BREAKING THE BINARY:BitNet b1.58

Discussion and Application of 1.58bit LLMs

LLM PERFORMANCE COMES AT A COST

130GB
Llama-2

326GB
GPT-3

DUE TO ITS LARGE SIZE

Slow Response Times

Large Memory Usage

High Energy Consumption

QUANTIZATION

is a technique to optimize deep learning models by reducing their computational demand.

	No Quantization	GGML/GGUF	GPTQ	EXL2
Model Size	16.07GB	4.92GB	5.74GB	5.21GB
Inference Time	23.602 seconds	32.274 seconds	5.7416 seconds	2.530 seconds
Result	I am a Filipino and I am proud of my heritage. I believe that the Philippines is a beautiful country with a rich culture and history. I am proud of our national heroes, our traditions and our people	I am a Filipino and I am proud to be one. I am proud of our rich culture, our beautiful language, our delicious food, and our warm hospitality. I am proud of our history, our heroes, and our struggles. I am	I am a Filipino and/or ulus FiorEqualityComp arereturtlehevoa dycastle herself蜗 Gioholdiram族自治 lite Franklinlow ? reyavouritesjonkt opiлруги93y mebrig ourselves cinsFormatExcept ionoonardy#a donomyeping Bry シ _TScasecmpwickA TRIXレス Trib	I am a Filipino and my native language is Tagalog. I love to write and share my thoughts and experiences with others. I also enjoy reading books, watching movies, and learning new things. I am a devout Catholic and I strive to live my faith every day. I am a member of the Filipino community here in the United States and I am proud to be part of it. I hope that through this blog, I can share my culture, my faith, and my experiences with others and learn from them as well. I am excited to start this journey and I hope you will join me along the way. Mabuhay! (Long live!)

CAN WE PUSH THE BOUNDARIES OF LLM QUANTIZATION FURTHER TO ITS EDGE?

BitNet: Scaling 1-bit Transformers for Large Language Models

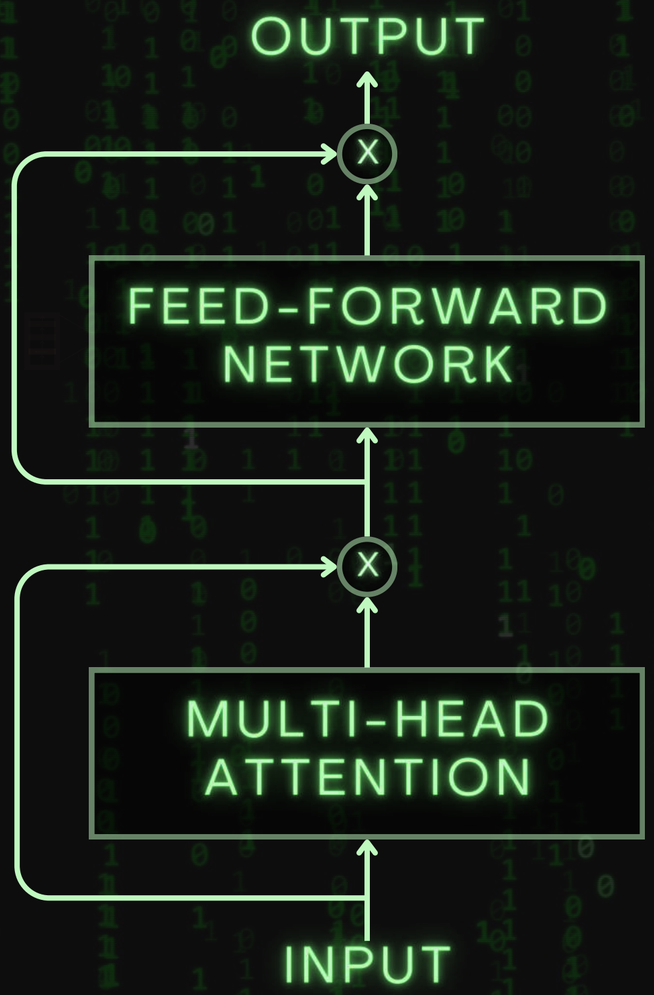
Hongyu Wang[*†‡] **Shuming Ma**[*†] **Li Dong**[†] **Shaohan Huang**[†]
Huaijie Wang[§] **Lingxiao Ma**[†] **Fan Yang**[†] **Ruiping Wang**[‡] **Yi Wu**[§] **Furu Wei**[†◇]
[†] Microsoft Research [‡] University of Chinese Academy of Sciences [§] Tsinghua University
https://aka.ms/GeneralAI

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., Yang, F., Wang, R., Wu, Y., & Wei, F. (2023). BitNet: Scaling 1-bit Transformers for Large Language Models. arXiv. https://arxiv.org/abs/2310.11453

