Mokey

Enabling Narrow Fixed-Point Inference for Out-of-the-Box Floating-Point Transformer Models

49th IEEE/ACM International Symposium on Computer Architecture (ISCA '22)

Ali Hadi Zadeh\textsuperscript{1,2}, Mostafa Mahmoud\textsuperscript{1}, Ameer Abdelhadi\textsuperscript{1}, and Andreas Moshovos\textsuperscript{1,2}

\textsuperscript{1}University of Toronto, \textsuperscript{2}Vector Institute
Transformers for Text Generation

My name is Ali. I am a PhD candidate at UofT. Today I am presenting my paper at the 49th IEEE/ACM International Symposium on Computer Architecture (ISCA '22).
My name is Ali. I am a PhD candidate at UofT. Today I am presenting my paper at the 49th IEEE/ACM International Symposium on Computer Architecture (ISCA '22). To tell you the truth, I'm nervous, but excited to show off my ideas!
Challenges

Weights

2018
1.2GB

BERT

Activations

2021
2TB

MT-NLG

A: 5%

A: 80%
Challenges

Weights

Activations

Memory: Performance & Energy Bottleneck

A: 80%
Challenges

Weights

- **2018**
  - 1.2GB

- **2021**
  - 2TB

Activations

- **A: 5%**
  - BERT

- **A: 80%**
  - MT-NLG

FP Compute

- 0.73
- 0.0023
- 0.02
- -0.054
- ~100T FP MACs
Mokey: BERT’s Better Self

Floating point

4b Int index
8x vs FP32
4x vs FP16

*Not your typical 4b quantization ;}
Mokey

4-bit Quantization: W+A

\[ W, A = f(idx) \]

\[ += A \times W. \Rightarrow \text{Count } idx \]

Post-training

Fixed-point compute
Mokey HW Accelerator

Vs. Tensor Cores: **15x** Faster + **100x** Energy Efficient

Mokey Memory Compression

For Tensor Cores:
- Off-chip Only: **4x** Faster + **8x** Energy Efficient
- Off- and on-chip*: **10x** Faster + **50x** Energy Efficient
Roadmap

Mokey Quantization
Computation on indices
Models’ Accuracy
Hardware Evaluation
Conclusion
Dictionary-Based Quantization

![Graph showing distribution of values with occurrence on the y-axis and values on the x-axis. The graph has a peak at zero and decreases symmetrically on either side.]
Dictionary-Based Quantization

Values

Occurrence

Index | Value
--- | ---
I | 0.02
II | 0.07
III | 0.12
IV | 0.25
... | ...
Dictionary-Based Quantization

Clustering: Iterative ☹
Not Feasible for Activations
A Dictionary for All Layers

### Reference Distribution

- Mean ($\mu$) = 0
- SD ($s$) = 1

| Index | Value |
|-------|-------|
| I     | -2.7  |
| II    | -1.98 |
| ...   | ...   |
| VI    | 1.98  |
| XVI   | 2.7   |

Golden Dict. (GD)

Scale and Shift is All You Need!
Weights & Embeddings: Offline
Activation: Profiling

Scale and Shift is All You Need!
Inference Computation

**Original**

\[ A = 0.2 \quad W = 0.7 \]

\[ A \times W += 0.2 \times 0.7 = 0.14 \]

**Dictionary Quant.**

\[ A = I \quad W = II \]

\[
\begin{array}{|c|c|}
\hline
\text{Index} & \text{Value} \\
\hline
I & 0.2 \\
II & 0.7 \\
III & 1.1 \\
IV & 1.4 \\
\ldots & \ldots \\
\hline
\end{array}
\]

\[ A \times W += I \times II = 0.14 \]

