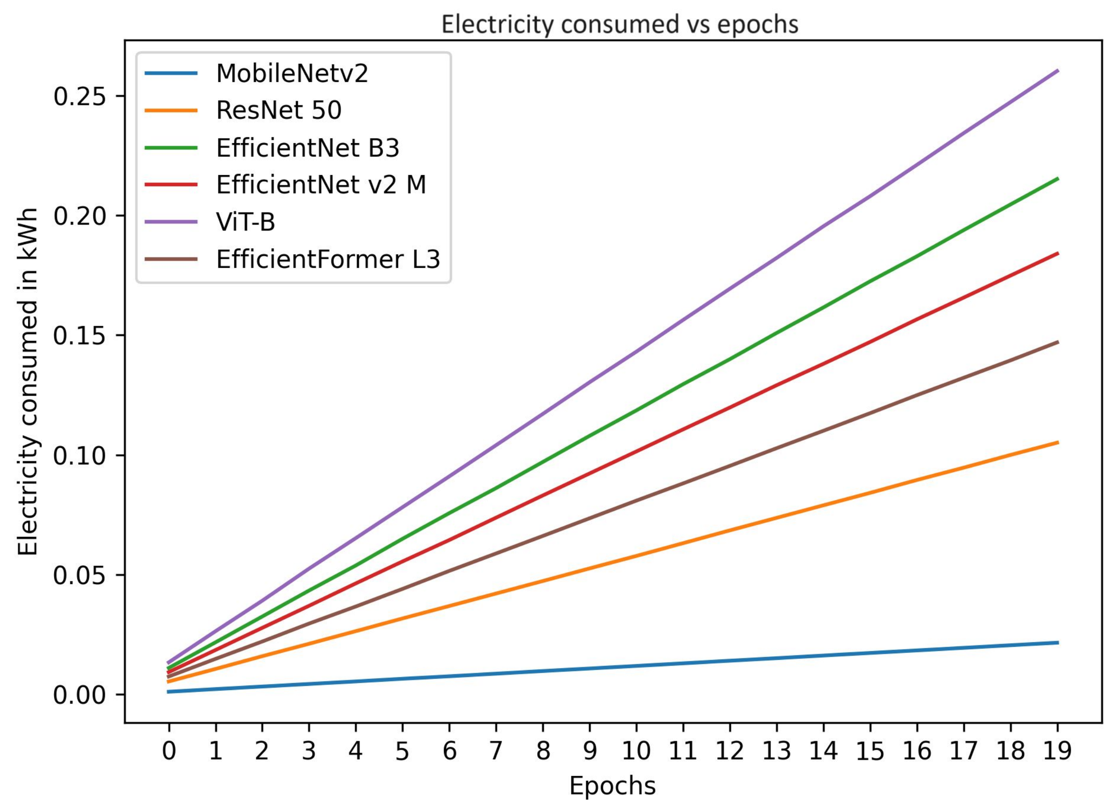
(a)

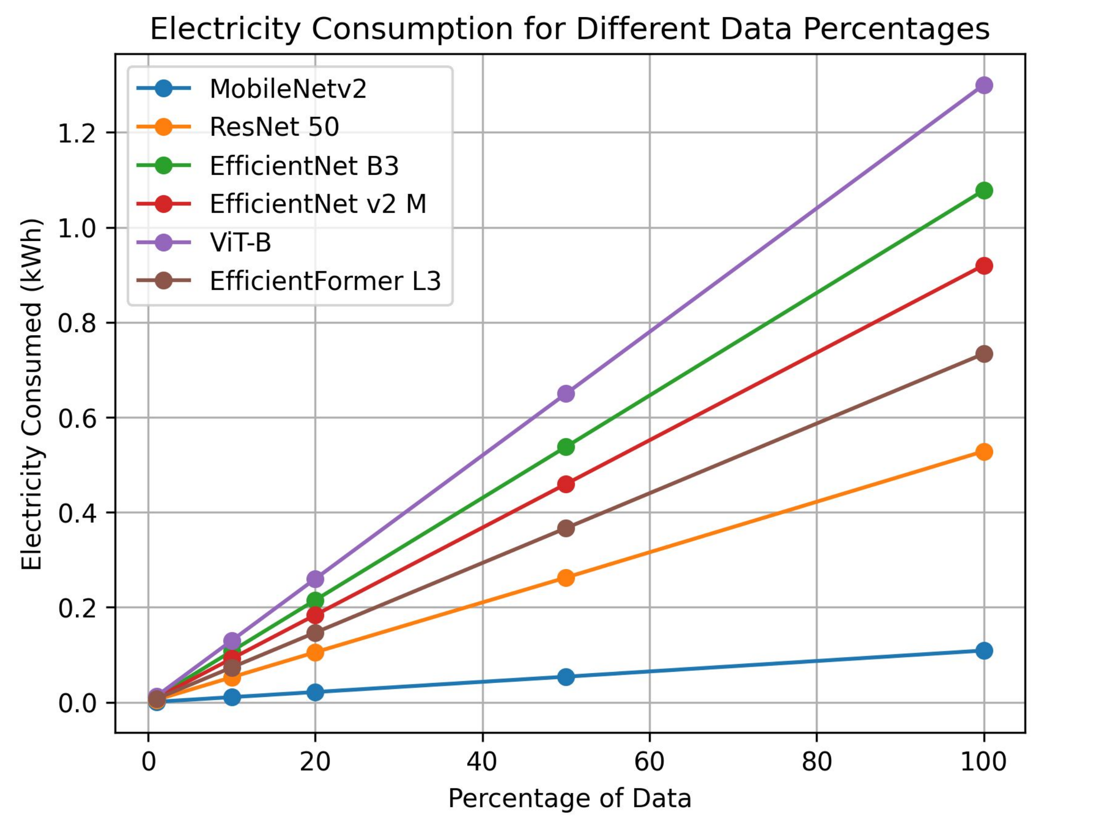


(b)

- Using SAM achieves a better balance between accuracy and electricity. In (a), lower accuracy models like MobileNet, ResNet, and EfficientNet are depicted poor compared to ViT-H. In (b), ViT-H and CLIP are penalized for high electricity usage, while EfficientNet and MobileNet rise.
  Marker points = electricity consumption.

How do we scale up?





- Given the time to reproduce results, a natural question is how can we scale things up in order to reach these results faster. (Left) Electricity Consumed Across Epochs: Linear Relationship Observed. (Right) Electricity Consumed Across Different Percentages of Data: Linear Relationship Observed.

Conclusion

- **Energy Efficiency Priority:** Balancing model accuracy with energy consumption is crucial to mitigate the environmental impact of deep learning.
- **New Metric:** The proposed metric offers a fair way to assess models by considering both accuracy and electricity use, encouraging more sustainable AI development.
- **Promoting Equity:** This approach levels the playing field, allowing smaller research entities to compete with well-resourced organizations based on efficiency.
- **Sustainable AI Future:** Encouraging energy-efficient practices in deep learning will reduce environmental harm while advancing innovation in a more equitable way.

Full paper: With results on Action Recognition and Instance Segmentation! Scan for more details:

Contact

Shreyank N Gowda: shreyank.narayanagowda@eng.ox.ac.uk

References

[1] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).
[2] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[3] Tan, M., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946.
[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words. arXiv preprint arXiv:2010.11929, 7.
[5] Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y. and Ren, J., 2022. Efficientformer: Vision transformers at mobilenet speed. Advances in Neural Information Processing Systems, 35, pp.12934-12949.