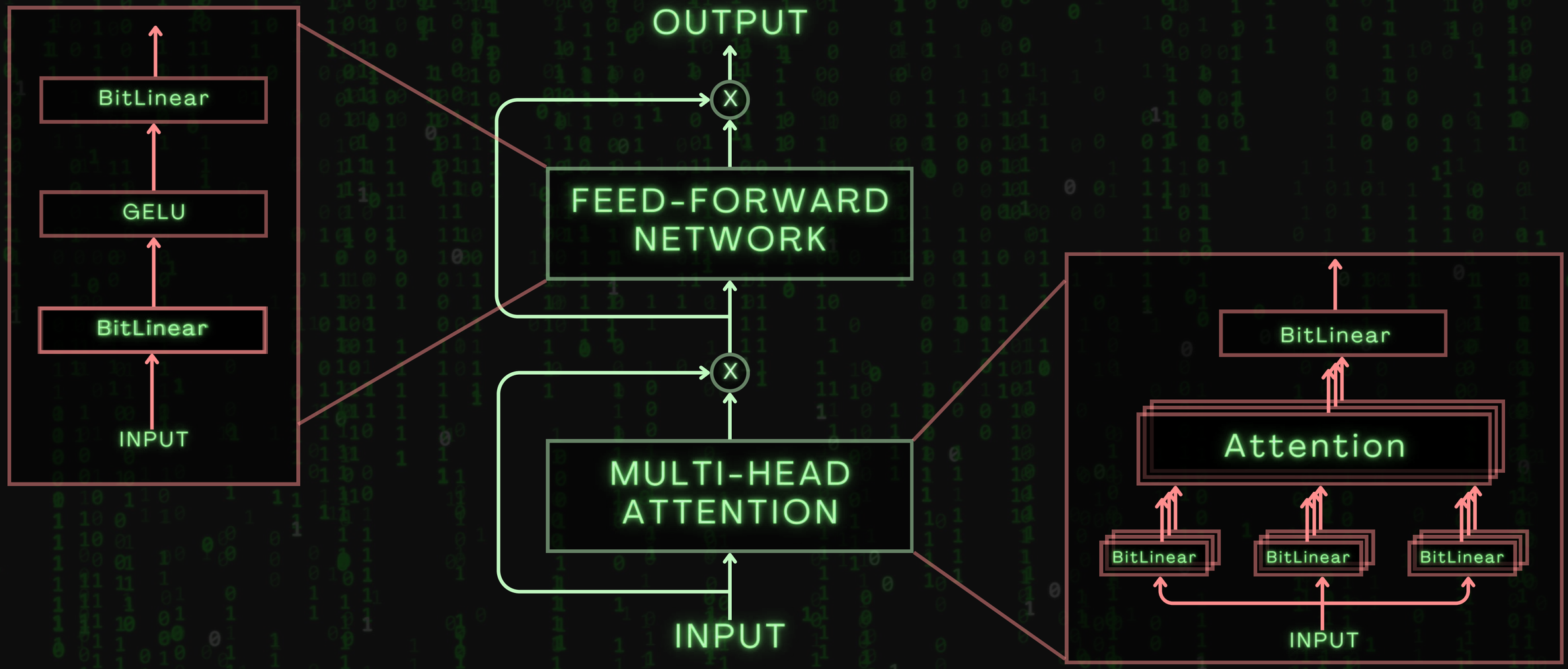
Post-Quantization

Quantization in Training

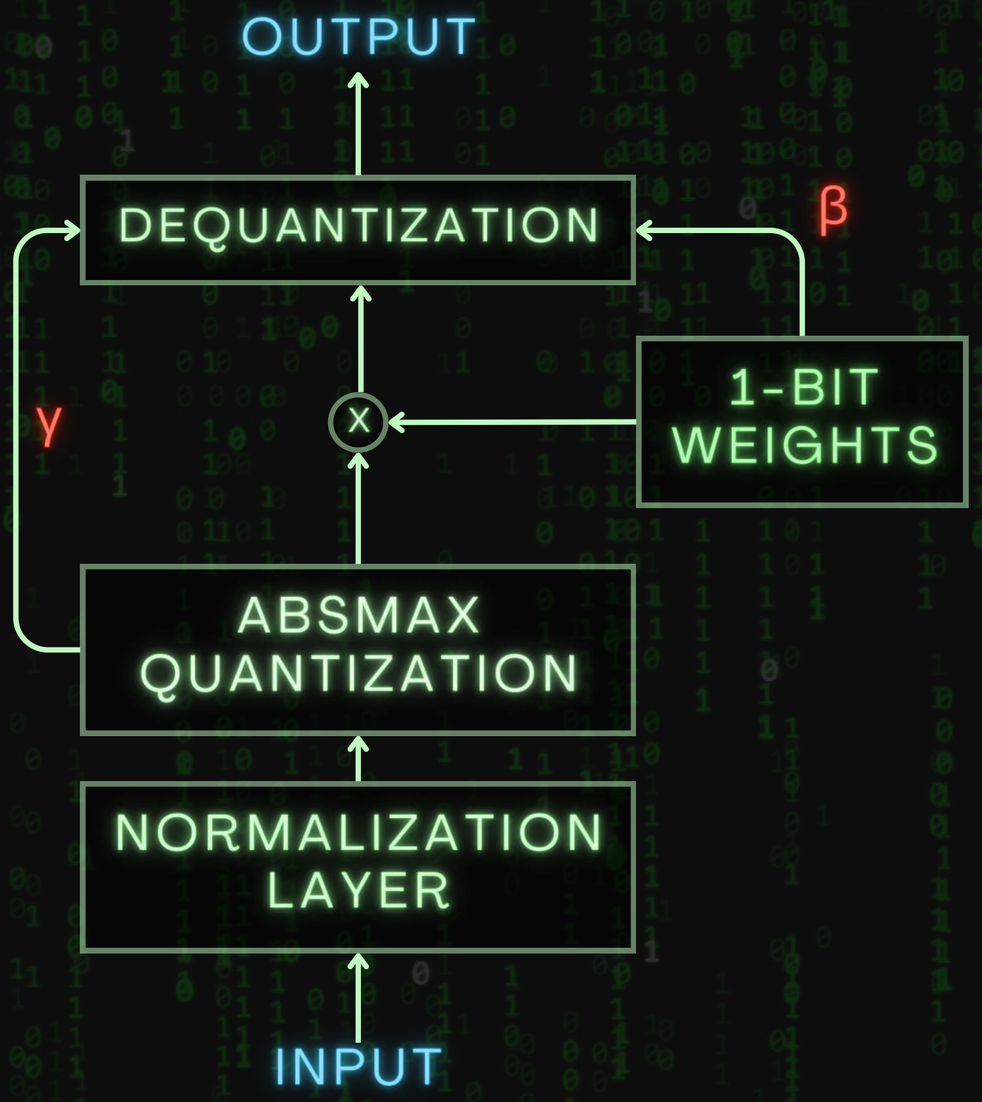
LLM TRANSFORMER



BITNET

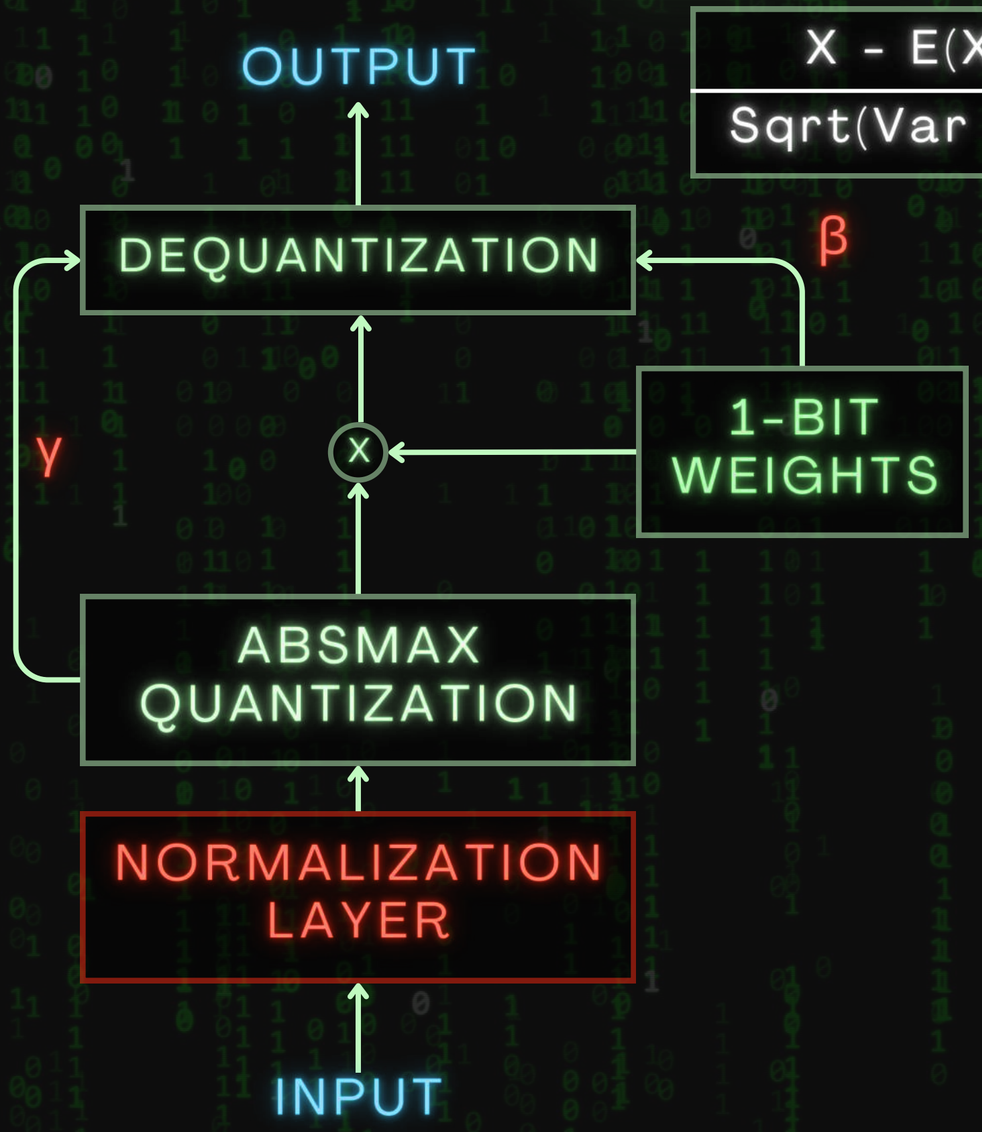

OUTPUT
X
FEED-FORWARD NETWORK
X
MULTI-HEAD ATTENTION
INPUT
BitLinear
GELU
BitLinear
INPUT
BitLinear
Attention
BitLinear
BitLinear
BitLinear
INPUT

BitLinear