**Mokey Quant.**

\[ A = I \quad W = II \]

\[ A \times W += I \times II = 0.14 \]
**How to Use Indices for Computation**

| Index | Value |
|-------|-------|
| I     | 0.2   |
| II    | 0.7   |
| III   | 1.1   |
| IV    | 1.4   |
| ...   | ...   |

\[
A \times W += I \times II \\
A \times W += 0.2 \times 0.7 = 0.14
\]

---

**Let’s See in Practice!**

Original | Dictionary Quant. | Mokey Quant.
FP16 Baseline

Original

\[ A = 0.2 \quad W = 0.7 \]

\[ A \times W = 0.2 \times 0.7 = 0.14 \]
Dictionary Quantization

| Index | Value |
|-------|-------|
| I     | 0.2   |
| II    | 0.7   |
| III   | 1.1   |
| IV    | 1.4   |
| …     | …     |

\[ A \times W += I \times II \]
\[ A \times W += 0.2 \times 0.7 = 0.14 \]

Chip

On chip Buffers (FP16)

Compute Units

DRAM (4-bit) 4x Compression

4bit
Dictionary Quantization

No on chip mem comp
Limited performance/energy gain

\[ A \times W += I \times II \]
\[ A \times W += 0.2 \times 0.7 = 0.14 \]
Dictionary Quantization

| Index | Value |
|-------|-------|
| I     | 0.2   |
| II    | 0.7   |
| III   | 1.1   |
| IV    | 1.4   |
| ...   | ...   |

\[ A = I \quad W = II \]

\[ A \times W += I \times II \]

\[ A \times W += 0.2 \times 0.7 = 0.14 \]
Dictionary Quantization

Many instances of LUT
Waste of Area & Energy 😞

\[ A = I, \quad W = II \]

\[ A \times W = 0.2 \times 0.7 = 0.14 \]
Mokey Quantization

“Simple” Relationship between Index and Value
Mokey Quantization
Symmetrical Dictionary

Golden Dict. (GD)

| Index | Value |
|-------|-------|
| I     | 0.05  |
| II    | 0.35  |
| …     | …     |
| VI    | 1.97  |
| VII   | 2.61  |
Exponential Function

Golden Dict. (GD)

| Index | Value |
|-------|-------|
| I     | 0.05  |
| II    | 0.35  |
| ...   | ...   |
| VI    | 1.97  |
| VII   | 2.6   |

GD = \alpha^i + b
Values Format

| Index | Value |
|-------|-------|
| I     | 0.05  |
| II    | 0.35  |
| ...   | ...   |
| VI    | 1.97  |
| VII   | 2.6   |

Golden Dict. (GD)

Values = ±(a_i + b)

per value Fixed
**Values Format**

Golden Dict. (GD)

| Index | Value |
|-------|-------|
| I     | 0.05  |
| II    | 0.35  |
| ...   | ...   |
| VI    | 1.97  |
| VII   | 2.6   |

Values = \( \pm (a_i + b) \times s + \mu \)

- **W**
  - \( \pm \)
- **A**
  - \( \pm \)

**per value**  
**Fixed**  
**Per layer**
Revisiting Computation

\[ A_0 = a_1^1 + b \quad W_0 = a_2^2 + b \]
\[ A_1 = a_4^4 + b \quad W_1 = a_5^5 + b \]

... \quad ...

Revisiting Computation

\[ \sum_{N} AW = A_0 W_0 + A_1 W_1 + \ldots \]

\[ = \alpha^1 + \beta + \alpha^4 + \beta + \ldots \]

\[ + \alpha^2 + \beta + \alpha^5 + \beta + \ldots + \]

Pre-computed Range [0,14]

