

Adaptive Sampling for Efficient Training of Large Vision Models

First Author¹ Second Author² Third Author¹¹Department of Computer Science, Example University²Institute for Applied Mathematics, Another University

Introduction

Training large vision models is dominated by the cost of passing every image through the network at every epoch, even though most examples stop contributing useful gradient early.

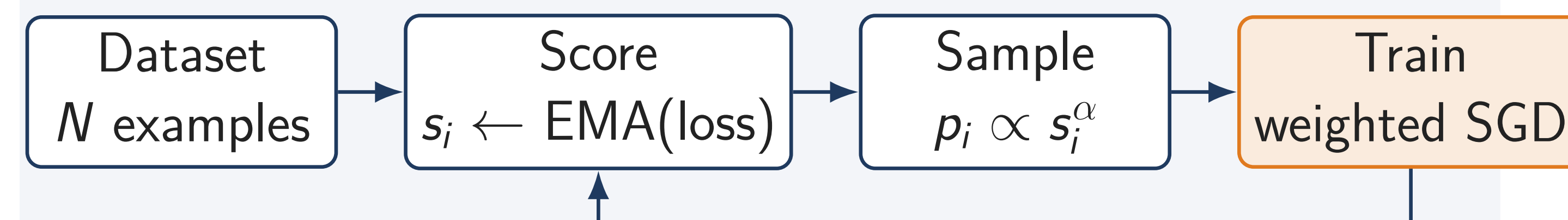
- **Problem:** uniform sampling wastes 40–60 % of compute on examples the model already fits.
- **Idea:** score examples online with a cheap proxy and sample proportionally to their estimated learning value.
- **Result:** same accuracy in 0.55× the wall-clock time on three benchmarks, with no change to the architecture.

Method

Each example i receives a score s_i from an exponential moving average of its recent loss, and the sampling distribution is

$$p_i = (1 - \varepsilon) \frac{s_i^\alpha}{\sum_j s_j^\alpha} + \frac{\varepsilon}{N}, \quad (1)$$

where α controls sharpness and ε guarantees that every example is still visited. Gradients are importance-weighted by $1/(Np_i)$ so the estimator stays unbiased.



The scorer runs on the forward pass that training performs anyway, so the overhead is below 2% of an epoch.

