FlyNet: Max it, Excite it, Quantize it
Luis Guerra¹, Tom Drummond²
¹Monash University, ²University of Melbourne

Comparison of FlyNet-h3 (3 MDW heads) architecture with different width multipliers against state-of-the-art models. FlyNet performs favorably in terms of parameter count against both popular and recently published works. Particularly, our method excels as the model shrinks, towards the extremely low parameter regime.

Why trading parameters with FLOPS?

We developed a series of simple but effective architectural modifications that can be integrated into any neural architecture to provide accuracy boosts at very little overhead. Particularly, our contributions are aimed at compressed networks in the extremely low parameter regime (sub 1M).

Energy Consumption of different processor operations