Bedot: Bit Efficient Dot Product for Deep Generative Models

CoNGA 2023

Nhut-Minh Ho, Duy-Thanh Nguyen, John L. Gustafson, and Weng-Fai Wong

Overview

- ❏ Background
- ❏ Bedot design
- ❏ Algorithms for optimizing set entries
- ❏ Rounding hints
- ❏ Evaluation
- ❏ Sample results

Overview

- ❏ **Background**
- ❏ Bedot design
- ❏ Algorithms for optimizing set entries
- ❏ Rounding hints
- ❏ Evaluation
- ❏ Sample results

Generative models

Generative Pre-trained Transformer 3 (GPT-3)

Generative Adversarial Networks (GAN)

Style transfer, image super resolution

- ❑ Diverse architectures
- ❑ Practical applications
- ❑ Sensitive to input output error and user perception?


Forbes

Microsoft Confirms Its $10 Billion Investment Into ChatGPT, Changing How Microsoft Competes With Google, Apple And Other Tech Giants


'Google killer' ChatGPT sparks AI chatbot race

Original
(W:H)

Low quality image
(W/4 : H/4)

ESRGAN reconstruct X4
(W:H)

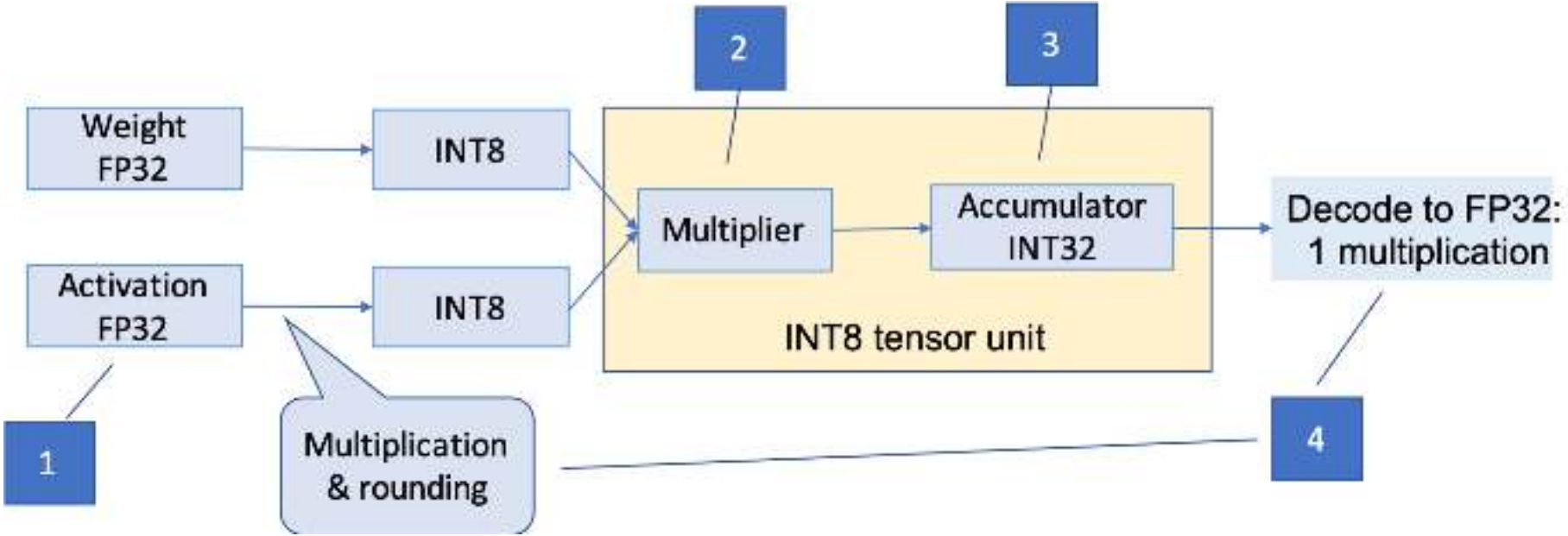
CycleGAN Monet Style

Sentence hint :
"**A robot was created** …."

GPT-2 text completion

"**A robot was created** by the University of Washington's Center for Robotics and Intelligent Systems (CRIS) to help with the construction of a new bridge..."