Input [3.2, -1.5, 2.8, 0.9]
Weights [2.5, -1.8, 3.2, 0.7]

BitLinear



$$\frac{X - E(X)}{\text{Sqrt}(\text{Var} + \varepsilon)}$$

Input [3.2, -1.5, 2.8, 0.9]

Weights [2.5, -1.8, 3.2, 0.7]

Mean = (3.2 - 1.5 + 2.8 + 0.9) / 4 = 1.35

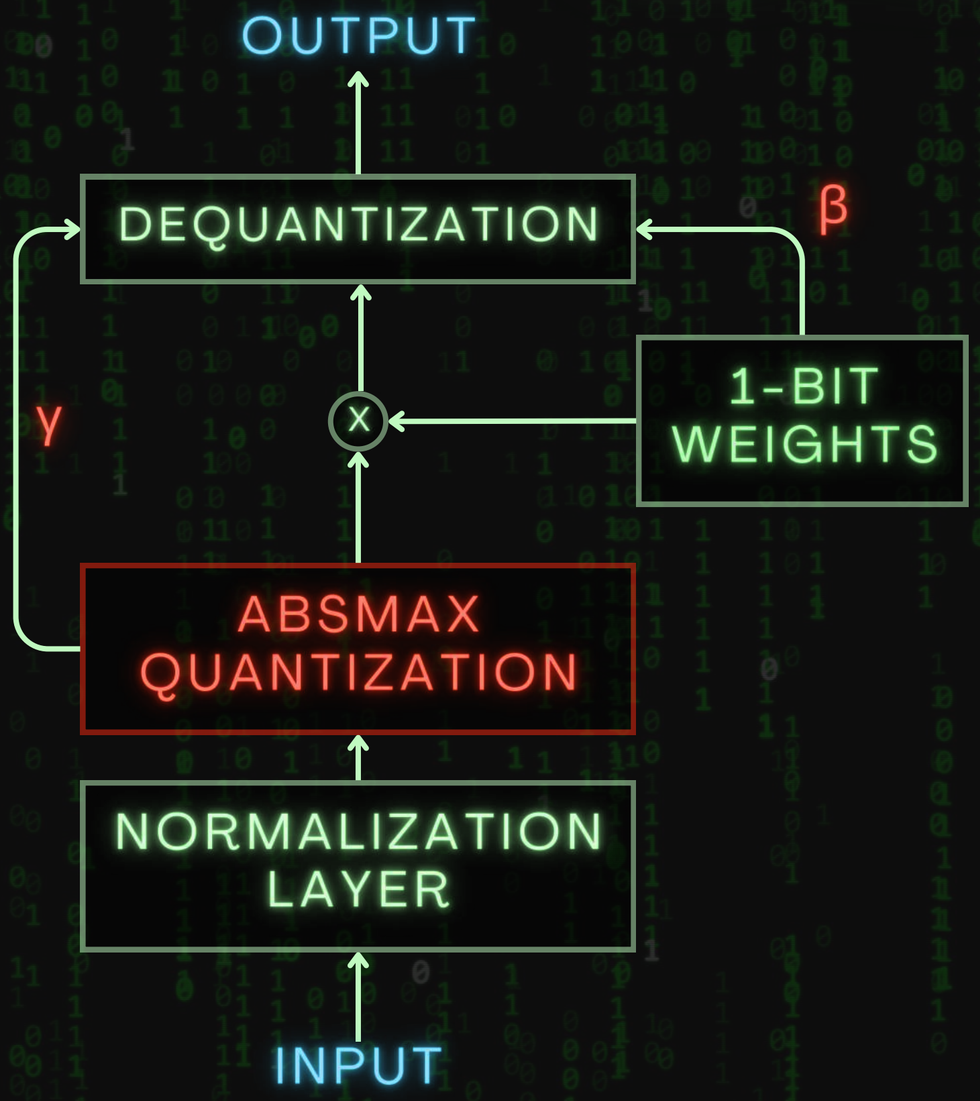
Variance = ((3.2 - 1.35)² + (-1.5 - 1.35)² + (2.8 - 1.35)² + (0.9 - 1.35)²) / 4 = 3.4225

