Overlapping Schwarz Attention

Oliver Rheinbach[1] Stephan Köhler[2]

We investigate a hierarchical attention mechanism for use in Transformer architectures, based on additive two-level overlapping Schwarz domain decomposition. The construction replaces a single global attention interaction operator by local attention blocks on overlapping subdomains and a coarse attention level for long-range information exchange,

$$\Phi Q_0 K_0^T \Phi^T + \sum_{i=1}^{N} R_i^T D_i^{1/2} Q_i K_i^T D_i^{1/2} R_i,$$

where R_i restricts to an overlapping subdomain, D_i defines partition-of-unity weights, and Φ is a coarse interpolation matrix. We test the method in a linear operator-learning setting and compare it with global softmax-free low-rank attention.

Experiments with synthetic Fourier right-hand sides indicate faster convergence and higher accuracy than the global baseline, despite substantially fewer parameters.

References:

[1] https://arxiv.org/abs/2606.18525

[1]Fakultät für Mathematik und Informatik
oliver.rheinbach@math.tu-freiberg.de

[2]Fakultät für Mathematik und Informatik
stephan.koehler@math.tu-freiberg.de