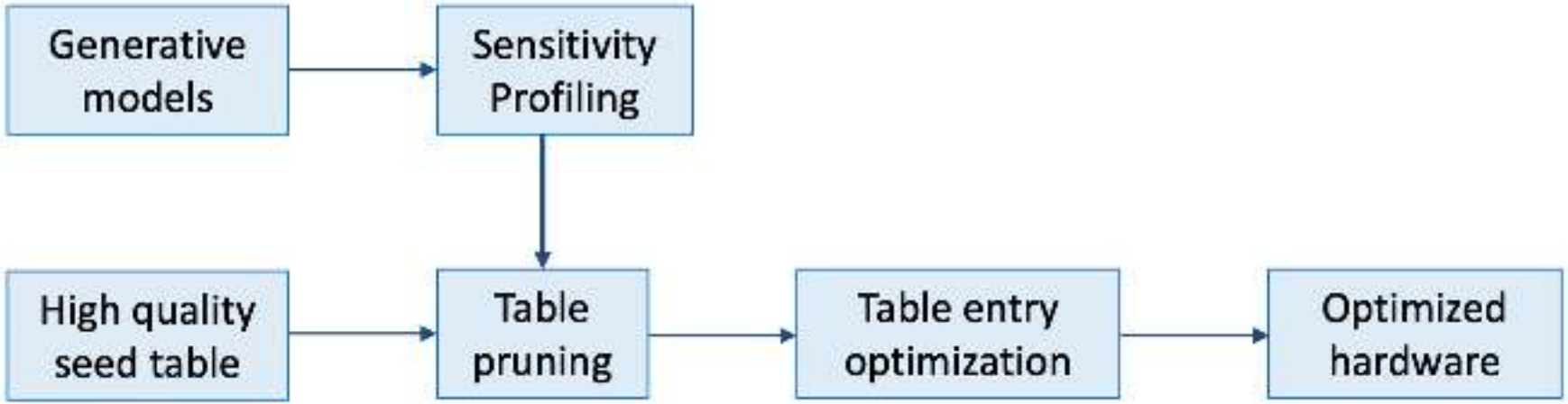
Quantization



Overview

- ❏ Background
- ❏ **Bedot**
- ❏ Algorithms for optimizing set entries
- ❏ Rounding hints
- ❏ Evaluation
- ❏ Sample results

Bedot


Generative models
Sensitivity Profiling
High quality seed table
Table pruning
Table entry optimization
Optimized hardware