$$\text{Normalized Input} = \frac{\text{Input} - \text{Mean}}{\text{Sqrt(Var)}}$$

Normalized Input

[1.00, -1.54, 0.78, -0.24]

BitLinear



8-bits

$$\text{ROUND}\left[\frac{\text{Normalized Input}}{\text{AbsMax}} * 128\right]$$

Normalized Input [1.00, -1.54, 0.78, -0.24]

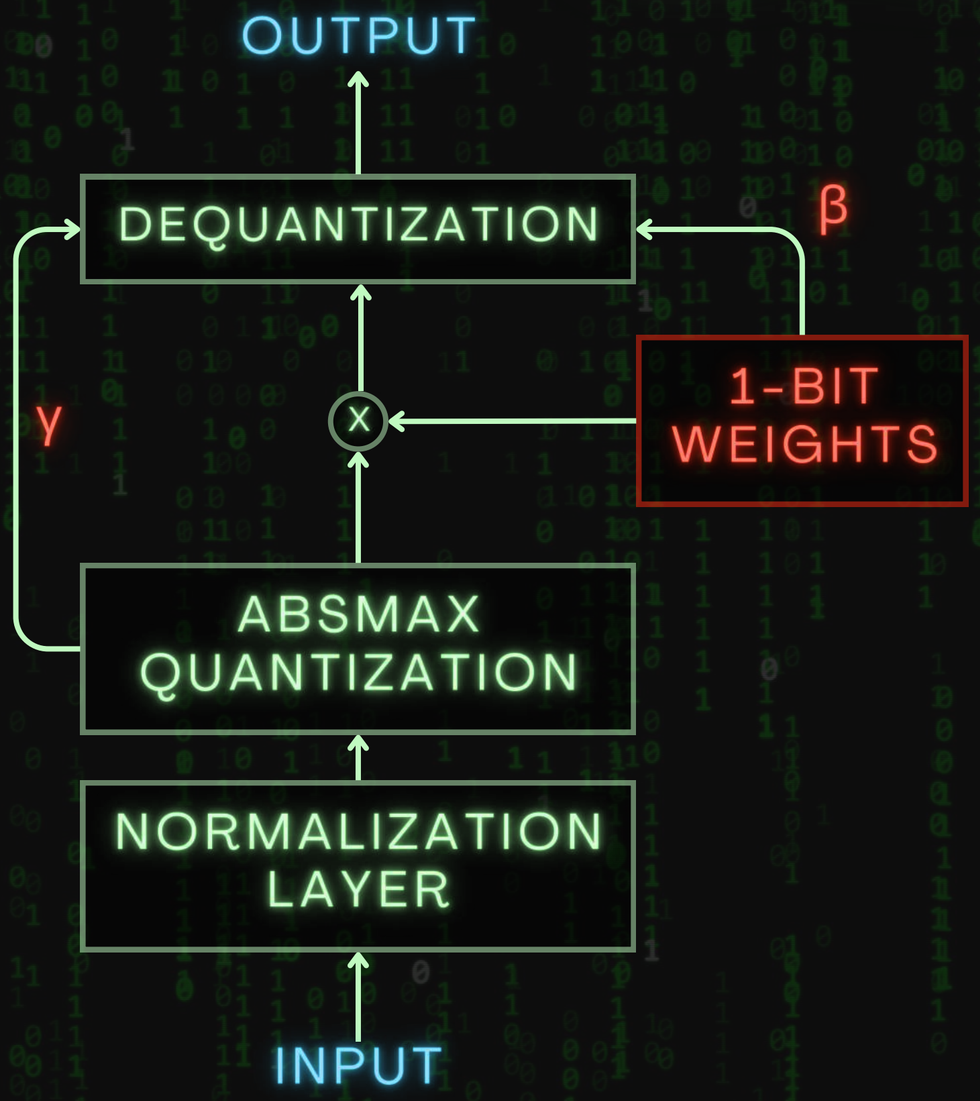
absmax = max(|1.00|, |-1.54|, |0.78|, |-0.24|)
= 1.54

Quantized = ROUND([1.00, -1.54, 0.78, -0.24] * 128 / 1.54)
= [83, -128, 65, -20]

8-bits

y = 1.54 / 128 = 0.012

BitLinear



$$\tilde{W} = \text{Sign}(W - \alpha)$$

Weights [2.5, -1.8, 3.2, 0.7]

