A. NAS Benchmarks

A.1. NAS-Bench-101

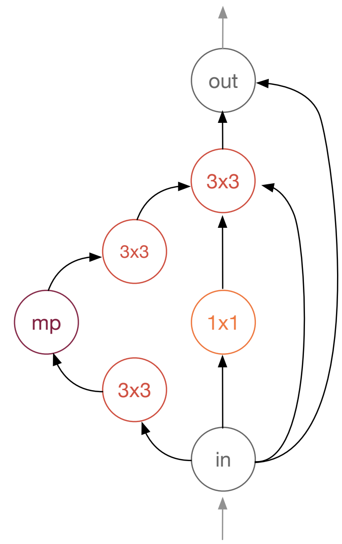
In NAS-Bench-101, the search space is restricted as follows: algorithms must search for an individual cell which will be repeatedly stacked into a pre-defined skeleton, shown in Figure 8c. Each cell can be represented as a directed acyclic graph (DAG) with up to 9 nodes and up to 7 edges. Each node represents an operation, and each edge represents a state. Operations can be chosen from: 3×3 convolution, 1×1 convolution, 3×3 max pool. An example of this is shown in Figure 8a. After de-duplication, this search space contains 423,624 possible neural networks. These have been trained exhaustively, with three different initialisations, on the CIFAR-10 dataset for 108 epochs.

A.2. NAS-Bench-201

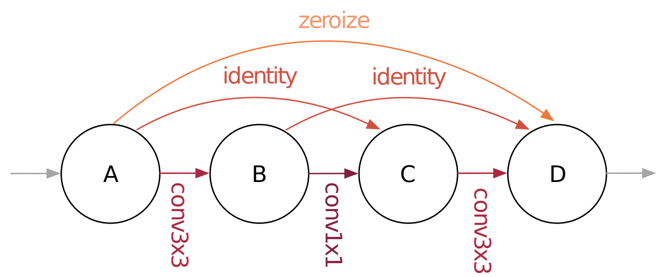
In NAS-Bench-201, networks also share a common skeleton (Figure 8c) that consists of stacks of its unique *cell* interleaved with fixed residual downsampling blocks. Each cell (Figure 8b) can be represented as a densely-connected DAG of 4 ordered nodes (A, B, C, D) where node A is the input and node D is the output. In this graph, there is an edge connecting each node to all subsequent nodes for a total of 6 edges, and each edge can perform one of 5 possible operations (Zeroise, Identity, 3×3 convolution, 1×1 convolution, 3×3 average pool). The search space consists of every possible cell. As there are 6 edges, on which there may be one of 5 operations, this means that there are $5^6 = 15,625$ possible cells. This makes for a total of 15,625 networks as each network uses just one of these cells repeatedly. The authors have manually split CIFAR-10, CIFAR-100, and ImageNet-16-120 (Chrabaszcz et al., 2017) into train/val/test, and provide full training results across all networks for (i) training on train, evaluation on val, and (ii) training on train/val, evaluation on test. The split sizes are 25k/25k/10k for CIFAR-10, 50k/5k/5k for CIFAR-100, and 151.7k/3k/3k for ImageNet-16-120.

A.3. NATS-Bench

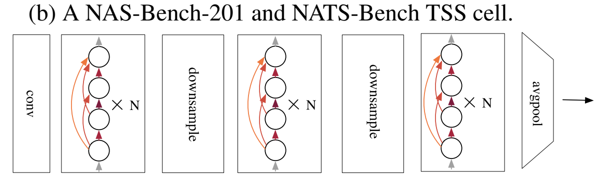
NATS-Bench (Dong et al., 2021) comprises two search spaces: a topology search space and a size search space. The networks in both spaces share a common skeleton which is the same as the skeleton used in NAS-Bench-201. The topology search space (NATS-Bench TSS) is the same as NAS-Bench-201 whereby networks vary by operation comprising the network cell. The size search space instead varies the channels of layers in 5 blocks of the skeleton architecture. Every network uses the same cell operations. The choice of operations corresponds to the best performing network in the topology search space with respect to the CIFAR-100 dataset. For each block the layer channel size is chosen from 8 possible sizes (8, 16, 24, 32, 40, 48, 56, 64). This leads to $8^5 = 32768$ networks in the size search space (NATS-Bench SSS).



(a) A NAS-Bench-101 cell.



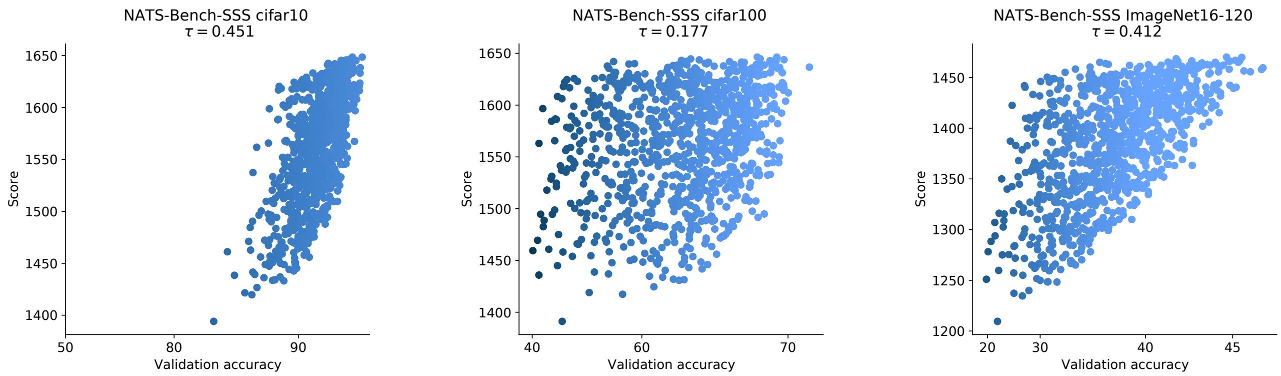
(b) A NAS-Bench-201 and NATS-Bench TSS cell.