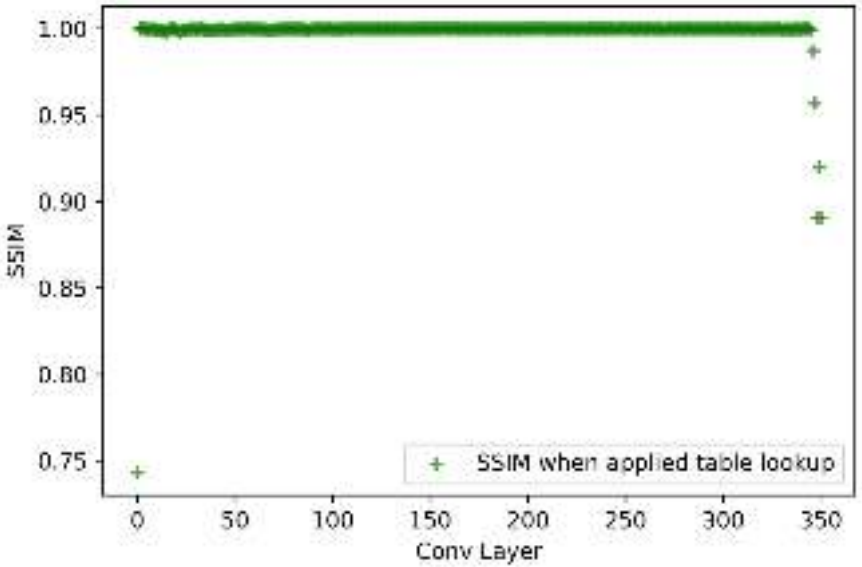
Sensitivity analysis

Table lookup for 345 middle layers
SSIM : **0.98**

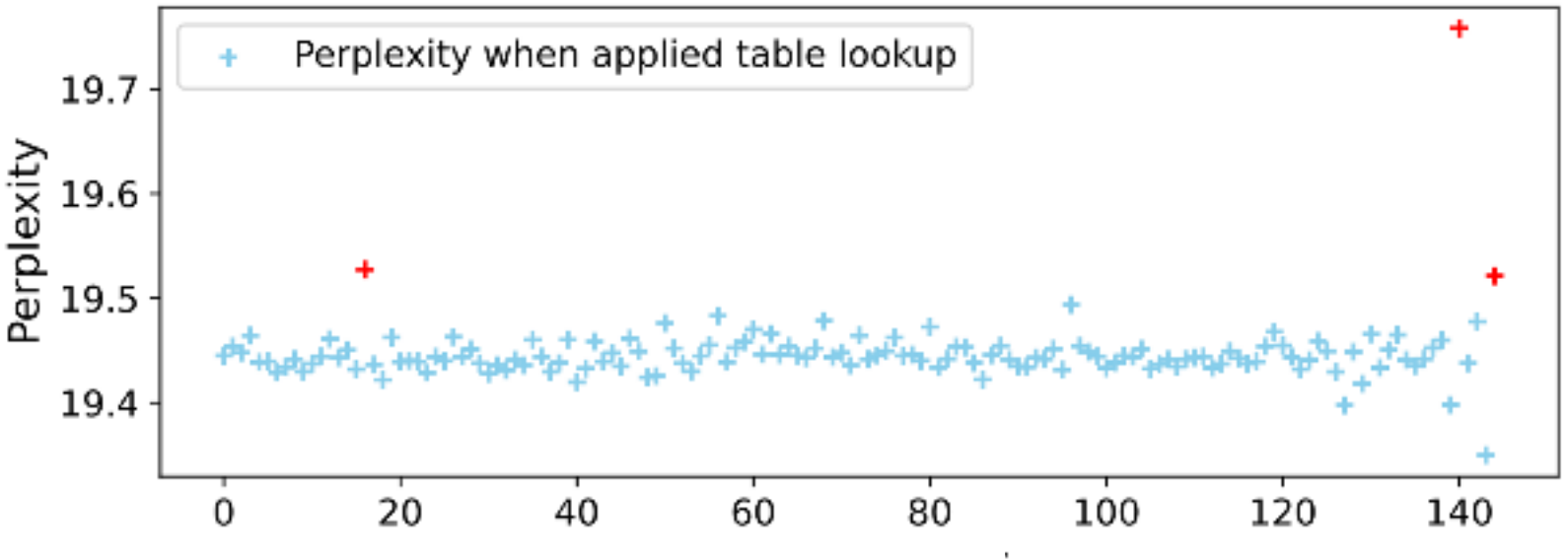
Table lookup for 345 + first layer
SSIM : **0.743**

Table lookup for all 351 layers
SSIM : **0.648**

Table lookup for 345 + last
SSIM : **0.878**



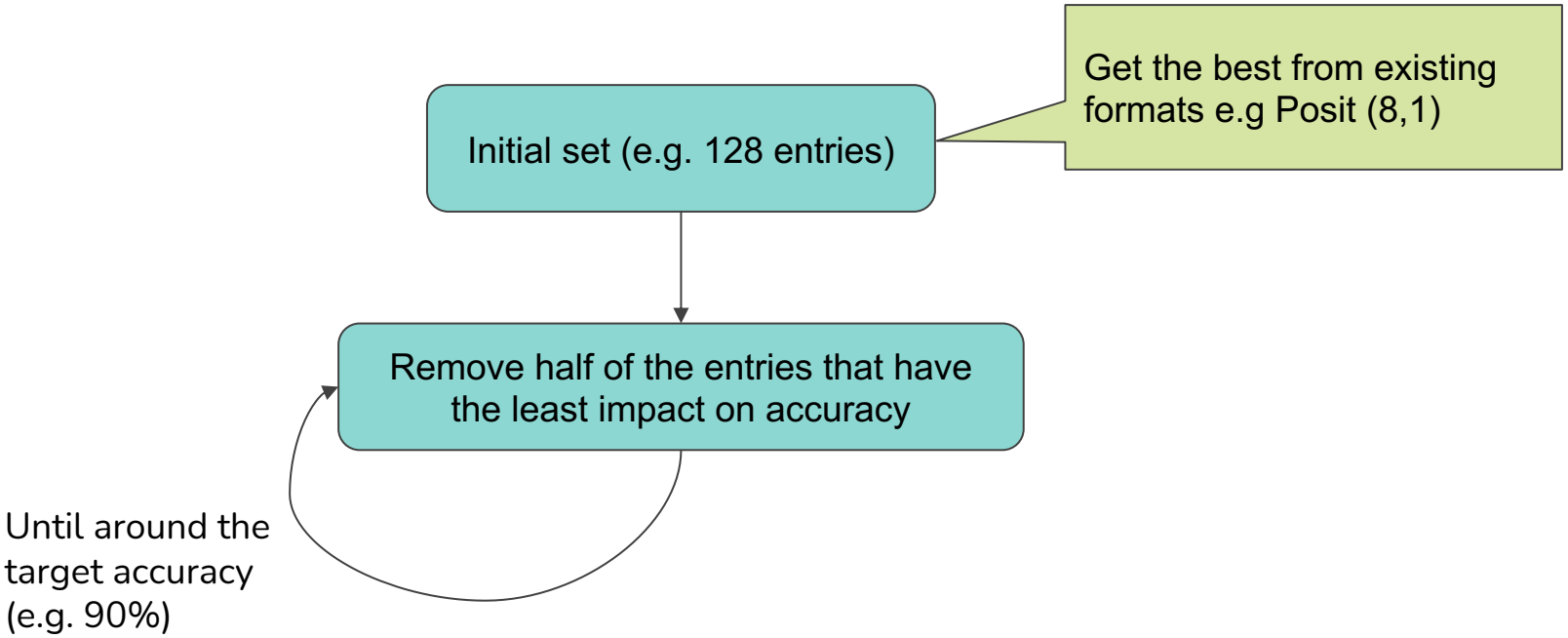
Sensitivity analysis (cont)