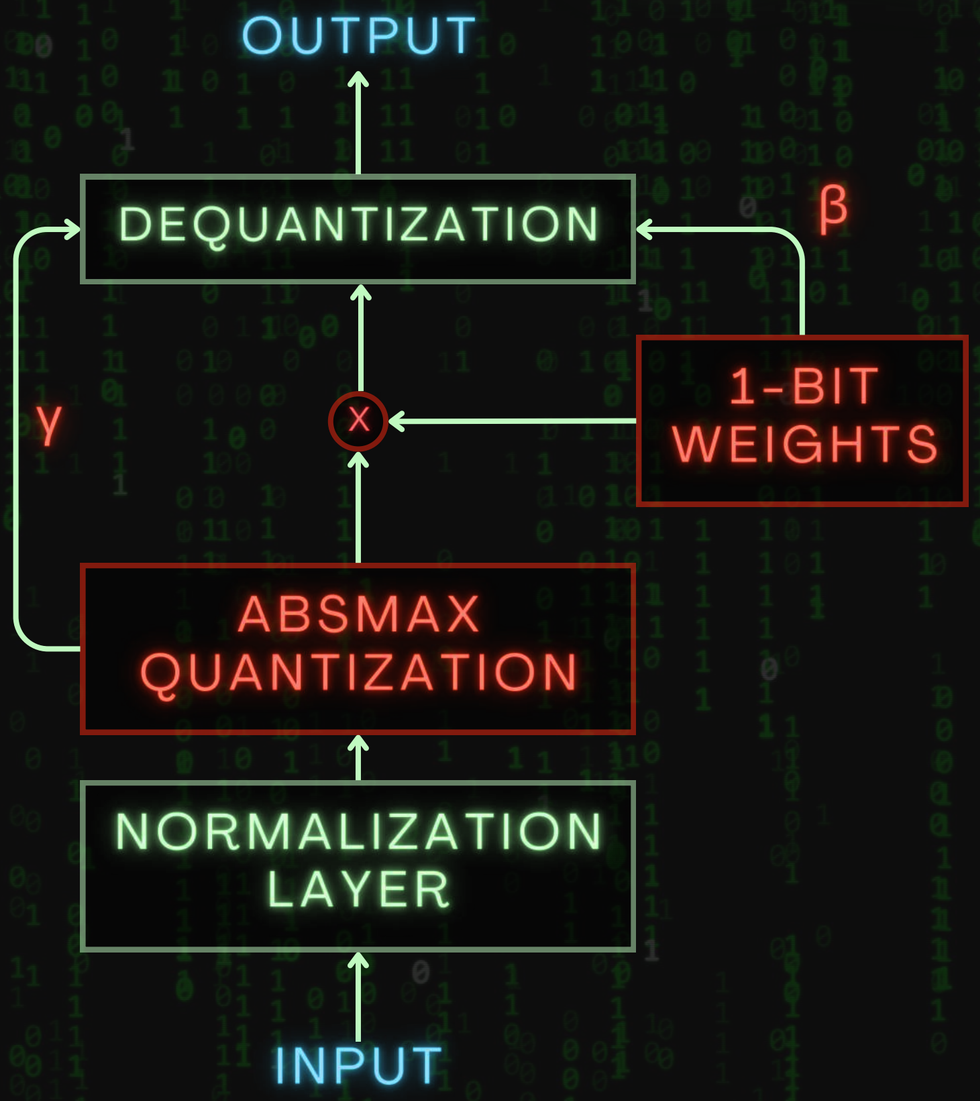
a. Calculate mean (α):
α = (2.5 - 1.8 + 3.2 + 0.7) / 4 = 1.15

b. Center weights:
Centered = [0.3, -0.5, 0.5, -0.3]

c. Apply sign function:
Binarized = [+1, -1, +1, -1]

