EFFICIENT QUANTIZATION FOR CPU-BASED DIFFUSION MODELS

A. L. Hanees[1] and E. Elango[2*]
[1]Department of Computer Science, Faculty of Applied Sciences
South Eastern University of Sri Lanka, Sri Lanka.
[2]Department of Computer Science, Government Arts College for Women,
Sivagangai, Tamilnadu, India
[*]alhanees@seu.ac.lk

The utilization of diffusion models to create visuals from textual descriptions has grown in popularity. However, the significant requirement for computing power still poses a significant obstacle and adds time to procedures. Diffusion models provide difficulties when quantization, a method used to reduce deep learning models for increased efficiency, is used. Comparing to other model types, these models are noticeably more susceptible to quantization, which could lead to deterioration in image quality. In this research, we present a unique method that uses distillation along with quantization-aware training to measure the diffusion models. Our findings demonstrate that quantized models can provide inference efficiency on CPUs while retaining great image quality. At https://github.com/intel/intel-extension-for-transformers, the source is accessible to the general public.

Keywords: *Quantization, Diffusion Models, U-Net Architecture*