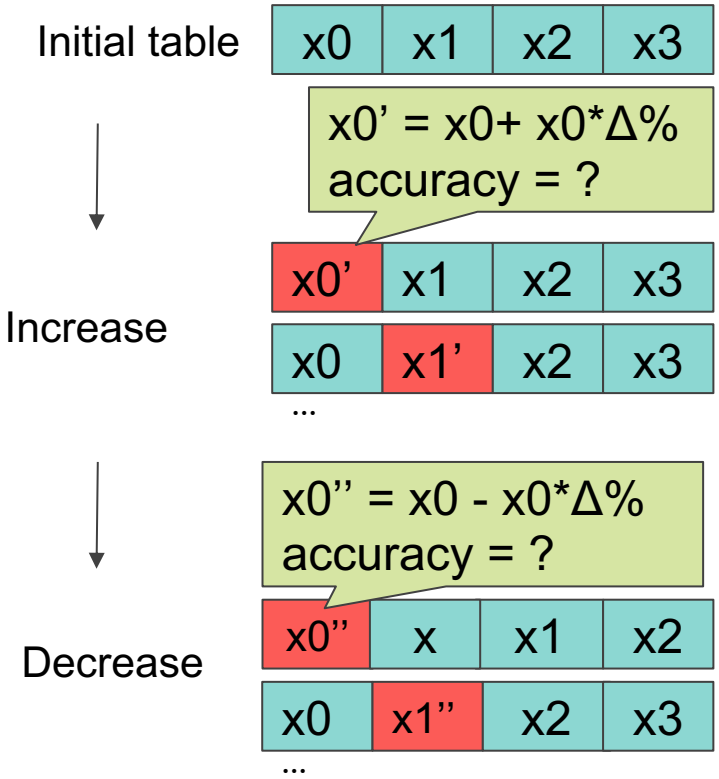
Overview

- ❏ Background
- ❏ Bedot design
- ❏ **Optimizing set entries**
- ❏ Rounding hints
- ❏ Evaluation
- ❏ Sample results