d. Calculate scaling factor (β):
β = (|2.5| + |-1.8| + |3.2| + |0.7|) / 4 = 2.05

BitLinear

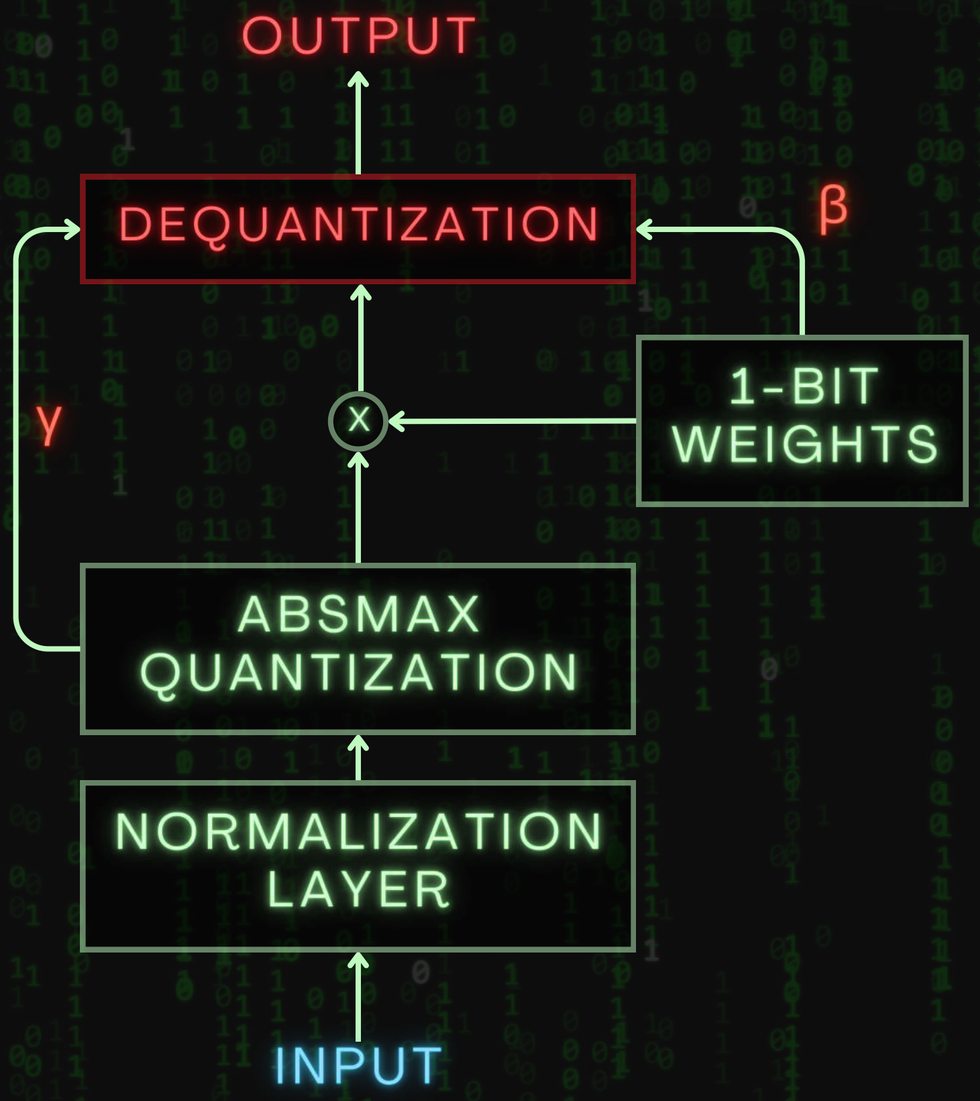


Result = Quantized Input * Binarized Weights

$$\text{Result} = [83, -128, 65, -20] \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$$

$$= 296$$

BitLinear

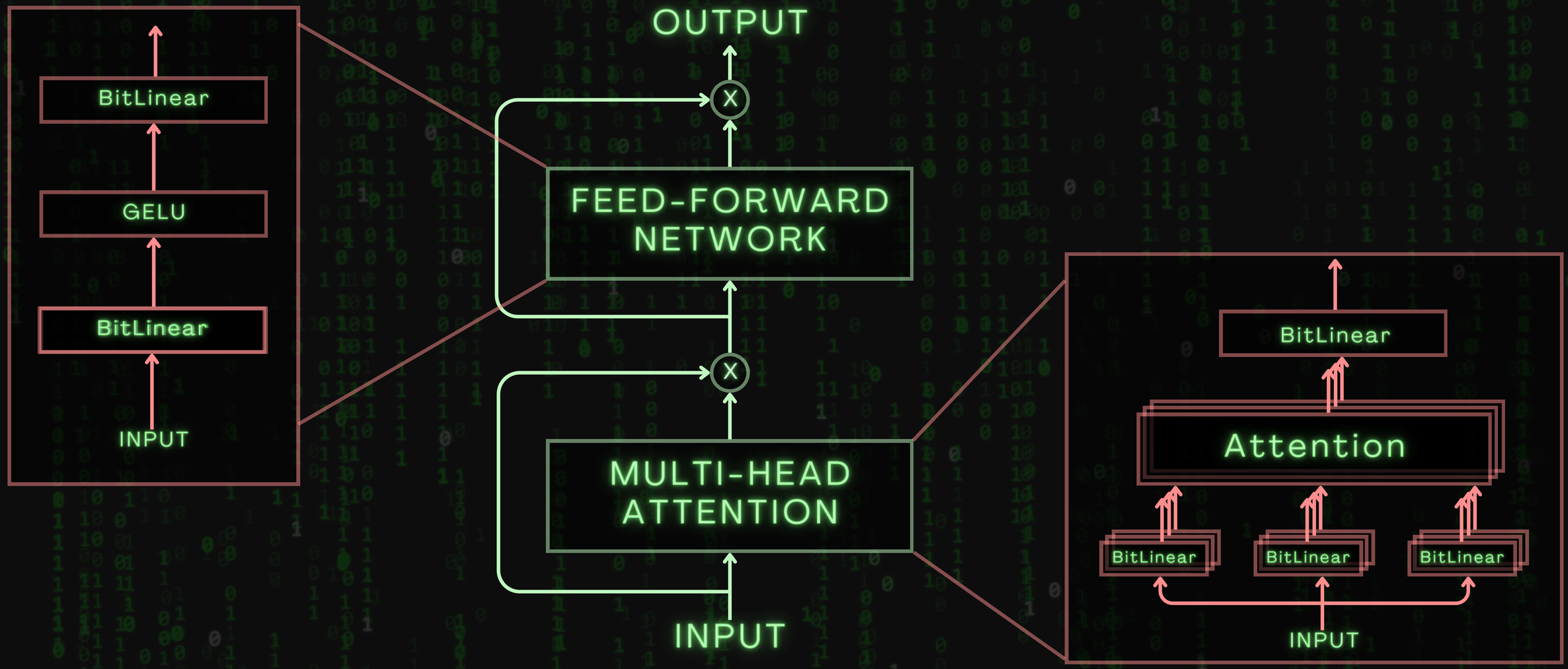


Final = Result * γ * β
Output

Dequantized Output = 296 * 0.012 * 2.05

= 7.28

BITNET


OUTPUT
X
FEED-FORWARD NETWORK
X
MULTI-HEAD ATTENTION
INPUT
BitLinear
GELU
BitLinear
INPUT
BitLinear
Attention
BitLinear
BitLinear
BitLinear
INPUT

The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits

Shuming Ma* **Hongyu Wang*** **Lingxiao Ma** **Lei Wang** **Wenhui Wang**
Shaohan Huang **Li Dong** **Ruiping Wang** **Jilong Xue** **Furu Wei◇**
https://aka.ms/GeneralAI

