Entropic Score metric: Decoupling Topology and Size in Training-free NAS

Supplementary Material

Niccolò Cavagnero Luca Robbiano Francesca Pistilli Barbara Caputo

Giuseppe Averta

name.surname@polito.it

Politecnico di Torino, Corso Duca degli Abruzzi, 24 — 10129 Torino, ITALIA

Abstract

This supplementary material provides additional details on the specific design of ESFormer architectures.

1. ESFormer Architectures

Figure 1 shows the detailed architectures of ESFormer model family: S0 (left), S1 (center), S2 (right).

It is possible to notice how IBN [4] constitutes the main building block, especially in the first stages. ConvNeXt [2] is instead preferred in the last stages, with stage 4 usually dominated by this kind of block.

Similarly, another interesting pattern can be seen in the configuration choice for the FNN in Transformer Encoder layers, with Inverted Bottlenecks usually preferred in stage 3 and ConvNeXt blocks in stage 4.

With regard to kernel sizes, there is a tendency in the adoption of larger kernels as the depth of the network increases, especially in S0 and S1. This is the same arrangement explicitly devised by the authors of EdgeNeXt [3], where the kernel size is maintained small in the first stages to capture low-level features and it is gradually increased to deal with more high-level features.

Finally, we can appreciate how the QKV-dimension is always consistenly lower than the embedding dimension, except for S2 stage 3. Indeed, it is shown that adopting a smaller QKV-dimension enables a decrease in model complexity without affecting the performance [1].

Notably, while these behaviours are manually encoded in state-of-the-art architectures, they spontaneously emerge in our models during the search thanks to Entropic Score.

References

[1] Minghao Chen, Kan Wu, Bolin Ni, Houwen Peng, Bei Liu, Jianlong Fu, Hongyang Chao, and Haibin Ling. Searching the search space of vision transformer. *Advances in Neural Information Processing Systems*, 34:8714–8726, 2021. 1

[2] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, 2022. 1

[3] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *ECCV*, 2023. 1

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1

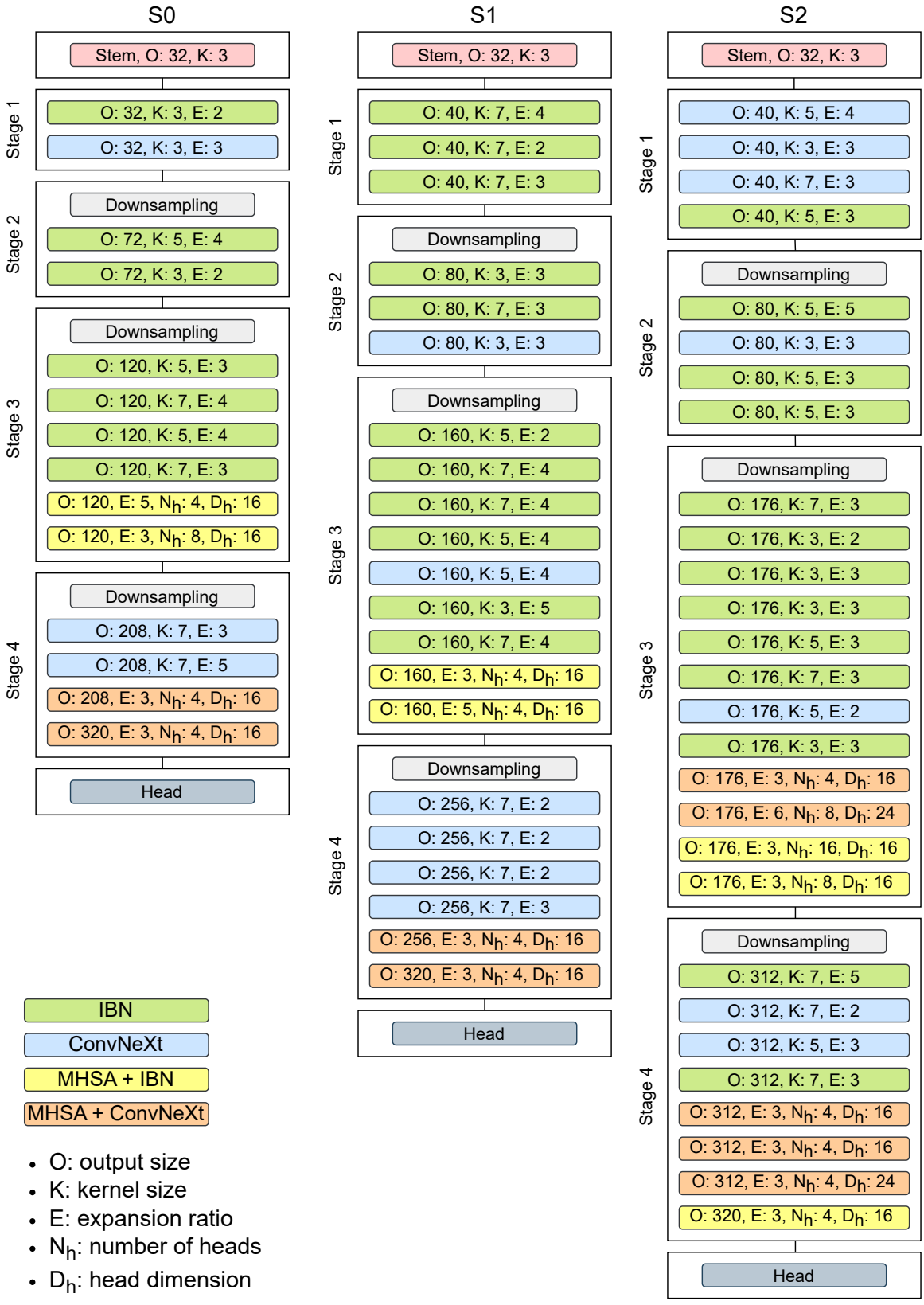

Figure 1: Detailed architectures for ESFormer-S0 (left), S1 (center), S2 (right).