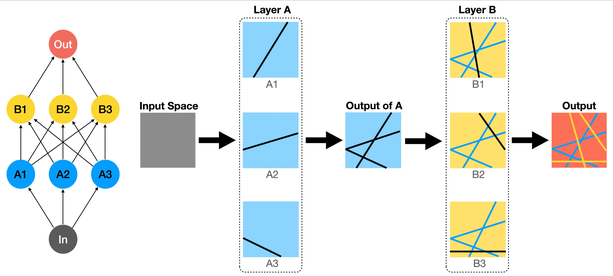
# SWAP-NAS: Sample-Wise Activation Patterns for Ultra-Fast NAS

Yameng Peng, Andy Song*, Haytham M. Fayek, Vic Ciesielski, Xiaojun Chang

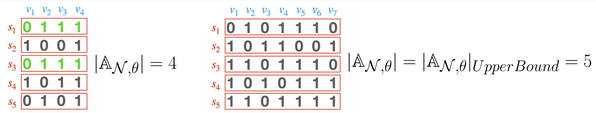
Scan QR code to access full paper and source code

Paper | Source Code

## Can We Assess the Performance of Neural Networks without Training?



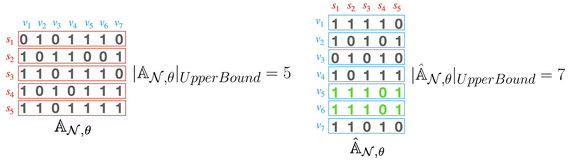

Neural networks employing piecewise linear activation functions, like ReLU, partition its input space into regions. The number of **distinct regions** serves as an indicator of the network's functional complexity. A network with more distinct regions is capable of capturing more complex features, thereby exhibiting higher expressivity.




Define a batch of inputs contains ***S*** samples, ***V*** denotes pre-activation values, $p_v^{(s)}$ is a single post-activation value from $v^{th}$ pre-activation value at $s^{th}$ sample, a set of post-activation values is defined as:

$$\mathbb{A}_{\mathcal{N},\theta} = \left\{ \mathbf{p}^{(s)} : \mathbf{p}^{(s)} = \mathbb{1}(p_v^{(s)})_{v=1}^{V},\ s \in \{1, \ldots, S\} \right\}$$

The number of distinct regions is the cardinality of $\mathbb{A}_{\mathcal{N},\theta}$. In this case, the **upper bound** of $|\mathbb{A}_{\mathcal{N},\theta}|$ is equal to ***S***. Thus, when ***V*** is large, $|\mathbb{A}_{\mathcal{N},\theta}|$ can easily reaches the upper bound (visualised above).




We swap the activation matrix from pre-activation value-wise to sample-wise (visualised above), the set of sample-wise activation values is defined as:

$$\hat{\mathbb{A}}_{\mathcal{N},\theta} = \left\{ \mathbf{p}^{(v)} : \mathbf{p}^{(v)} = \mathbb{1}(p_s^{(v)})_{s=1}^{S},\ v \in \{1, \ldots, V\} \right\}$$

In **sample-wise activation patterns (SWAP)**, the upper bound of $|\hat{\mathbb{A}}_{\mathcal{N},\theta}|$ is ***V***, which would be much larger than ***S***, thus, $|\hat{\mathbb{A}}_{\mathcal{N},\theta}|$ offers much higher capacity for distinct regions than $|\mathbb{A}_{\mathcal{N},\theta}|$.

The process of collecting activation patterns requires only a feedforward pass with a mini-batch of inputs and without back-propagation, thus, SWAP-Score is a **training-free** performance indicator.

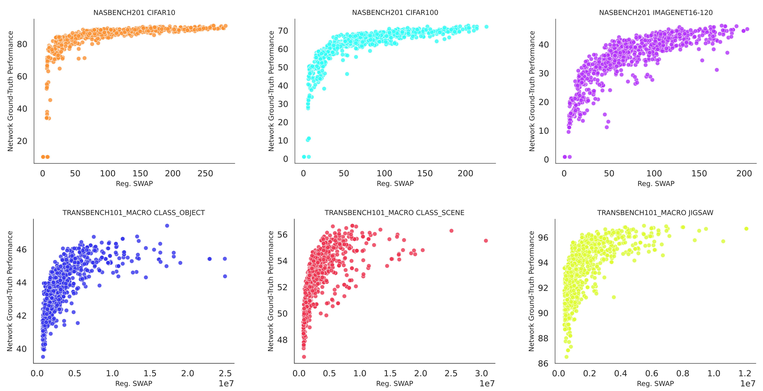
*Corresponding author: andy.song@rmit.edu.au

