Exploiting Network Compressibility and Topology in Zero-Cost NAS

[1,★] Lichuan Xiang [1,★] Rosco Hunter [2,★] Łukasz Dudziak [1] Minghao Xu [1,2] Hongkai Wen

[1] University of Warwick [2] Samsung AI Center, Cambridge [★] Equal Contribution

Introduction

Neural Architecture Search (NAS) streamlines and systematises the design of high performance neural networks. Zero-cost metrics have emerged as a low-latency approach within NAS, enabling the prediction of a trained network's performance by probing it at initialization. In this paper, we consider how compressibility and layer-wise partitions can explain and extend a number of existing metrics. Drawing from these insights, we produce two novel metrics that achieve state-of-the-art (SOTA) results.

Gradient Centric Compressibility (SSNR)

Many metrics [1] sum up the saliency (importance) scores of the parameters of a network, which discards potentially important statistical information. ZiCo [2] stands out as a method that goes beyond mere saliency aggregation.

$$ZiCo = \sum_l \log\left(\sum_{\omega\in\Theta^l} \frac{\mathbb{E}[|\nabla_\omega L(\boldsymbol{X_i},\boldsymbol{y_i};\Theta)|]}{\sqrt{\mathrm{Var}(|\nabla_\omega L(\boldsymbol{X_i},\boldsymbol{y_i};\Theta)|)}}\right)$$

Where Θ are the parameters of the network, which are partitioned by the layers. We frame ZiCo as a measure of gradient compressibility over data samples. The intuition is that a high-variance gradient is suggestive of poor training dynamics that are dominated by a small subset of the inputs. Motivated by the insight that compressibility over the data plays an important role in network performance, we ask whether compressibility over the parameters also proves insightful. Intuitively, if two networks have similar saliency sums, the one whose parameters have more diverse scores has a better chance of being successfully pruned.

Therefore, we propose the Saliency Signal-to-Noise Ratio (SSNR):

$$S_n^l = \sum_{\omega\in\Theta^l} S(\omega) \quad and \quad \sigma^l = \sqrt{\frac{1}{|\Theta^l|}\sum_{\omega\in\Theta^l}\left(S(\omega) - \frac{S_n^l}{|\Theta^l|}\right)^2}$$

$$SSNR = \sum_l \frac{S_n^l}{\sigma^l}$$

Where S denotes any parameter saliency score, as opposed to the saliency instance considered by ZiCo, $S = |\nabla L|$. Motivated by the Conservation of Synaptic Saliency [3] and ZiCo's formula, we separately calculate the SNR for each layer of the network. But can this notion of layer-wise compressibility be applied beyond just saliency score to also probe activation patterns?

Combining Gradient and Activation Centric Compressibility (T-CET)

An existing metric, NASWOT [4], already measures the compressibility of activation patterns over the data:

$$NASWOT = \log|K|$$

Where K is the Gram matrix of the activation patterns. However, we observe that NASWOT might too be improved through layer-wise partitions. As such, we generate a separate Gram for layer's activation patterns and sum their log-determinants:

$$Layerwise \quad NASWOT = \sum_l \log|K^l|$$

Motivated by their differences, we combine the layer-wise measures of activation and gradient compressibility via a dot product:

$$T-CET = \sum_l \frac{S_n^l}{\sigma^l} \cdot \log|K^l|$$

Results

SSNR and T-CET outperformed SOTA metrics across a variety of search spaces. For example, in NASBench-201 [5] taking SNIP's signal-to-noise ratio (SSNR) far outperformed simple aggregation (SNIP):

	Synflow[3]	SNIP[6]	ZiCo[2]	ZenScore[7]	NASWOT[4]	SSNR	T-CET
CIFAR-10	0.54	0.46	0.61	0.29	0.58	0.68	**0.69**
CIFAR-100	0.57	0.46	0.61	0.28	0.62	**0.65**	**0.65**
ImageNet 16	0.56	0.43	0.60	0.29	0.60	**0.63**	0.62

Table 1. Kendall-tau correlation between different zero-cost metrics and model accuracy on NASBench-201. SNIP is used as the saliency score for SSNR and T-CET.

In the practical setting of the ZenNAS search space, T-CET outperformed all the other proxies in identifying high-performance architectures:

	Random	Synflow	ZiCo	ZenScore	NASWOT	TE-Score[8]	T-CET
CIFAR-10	93.5	95.1	97.0	96.2	96.0	96.1	**97.2**
CIFAR-100	71.1	75.9	80.2	80.1	77.5	77.2	**80.4**

Table 2. Top-1 Acc. % for zero-cost proxies on ZenNAS Search-Space. Budget: model size N$<$1M. SNIP is used as the saliency score for T-CET.

References

[1] Abdelfattah, M. S., Mehrotra, A., Dudziak, Ł., and Lane, N. D. (2020). Zero-Cost Proxies for Lightweight NAS. In International Conference on Learning Representations (ICLR).
[2] Li, G., Yang, Y., Bhardwaj, K., and Marculescu, R. (2023). ZiCo: Zero-shot NAS via inverse Coefficient of Variation on Gradients. In International Conference on Learning Representations (ICLR).
[3] Tanaka, H., Kunin, D., Yamins, D. L., and Ganguli, S. (2020). Pruning neural networks without any data by iteratively conserving synaptic flow. In Advances in Neural Information Processing Systems (NeurIPS).
[4] Mellor, J., Turner, J., Storkey, A., and Crowley, E. J. (2021). Neural architecture search without training. In International Conference on Machine Learning (ICML).
[5] Dong, X. and Yang, Y. (2020). NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search. In International Conference on Learning Representations (ICLR).
[6] Lee, N., Ajanthan, T., and Torr, P. H. (2019). Snip: Single-shot network pruning based on connection sensitivity. In International Conference on Learning Representations (ICLR).
[7] Lin, M., Wang, P., Sun, Z., Chen, H., Sun, X., Qian, Q., Li, H., and Jin, R. (2021). Zen-nas: A zero-shot nas for high-performance deep image recognition. In International Conference on Computer Vision (ICCV).
[8] Chen, W., Gong, X., and Wang, Z. (2021). Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. In International Conference on Learning Representations (ICLR).