Scaling Knowledge Processing from 2D Chips to 3D Brains

Kwabena Boahen

Stanford University, USA
E-mail address: boahen@stanford.edu

Artificial intelligence (AI) realizes a synaptocentric conception of the learning brain with dot-products and advances by performing twice as many multiplications every two months. But the semiconductor industry tiles twice as many multipliers on a chip only every two years. Moreover, the returns from tiling these multipliers ever more densely now diminish, because signals must travel relatively farther and farther, expending energy and exhausting heat that scales quadratically. As a result, communication is now much more expensive than computation. Much more so than in biological brains, where energy-use scales linearly rather than quadratically with neuron count. That allows an 86-billion-neuron human brain to use as little power as a single lightbulb (25W) rather than as much as the entire US (3TW). Hence, rescaling a chip's energy-use from quadratic to linear is critical to scale AI sustainably from 10^{12} parameters (mouse scale) today to 10^{15} parameters (human scale) in the next five years. But this would require communication cost to be reduced radically. Towards that end, I will present a recent re-conception of the brain's fundamental unit of computation that sparsifies signals by moving away from synaptocentric learning with dot-products to dendrocentric learning with sequence detectors.

References

1. K. Boahen, *Nature* 612, 43-5056 **(2022).**
2. H. Chen et al., IEDM 2023, 1-4 **(2023).**