Reducing the number of entries

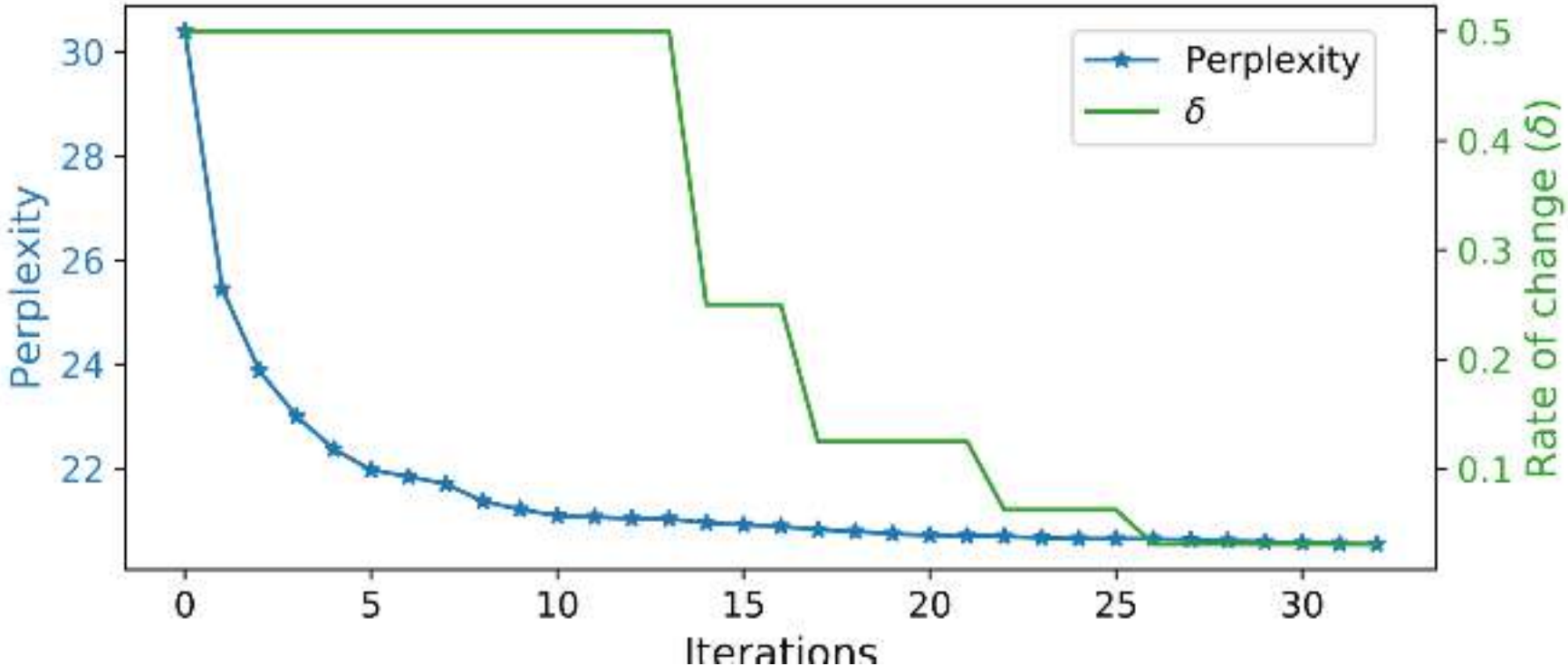


Optimizing set entries



- ❏ Increase each entry Δ% in turn, record the accuracies
- ❏ Decrease each entry Δ% in turn, record the accuracies
- ❏ Find the **best** change that **increases accuracy** the most => apply the change
- ❏ If no improvement, reduce Δ = Δ/2
- ❏ Repeat until converged (Δ<1%)