Histogram

(2) Weighted reduction

Computed during last layer’s quantization

\[ A_0 = \alpha^{\text{skew}} + \beta \]
\[ W_0 = \alpha^{\text{bias}} + \beta \]

\[ A_1 = \alpha^{\text{skew}} + \beta \]
\[ W_1 = \alpha^{\text{bias}} + \beta \]
Revisiting Computation

\[ \sum_{N} AW = A_0W_0 + A_1W_1 + \ldots \]

\[ = a^{1+2} + ba^{1} + ba^{2} + b^{2} + \]

\[ + a^{4+5} + ba^{4} + ba^{5} + b^{2} + \ldots \]

\[ = (a^{3} + a^{9} + \ldots) + b(a^{1} + a^{4} + \ldots) + b(a^{2} + a^{3} + \ldots) + Nb^{2} \]

Pre-computed

Range [0, 7]
Revisiting Computation

\[ \Sigma_N AW = A_0W_0 + A_1W_1 + \ldots \]

\[ = a_1^2 + b + ba_1 + ba_2 + b^2 + ba_4 + ba_5 + b^2 + \ldots \]

\[ = (a_3 + a_9 + \ldots) + b(a_1^2 + a_4^2 + \ldots) + b(a_2^2 + a_5^2 + \ldots) + Nb^2 \]

Range \([0, 7]\)

Comprehension during last layer’s quantization

Pre-computed
Revisiting Computation

\[ \sum_N A W = A_0 W_0 + A_1 W_1 + \ldots \]

\[ = a^1 + b + ba^1 + b^2 + b^2 + ba^4 + ba^5 + b^2 + \ldots \]

\[ = (a^3 + a^9 + \ldots) + b(a^1 + a^4 + \ldots) + b(a^2 + a^5 + \ldots) + N b^2 \]

- **Range** [0,7]+[0,7]=[0,14]
- **(1) Histogram**
- **(2) Weighted reduction**
- **Computed during last layer’s quantization**
- **Pre-computed**
Revisiting Computation

Histogram: Most of the computation => 3-bit INT Add
Reduction: Post processing => 16-bit Fixed-point

\[ \sum A W = A_0 W_0 + A_1 W_1 + \ldots = 1 + 2 + 1 + 2 + 2 + 4 + 5 + 4 + 5 + 2 + \ldots = (a^3 + a^9 + \ldots) + b(a^1 + a^4 + \ldots) + b(a^2 + a^5 + \ldots) + N b^2 \]

Range \([0,7] + [0,7] = [0,14]\)
One Size Fits All

Approach
Evaluation

- FP16 Tensor Cores baseline
- Wide range of on-chip buffers
- 110M - 750M parameter models

- Custom cycle accurate simulator.
  - DRAMsim3: Dual Channel DDR4-3200
- On-chip Memory: CACTI
- Synthesis: Synopsis Design Compiler
  - 65nm TSMC – 1Ghz
- Layout: Cadence Innovus
- Signal Activity: Modelsim
- Power Estimation: Cadence Innovus
Quantization Accuracy

Score

FP Mokey

| Models/Tasks | MNLI (m ACC) | SQuAD (F1) |
|--------------|--------------|-------------|
| BERT-Base    |              |             |
| BERT-L       |              |             |
| RoBERTa-L    |              |             |
| DeBERTa-XL   |              |             |
| BERT-L       |              |             |
| RoBERTa-L    |              |             |
Accelerator Performance

Compression: 16-bit to 4-bit => 4x
Accelerator Performance

Compression: 16-bit to 4-bit => 4x

Baseline can almost fit activations on-chip
Accelerator Performance

**Compression:** 16-bit to 4-bit => 4x

![Graph showing speedup for different models with longer sequence length](image)
Accelerator Energy Efficiency

Memory Compression and more in paper 😊
Memory Compression
Conclusion

- Mokey Quantization:
  - 4-bit Dictionary-based
  - Focus on a subspace => closed-form representation
  - Fixed-point compute
  - No fine-tuning

- Mokey HW Accelerator:
  - Compute directly on indices
  - 1.6x Smaller tiles vs. Tensor Cores
  - 15x Faster and 100x Energy efficient
  - Can be adapted to other accelerators

Ali Hadi Zadeh
hadizade@ece.utoronto.ca