# CMSC828G: Special Topics in Systems for Machine Learning

# KV Cache Compression I

Instructor: Yafan Huang




DEPARTMENT OF
COMPUTER SCIENCE

# When you are compressing in a lossy manner…

- What to compress?
  - K cache, V cache, Pre-/after- RoPE.
- How to compress?
  - Sparsity, token eviction, attention approximation works. *But not our focus here.*
  - Quantization? Huffman? Run-length? Fixed-length? Delta? Min+difference?
- How much accuracy do we lose / can we accept?
  - Very task-specific.
- System integration.

While other compression methods work in current literature, we mainly talk about quantization for KV cache compression in this lecture.

# Quantization Error in Attention

- Attention computation: $X' = AV = \mathrm{softmax}(QK^T/\sqrt{d_k} + M)V$
- Quantizing keys: Perturbs attention scores, then attention weights.
- Quantizing values: Perturbs the weighted output directly.
- Hard to quantify which is more vulnerable to loss.

# Quantization Granularity: Token or Channel

| | | | | |
|---|---|---|---|---|
| Token1 | 10 | 0.1 | 0.2 | 0.3 |
| Token2 | 11 | 0.2 | 0.3 | 0.1 |
| Token3 | 9 | 0.2 | 0.2 | 0.2 |
| Token4 | 10 | 0.3 | 0.1 | 0.2 |

Per-token Grouping

| | | | | |
|---|---|---|---|---|
| Token1 | 10 | 0.1 | 0.2 | 0.3 |
| Token2 | 11 | 0.2 | 0.3 | 0.1 |
| Token3 | 9 | 0.2 | 0.2 | 0.2 |
| Token4 | 10 | 0.3 | 0.1 | 0.2 |

Per-channel Grouping

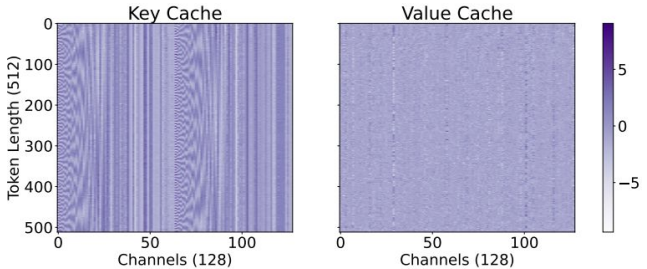



Fig. 3. Visualization of the Key cache and Value cache in the Llama 3.1 8B Instruct model for an input sequence of 512 tokens.

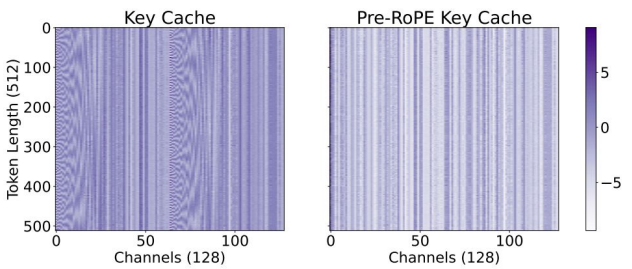



Fig. 4. Comparison of the key cache before and after applying RoPE.

# Quantization of a Streaming KV Cache

- Per-token quantization: A new token can be quantized independently.
- Per-channel quantization: A group may require values from multiple tokens.
    - Residual buffering: Keep incomplete groups in higher precision until they can be processed within a group.
    - Example: group size = 4; while decode loop has 6 tokens already, the first 4 can form a group and last 2 stays in residual buffer.

# Compression Ratio and Memory Overhead

- Ideal compression ratio: FP16 -> INT2: 16/2 = 8
- Practical memory overhead:
  - Quantization metadata (e.g., scales and zero-points).
  - Uncompressed data (e.g., residual tokens or outliers).
- Effective compression ratio depends on the actual storage required, not just quantization bandwidth.
- Example: 90% INT2 + 10% FP16 (assuming no metadata in this case)

$$b_{avg} = 0.9 \times 2 + 0.1 \times 16 = 3.4$$

$$R_{effective} = 16/3.4 \approx 4.71$$

# Beyond Uniform Quantization

| Target | Possible Solution |
| --- | --- |
| Outlier channels | Per-channel quantization |
| Extreme individual values | Sparse outlier representation (bit mask) |
| Different numerical distributions | Non-uniform quantization |
| RoPE changes Key distributions | Delta / Min sub. + Fixed-length in Pre-RoPE |
| Different layer sensitivities | Sensitivity-aware quantization |

# Student Presentations

- [ICML'24] KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache
- [NeurIPS'24] KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization