Neural Acceleration for GPU Throughput Processors

Amir Yazdanbakhsh    Jongse Park    Hardik Sharma
Pejman Lotfi-Kamran    Hadi Esmaeilzadeh

Alternative Computing Technologies (ACT) Lab
Georgia Institute of Technology

NGPU
Neurally Accelerated GPU
Many GPU applications are amenable to approximation.
On average, more than 55% of runtime and energy is spent on neurally approximable regions.

CUDA Code

uchar4 p = tex2D(img, x, y);
dst[img.width * y + x] = z;
Accelerator [Session E1]

Tuesday at 3:00pm

2.4×  2.8×  ≥90%

Speedup  Energy Reduction  Quality

≤1%

Area Overhead

NGPU