## SWAP-Scores Are Highly Correlated with the Actual Performance

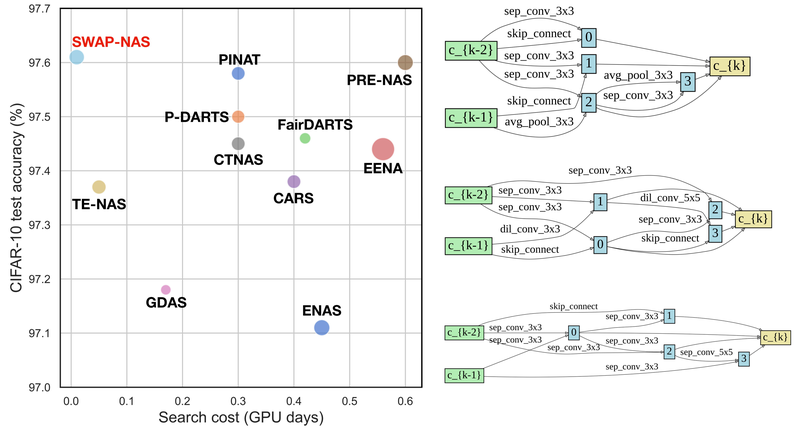
Spearman's Correlation between Training-free Metrics and Networks' Performance

| | TNB101_MICRO-AUTOENC | TNB101_MACRO-OBJECT | TNB101_MACRO-JIGSAW | NB101-CF10 | TNB101_MACRO-AUTOENC | TNB101_MACRO-SCENE | NB301-CF10 | TNB101_MICRO-JIGSAW | TNB101_MICRO-OBJECT | TNB101_MICRO-SCENE | NB201-IMGNT | NB201-CF10 | NB201-CF100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| plain | 0.07 | -0.19 | -0.30 | -0.32 | -0.11 | -0.19 | -0.33 | 0.35 | 0.34 | 0.23 | -0.24 | -0.25 | -0.19 |
| grasp | -0.12 | -0.64 | -0.27 | 0.27 | -0.02 | -0.43 | 0.34 | -0.13 | -0.23 | -0.27 | 0.54 | 0.51 | 0.53 |
| num_lr | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.02 | 0.00 | 0.00 | 0.00 | 0.07 | 0.07 | 0.07 |
| fisher | -0.58 | -0.30 | -0.26 | -0.28 | -0.18 | -0.12 | -0.27 | 0.32 | 0.46 | 0.67 | 0.47 | 0.49 | 0.53 |
| grad_norm | -0.32 | -0.55 | -0.27 | -0.24 | 0.32 | -0.34 | -0.03 | 0.37 | 0.40 | 0.66 | 0.55 | 0.57 | 0.62 |
| snip | -0.26 | -0.37 | -0.20 | -0.19 | 0.20 | -0.14 | -0.04 | 0.43 | 0.47 | 0.71 | 0.55 | 0.58 | 0.62 |
| epe_nas | 0.00 | -0.01 | 0.02 | -0.00 | 0.00 | -0.01 | 0.00 | 0.16 | 0.40 | 0.52 | 0.33 | 0.70 | 0.58 |
| l2_norm | 0.05 | 0.09 | 0.15 | 0.51 | -0.19 | 0.29 | 0.46 | 0.35 | 0.33 | 0.53 | 0.68 | 0.68 | 0.71 |
| zen | 0.14 | 0.10 | 0.24 | 0.60 | -0.01 | 0.28 | 0.44 | 0.52 | 0.55 | 0.72 | 0.38 | 0.33 | 0.34 |
| jacov | 0.18 | 0.07 | 0.19 | -0.29 | 0.46 | 0.19 | -0.05 | 0.56 | 0.51 | 0.75 | 0.70 | 0.74 | 0.70 |
| params | -0.00 | 0.17 | 0.17 | 0.38 | -0.18 | 0.33 | 0.46 | 0.44 | 0.46 | 0.63 | 0.69 | 0.72 | 0.73 |
| synflow | 0.00 | 0.12 | 0.34 | 0.31 | 0.00 | 0.28 | 0.18 | 0.48 | 0.49 | 0.72 | 0.75 | 0.73 | 0.76 |
| zico | 0.13 | 0.04 | 0.08 | 0.37 | 0.07 | 0.23 | 0.53 | 0.52 | 0.50 | 0.70 | 0.77 | 0.76 | 0.79 |
| flops | -0.01 | 0.79 | 0.64 | 0.37 | 0.76 | 0.85 | 0.43 | 0.45 | 0.46 | 0.65 | 0.67 | 0.69 | 0.71 |
| nwot | 0.03 | 0.84 | 0.76 | 0.32 | 0.66 | 0.89 | 0.48 | 0.43 | 0.40 | 0.60 | 0.77 | 0.77 | 0.80 |
| **swap** | -0.09 | 0.86 | 0.77 | 0.47 | 0.68 | 0.83 | 0.58 | 0.43 | 0.41 | 0.57 | 0.76 | 0.79 | 0.81 |
| **reg_swap** | -0.00 | 0.86 | 0.77 | 0.75 | 0.68 | 0.83 | 0.63 | 0.47 | 0.48 | 0.67 | 0.87 | 0.88 | 0.90 |