Results

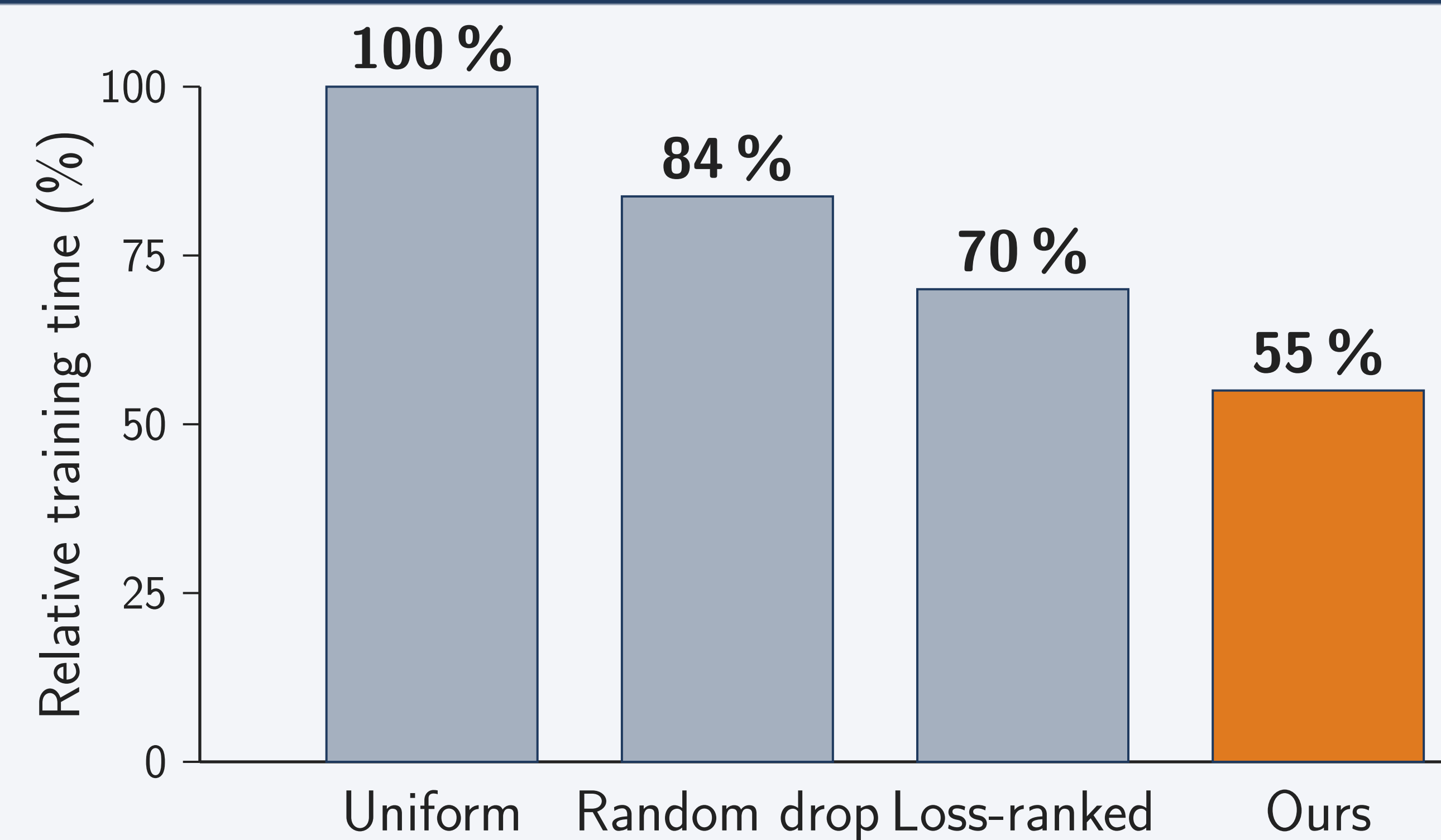


Figure 1: Wall-clock time to reach the baseline accuracy on ImageNet-1k (ResNet-50). Lower is better.

Method	CIFAR-100	ImageNet-1k	Places365
Uniform sampling	80.1	76.3	54.7
Random drop (50%)	77.9	73.8	52.1
Loss-ranked	80.0	76.0	54.4
Ours	80.4	76.5	54.9

Table 1: Top-1 accuracy (%) at a fixed compute budget of 0.55× the uniform baseline.

Ablation

- Sharpness $\alpha \in [0.5, 1.0]$ is robust; $\alpha > 2$ overfits hard examples.
- Removing the ε floor costs 0.8 points of accuracy.
- EMA half-life of 3 epochs works across all three datasets.

Key findings

- **1.8× faster** training to target accuracy.
- **No accuracy loss** on three datasets and two architectures.
- **Drop-in:** a data loader change only, no new hyperparameters beyond α and ε .

Conclusion

Adaptive sampling turns the loss signal the model already produces into a free curriculum. Because the estimator is unbiased, standard convergence guarantees still apply, and the practical speed-up grows with dataset redundancy.

Future work: extend the scorer to multimodal pre-training, study interactions with data augmentation, and release a framework-agnostic loader.

References

1. A. Katharopoulos and F. Fleuret. Not all samples are created equal: Deep learning with importance sampling. *ICML*, 2018.
2. Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. *ICML*, 2009.
3. T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. *ICLR*, 2016.
4. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016.

Acknowledgements

This work was supported by Example Research Council grant EX/0000001. Compute was provided by the Example University HPC centre.