(c) The skeleton for NAS-Bench-101 (N=3), 201 (N=5), and NATS-Bench TSS (N=5).

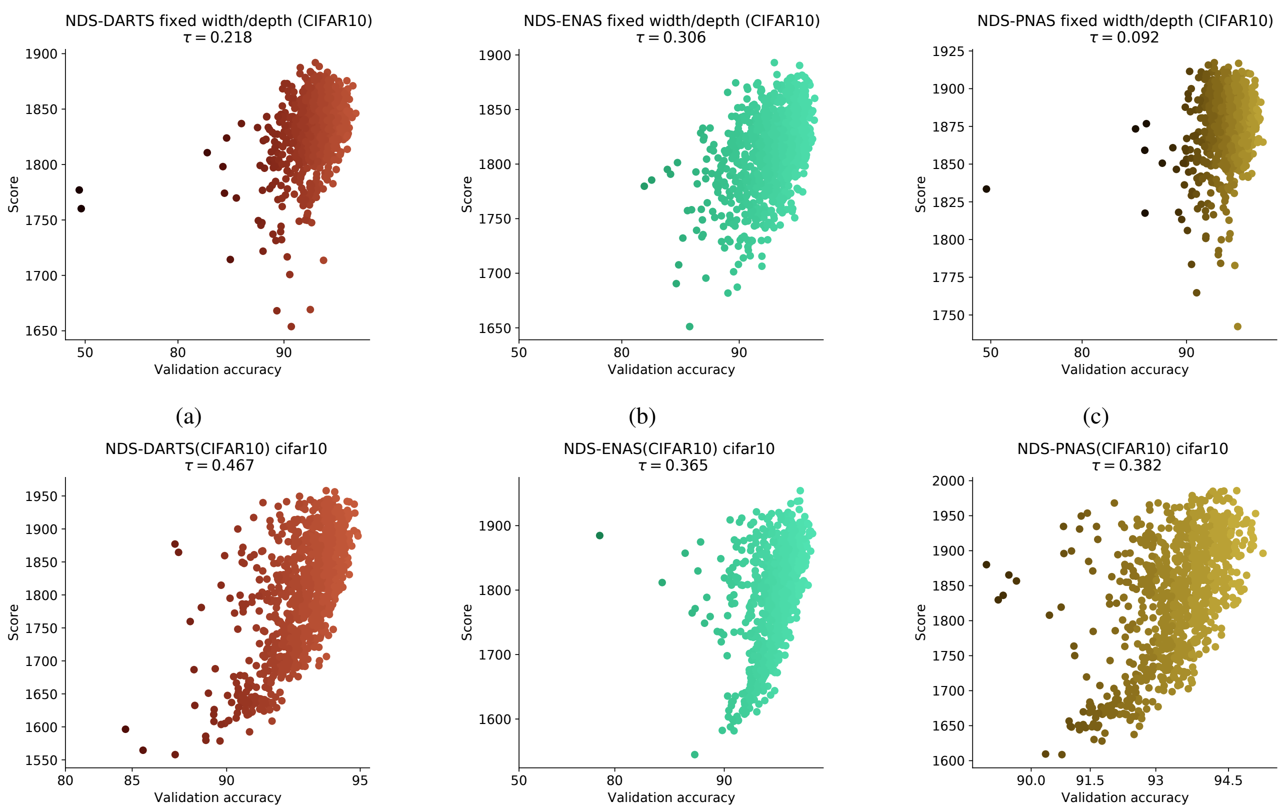
Figure 8. (a): An example cell from NAS-Bench-101, represented as a directed acyclic graph. The cell has an input node, an output node, and 5 intermediate nodes, each representing an operation and connected by edges. Cells can have at most 9 nodes and at most 7 edges. NAS-Bench-101 contains 426k possible cells. By contrast, (b) shows a NAS-Bench-201 (NATS-Bench TSS) cell, which uses nodes as intermediate states and edges as operations. The cell consists of an input node (A), two intermediate nodes (B, C) and an output node (D). An edge e.g. A→ B performs an operation on the state at A and adds it to the state at B. Note that there are 6 edges, and 5 possible operations allowed for each of these. This gives a total of 5^6 or 15,625 possible cells. (c): Each cell is the constituent building block in an otherwise-fixed network skeleton (where N=5). As such, NAS-Bench-201 contains 15,625 architectures.

B. Additional Plots



(a) (b) (c)

Figure 9. Further plots of our score (Equation 2) for 1000 randomly sampled **untrained** architectures in NATS-Bench SSS against validation accuracy when trained on (a) CIFAR-10, (b) CIFAR-100, and (c) ImageNet16-120.



(d) (e) (f)

Figure 10. Further plots of our score (Equation 2) for around 1000 randomly sampled **untrained** architectures in NDS-DARTS, NDS-ENAS, and NDS-PNAS against validation accuracy when trained. The top row shows the fixed width and depth variants of the search spaces, while the bottom row shows the variable width and depth spaces.