NAS-Bench-Suite-Zero, a standardised framework for fairly verifying the effectiveness of training-free metrics, incorporates five neural architecture spaces, NAS-Bench-101/201/301 and TransNAS-Bench-101-Micro and -Macro. On the NAS-Bench-Suite-Zero, the same setup is applied to all metrics, SWAP-Score and its regularised version outperform 15 other training-free metrics in the majority of the tasks.




## Sample-Wise Activation Patterns for Ultra-Fast Neural Architecture Search




We integrated the regularised SWAP-Score with an evolutionary search algorithm as SWAP-NAS, applied it to the DARTS search space. SWAP-NAS only requires approximately **6** minutes to search for neural networks for the CIFAR-10 dataset, and **9** minutes for the ImageNet dataset.

| | Test Error Top1/Top5 | Params (M) | GPU Days | Search Method | Evaluation | Searched Dataset |
|---|---|---|---|---|---|---|
| ProxylessNAS (Cai et al., 2018) | 24.9 / 7.5 | 7.1 | 8.3 | Gradient | One-shot | ImageNet |
| DARTS (Liu et al., 2019)† | 26.7 / 8.7 | 4.7 | 4 | Gradient | One-shot | CIFAR-10 |
| EvNAS (Sinha & Chen, 2021)† | 24.4 / 7.4 | 5.1 | 3.83 | Evolution | One-shot | CIFAR-10 |
| PRE-NAS (Peng et al., 2023)† | 24.0 / 7.8 | 6.2 | 0.6 | Evolution | Predictor | CIFAR-10 |
| FairDARTS (Chu et al., 2020)† | 26.3 / 8.3 | 2.8 | 0.42 | Gradient | One-shot | CIFAR-10 |
| CARS (Yang et al., 2020)† | 24.8 / 7.5 | 5.1 | 0.4 | Evolution | One-shot | CIFAR-10 |
| P-DARTS (Chen et al., 2019)† | 24.4 / 7.4 | 4.9 | 0.3 | Gradient | One-shot | CIFAR-10 |
| CTNAS (Chen et al., 2021b) | 22.7 / 7.5 | - | 0.1+50 | Reinforce | Predictor | ImageNet |
| ZiCo (Li et al., 2023) | 21.9 / - | - | 0.4 | Evolution | Training-free | ImageNet |
| PINAT-T (Lu et al., 2023)† | 24.9 / 7.5 | 5.2 | 0.3 | Evolution | Predictor | CIFAR-10 |
| GDAS (Dong & Yang, 2019a) | 27.5 / 9.1 | 4.4 | 0.17 | Gradient | One-shot | CIFAR-10 |
| TE-NAS (Chen et al., 2021a)† | 24.5 / 7.5 | 5.4 | 0.17 | Pruning | Training-free | ImageNet |
| QE-NAS (Sun et al., 2022)† | 25.5 / - | 3.2 | 0.02 | Evolution | Training-free | ImageNet |
| **SWAP-NAS ($\mu$=25, $\sigma$=25)** † | 24.0 / 7.6 | 5.8 | 0.006 | Evolution | Training-free | ImageNet |

**Future Work**. Our future work aims to extend the sample-wise activation patterns to other types of activation functions such as GELU. Considering the increasing popularity of research on Transformers, evaluating the performance of Transformers in a training-free fashion would benefit the research community and have practical significance.