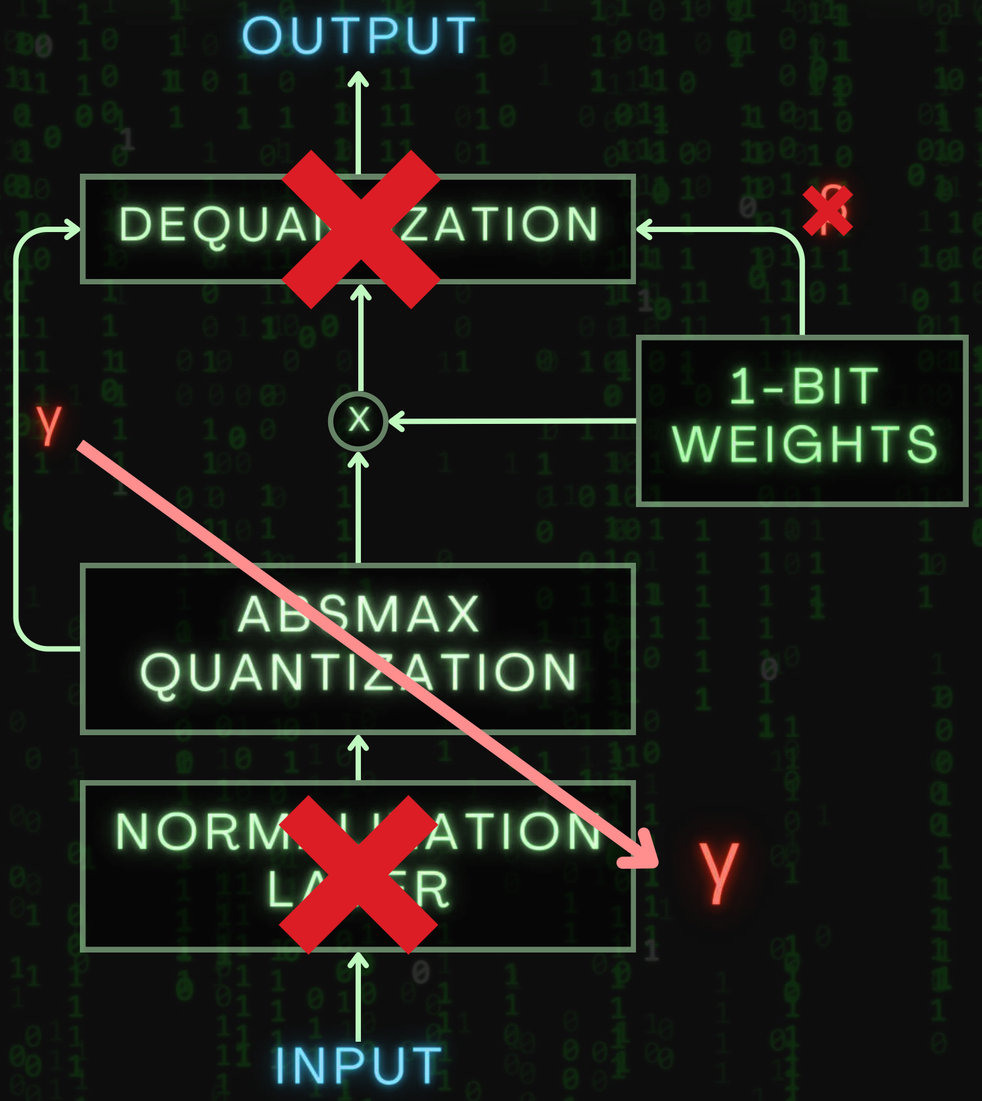
Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., & Wei, F. (2024). The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. arXiv. https://arxiv.org/pdf/2402.17764