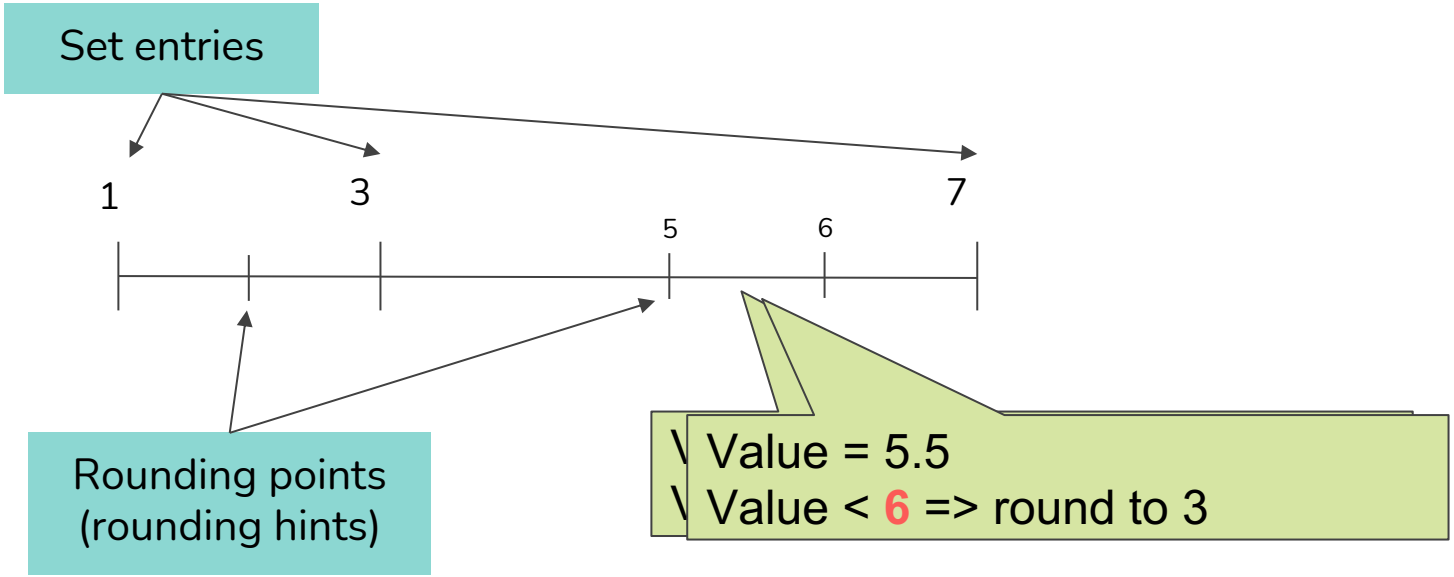
Optimizing set entries



Overview

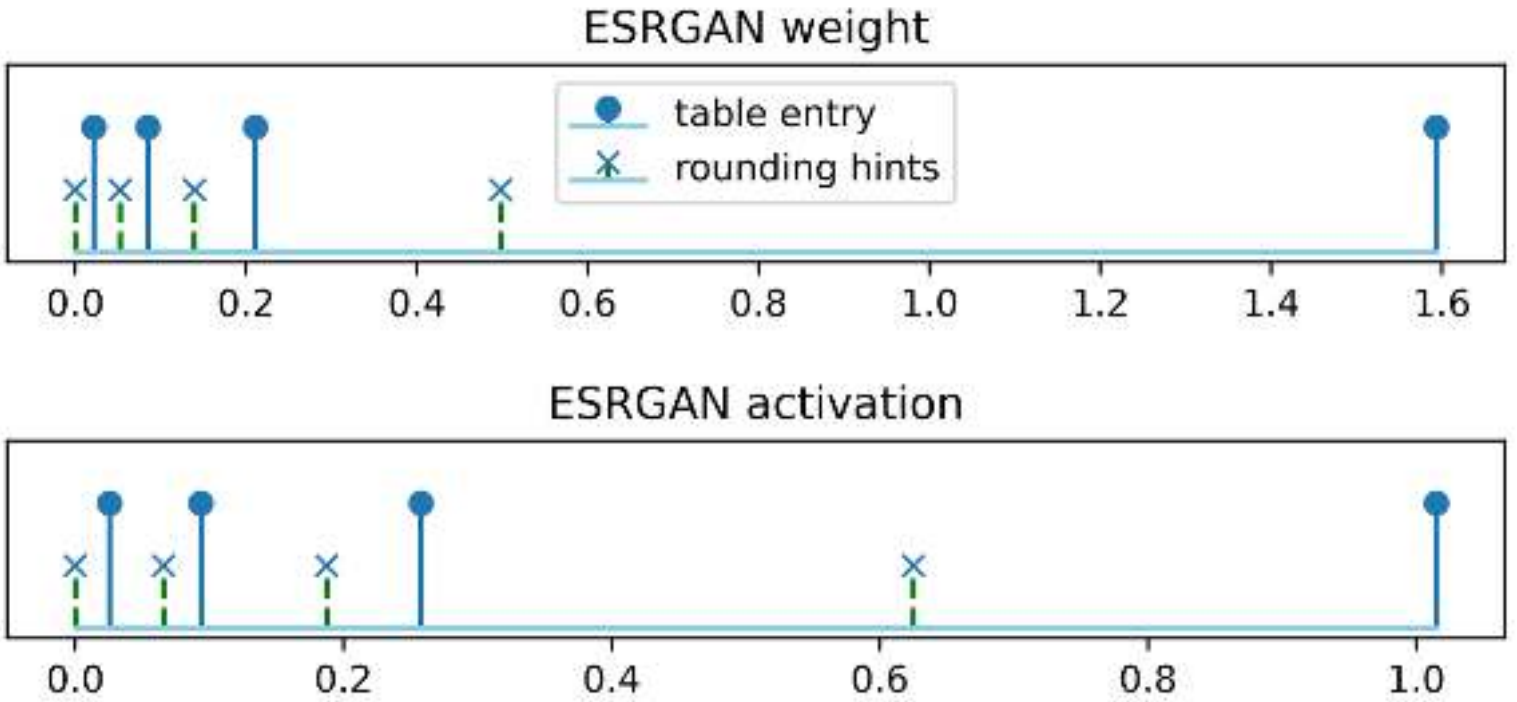
- ❏ Background
- ❏ Bedot design
- ❏ Algorithms for optimizing set entries
- ❏ **Rounding hints**
- ❏ Evaluation
- ❏ Sample results

Round to nearest and rounding hints



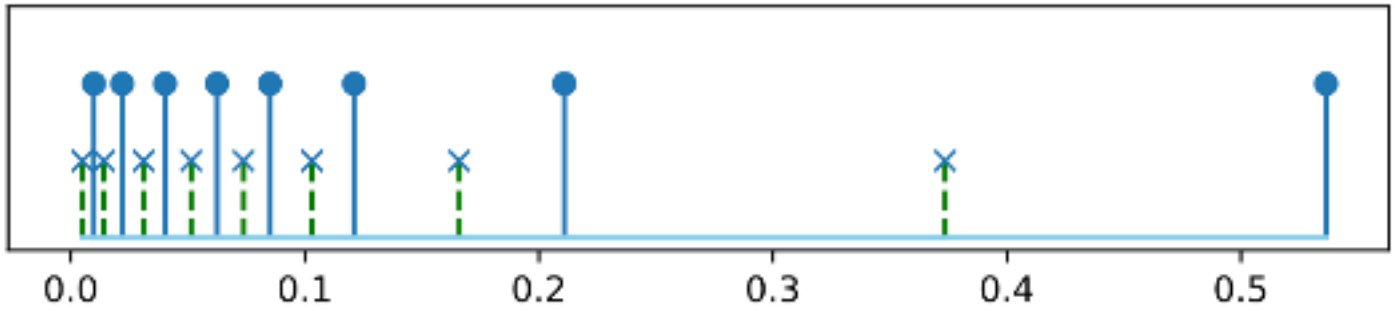
- **Rounding hints** are also **tunable** with our algorithm
- Let them move around and find the **best** location

Effect of tuning the mid points

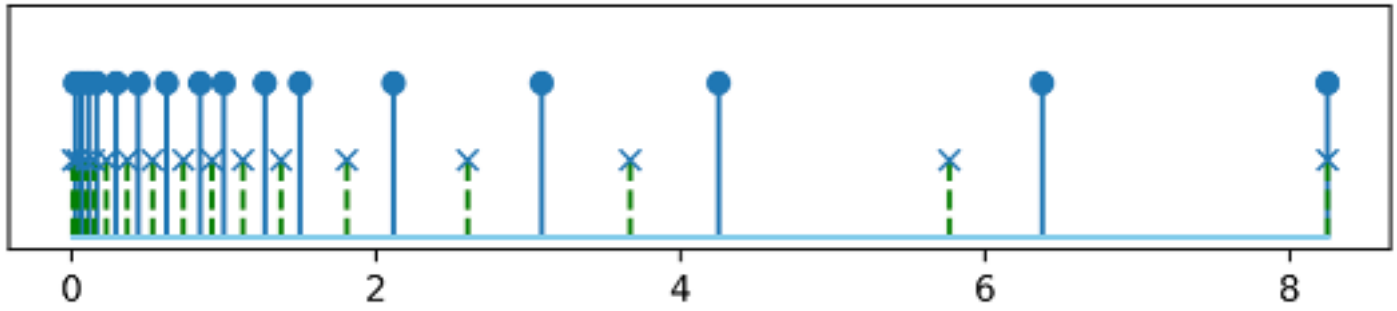


Effect of tuning the mid points

GPT2 weight



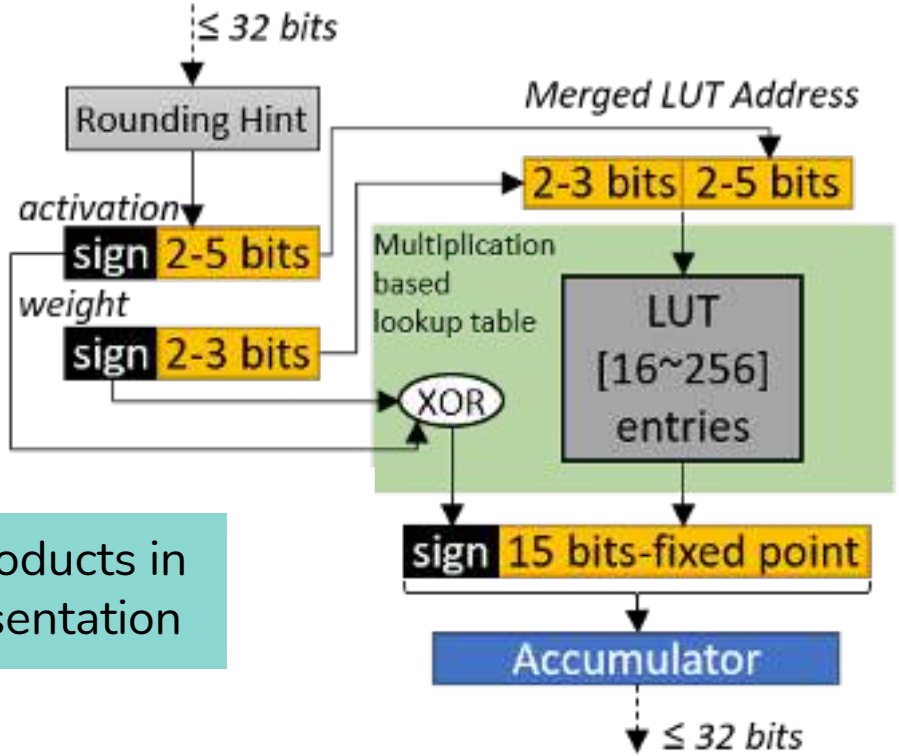
GPT2 activation



Overview

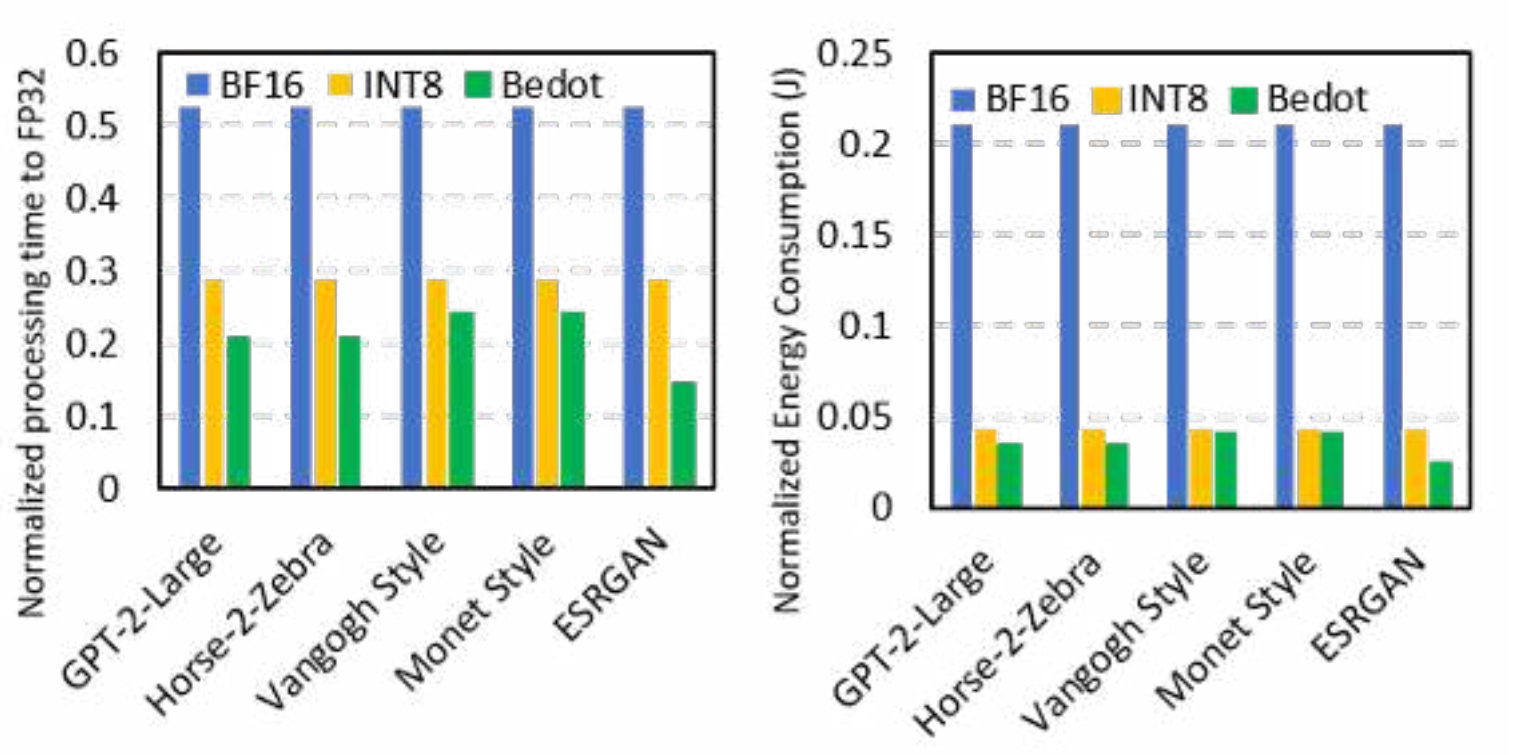
- ❏ Background
- ❏ Bedot design
- ❏ Algorithms for optimizing set entries
- ❏ Rounding hints
- ❏ **Evaluation**
- ❏ Sample results

MAC Unit



Lookup for the products in fixed point representation

Runtime + Energy



Overview

- ❏ Background
- ❏ Bedot design
- ❏ Algorithms for optimizing set entries
- ❏ Rounding hints
- ❏ Evaluation
- ❏ **Sample results**

Output quality

- **2-3** bit for weight, **2-5** bit for activation
- **0.94-0.95** SSIM for image tasks compared to **1.0** in FP32 (higher better)
- **20.435** perplexity for GPT-2 compared to **19.1** in original (lower better)

Hard to perceive how good is the result

Original

Style VanGogh
Reference FP32

Table lookup
3-bit weights, 5-bit acts
0.94 SSIM

Original

Style Monet
Reference FP32

Table lookup
3-bit weights, 5-bit acts
0.95 SSIM

2-bit lookup + rounding hints

Input FP32 Bedot

Text generation with GPT-2

Bedot & FP32

In the 19th century, the invention of the automobile was a major event. The automobile was a new form of transportation that was a revolution in the way people traveled. The automobile was a revolution in the way people traveled…

In the 19th century, the invention of the automobile and the development of the internal combustion engine led to a dramatic increase in the number of people who could afford to own a car. The automobile was a symbol of the American dream, and the automobile was…

A robot was created by the University of Washington's Center for Robotics and Intelligent Systems (CRIS) to help with the construction of a new bridge. The bridge is being built by the Seattle Department of Transportation (SDOT) and the Seattle…

A robot was created to help the elderly in a hospital in the Philippines. The robot, called "A-Bot," is able to walk, talk, and **even eat**. The robot was created by a team of researchers from the…

Conclusion & Demo Code

Lookup MAC with 2-4 bit is feasible for generative models, producing good quality output

Enabled by Qtorch+: Source code + Google Colab: https://github.com/minhhn2910/QPyTorch

- ❏ Posit
- ❏ Simulate new number formats
- ❏ Arbitrary number set + rounding hints

Thank you

Q & A