Pruning techniques for artificial intelligence networks: a deeper look at their engineering design and bias: the first review of its kind

Lopamudra Mohanty[1,2] · Ashish Kumar[1] · Vivek Mehta[1] · Mohit Agarwal[1] ·
Jasjit S. Suri[3]

Abstract

Trained Artificial Intelligence (AI) models are challenging to install on edge devices as they are low in memory and computational power. Pruned AI (PAI) models are therefore needed with minimal degradation in performance. This is the first study of its kind where we have characterized three types of PAI paradigms, namely, weight pruning (WP), channel pruning (CP), and hybrid pruning (HP). Further, the study also analyses the risk of bias (RoB) in PAI models, and explainable PAI (XPAI) networks. PRISMA model was adapted for the selection of the best **145** studies. The review had two hypotheses: (i) PAI had bias and (ii) the order of performance for the three PAI is: HP > CP > WP in terms of performance. Thirty-two AI attributes using each of two experts were categorized into four clusters, namely, architecture, optimization, dataset design, and scientific validation. The RoB was analyzed in these PAI studies using AP(**ai**)Bias 2.0 (AtheroPoint™, CA, USA), which was based on a *scoring-and-ranking* strategy. This was then compared against the Butterfly model, composed of *radial* and *regional* biases, and analyzed using a Venn diagram in three independent bins based on low-moderate and moderate-high cut-offs. Recommendations were laid out for low-bias studies. PAI studies that lack adequate AI attributes were more inclined to the RoB. It has been observed that 19.64%, 21.42%, and 58.92% studies were in low-bias, moderate-bias, and high-bias bins, respectively. Pruning made inference faster and reduced sparsity. We have observed low participation of the PAI studies in the XPAI framework. The study presented the challenges and five-step recommendations for mitigating bias in PAI models.

Keywords Artificial Intelligence · Resource constraint devices · Training model sizes · Pruning · Quantization · Risk of bias · Explainability

✉ Jasjit S. Suri
jasjit.suri@atheropoint.com

1   School of Computer Science Engineering and Technology, Bennett University, Gr. Noida, India

2   Department of Computer Science, ABES Engineering College, Ghaziabad, UP, India

3   Stroke Diagnostic and Monitoring Division, AtheroPoint™, Roseville, CA, USA

🖄 Springer