Increasing the Efficiency of the Transformer Architecture of ChatGPT During Inference Using Nanophotonics

Guo, Connie (School: James Clemens High School)

ChatGPT is a widely utilized resource with billions of users projected worldwide, simultaneously handling numerous tasks. However, its cost and efficiency have become a matter of concern due to its transformer architecture, which exhibits an inference time complexity of □(□^2) despite its ability to train on extensive datasets. To address these challenges, I propose employing the Receptance Weighted Key Value (RWKV) architecture—a recently introduced model that can be trained as a Transformer and used more efficiently like a recurrent neural network (RNN)—as an alternative, aiming to enhance efficiency and reduce expenses. Nevertheless, I discover that RWKV encounters the problem of vanishing gradients, which makes the training of RWKV sub-optimal. To overcome both vanishing and exploding gradients, I introduce a phase-coded parameterization in our algorithms inspired by nanophotonics. Furthermore, by leveraging optical processing units based on the nanophotonics computation modeled by phase-coded parameterizations, these text generative algorithms can achieve greater efficiency gains and cost reductions.