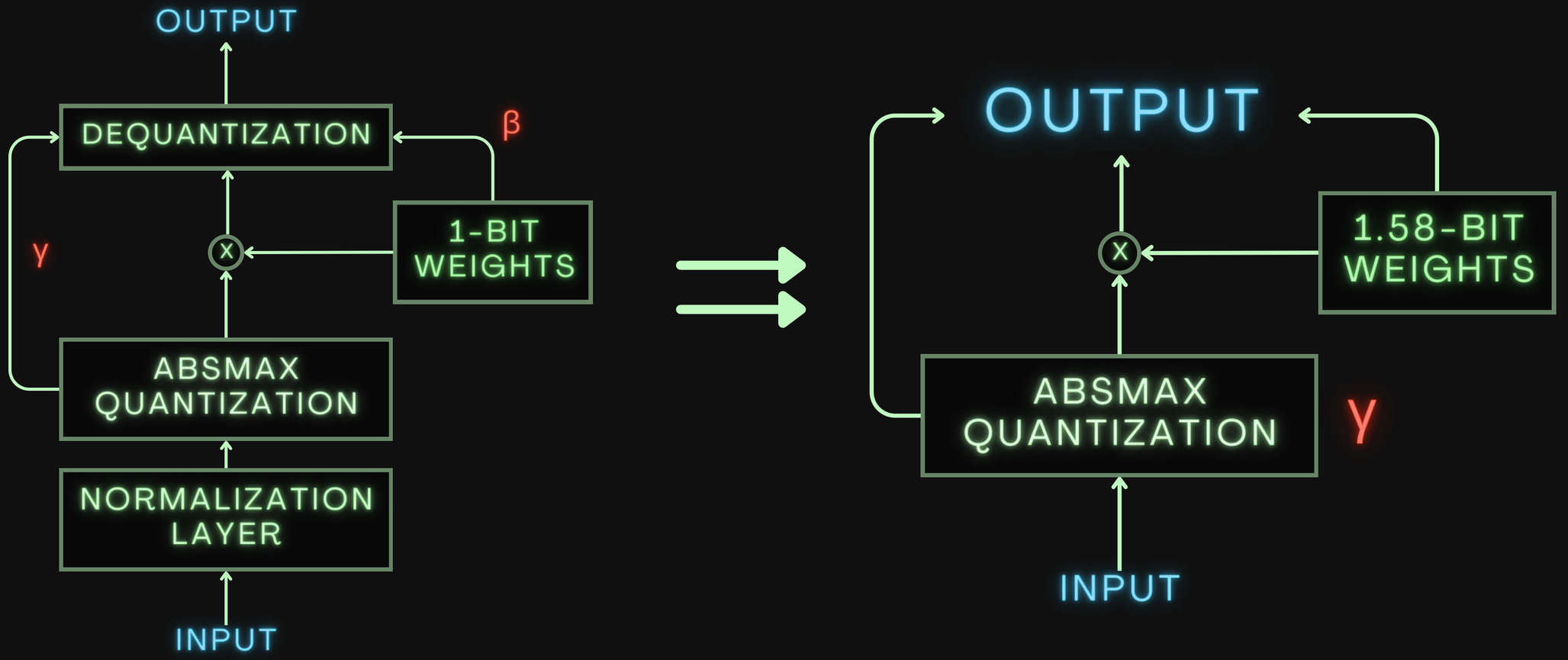
BitNet b1.58

Created by Microsoft Research in 2024

$$\{-1, 0, 1\} \rightarrow \log_2(3) = 1.58$$

BitLinear in Bitnet b.158




OUTPUT
DEQUANTIZATION
β
γ
X
1-BIT WEIGHTS
ABSMAX QUANTIZATION
NORMALIZATION LAYER
INPUT
OUTPUT
X
1.58-BIT WEIGHTS
ABSMAX QUANTIZATION
γ
INPUT

QUANTIZATION FUNCTION

$$\tilde{W} = \text{RoundClip}\left(\frac{W}{\gamma + \epsilon}, -1, 1\right)$$

$$\text{RoundClip}(x,a,b) = \max(a, \min(b, \text{round}(x)))$$

$$\gamma = \frac{1}{nm} \sum_{ij} |W_{ij}|$$

QUANTIZATION FUNCTION

Given Weight (W) Matrix

$$W = \begin{bmatrix} 0.5 & -0.8 & 1.2 \\ -1.5 & 0.3 & -0.4 \end{bmatrix}$$

QUANTIZATION FUNCTION

Given Weight (W) Matrix

$$W = \begin{bmatrix} 0.5 & -0.8 & 1.2 \\ -1.5 & 0.3 & -0.4 \end{bmatrix}$$

Calculate the Scaling Factor

$$\gamma = \frac{1}{nm} \sum_{ij} |W_{ij}|$$

QUANTIZATION FUNCTION

Given Weight (W) Matrix

$$W = \begin{bmatrix} 0.5 & -0.8 & 1.2 \\ -1.5 & 0.3 & -0.4 \end{bmatrix}$$

Calculate the Scaling Factor

$$\gamma = \frac{1}{nm} \sum_{ij} |W_{ij}|$$

$$\gamma = \frac{1}{2*3} (|0.5| + |-0.8| + |1.2| + |-1.5| + |0.3| + |-0.4|)$$

$$= \frac{1}{6} (4.7)$$

QUANTIZATION FUNCTION

Given Weight (W) Matrix

$$W = \begin{bmatrix} 0.5 & -0.8 & 1.2 \\ -1.5 & 0.3 & -0.4 \end{bmatrix}$$

Calculate the Scaling Factor

$$\gamma = \frac{1}{nm} \sum_{ij} |W_{ij}|$$

$$\gamma = \frac{1}{2*3} (|0.5| + |-0.8| + |1.2| + |-1.5| + |0.3| + |-0.4|)$$

$$= \frac{1}{6} (4.7)$$

$$= 0.7833$$

QUANTIZATION FUNCTION

Given Weight (W) Matrix

$$W = \begin{bmatrix} 0.5 & -0.8 & 1.2 \\ -1.5 & 0.3 & -0.4 \end{bmatrix}$$

Scaling Factor

$$\gamma = 0.7833$$

Constant

$$\epsilon = 0.01$$

QUANTIZATION FUNCTION

Given Weight (W) Matrix

$$W = \begin{bmatrix} 0.5 & -0.8 & 1.2 \\ -1.5 & 0.3 & -0.4 \end{bmatrix}$$

Scaling Factor

$\gamma = 0.7833$

Constant

$\epsilon = 0.01$

Scale the weight matrix

$$\frac{W}{\gamma + \epsilon} = \frac{W}{0.7833 + 0.01} = \frac{W}{0.7933}$$

QUANTIZATION FUNCTION

Given Weight (W) Matrix

$$W = \begin{bmatrix} 0.5 & -0.8 & 1.2 \\ -1.5 & 0.3 & -0.4 \end{bmatrix}$$

Scaling Factor

$$\gamma = 0.7833$$

Constant

$$\epsilon = 0.01$$

Scale the weight matrix

$$\frac{W}{\gamma + \epsilon} = \frac{W}{0.7833 + 0.01} = \frac{W}{0.7933}$$

$$W = \begin{bmatrix} \frac{0.5}{0.7933} & \frac{-0.8}{0.7933} & \frac{1.2}{0.7933} \\ \frac{-1.5}{0.7933} & \frac{0.3}{0.7933} & \frac{-0.4}{0.7933} \end{bmatrix}$$

QUANTIZATION FUNCTION

$$W = \begin{bmatrix} \frac{0.5}{0.7933} & \frac{-0.8}{0.7933} & \frac{1.2}{0.7933} \\ \frac{-1.5}{0.7933} & \frac{0.3}{0.7933} & \frac{-0.4}{0.7933} \end{bmatrix} = \begin{bmatrix} 0.63 & -1.01 & 1.51 \\ -1.89 & 0.38 & -0.50 \end{bmatrix}$$

QUANTIZATION FUNCTION

$$W = \begin{bmatrix} 0.63 & -1.01 & 1.51 \\ -1.89 & 0.38 & -0.50 \end{bmatrix}$$

Apply the Clip Function

$$\text{RoundClip}(x,a,b) = \max(a, \min(b, \text{round}(x)))$$

QUANTIZATION FUNCTION

Scaled Weights

$$W = \begin{bmatrix} 0.63 & -1.01 & 1.51 \\ -1.89 & 0.38 & -0.50 \end{bmatrix}$$

Apply the Clip Function to the Scaled Weights

RoundClip(0.63, -1, 1) = 1
RoundClip(-1.01, -1, 1) = -1
RoundClip(1.51, -1, 1) = 1
RoundClip(-1.89, -1, 1) = -1
RoundClip(0.38, -1, 1) = 0
RoundClip(-0.50, -1, 1) = 0

QUANTIZATION FUNCTION

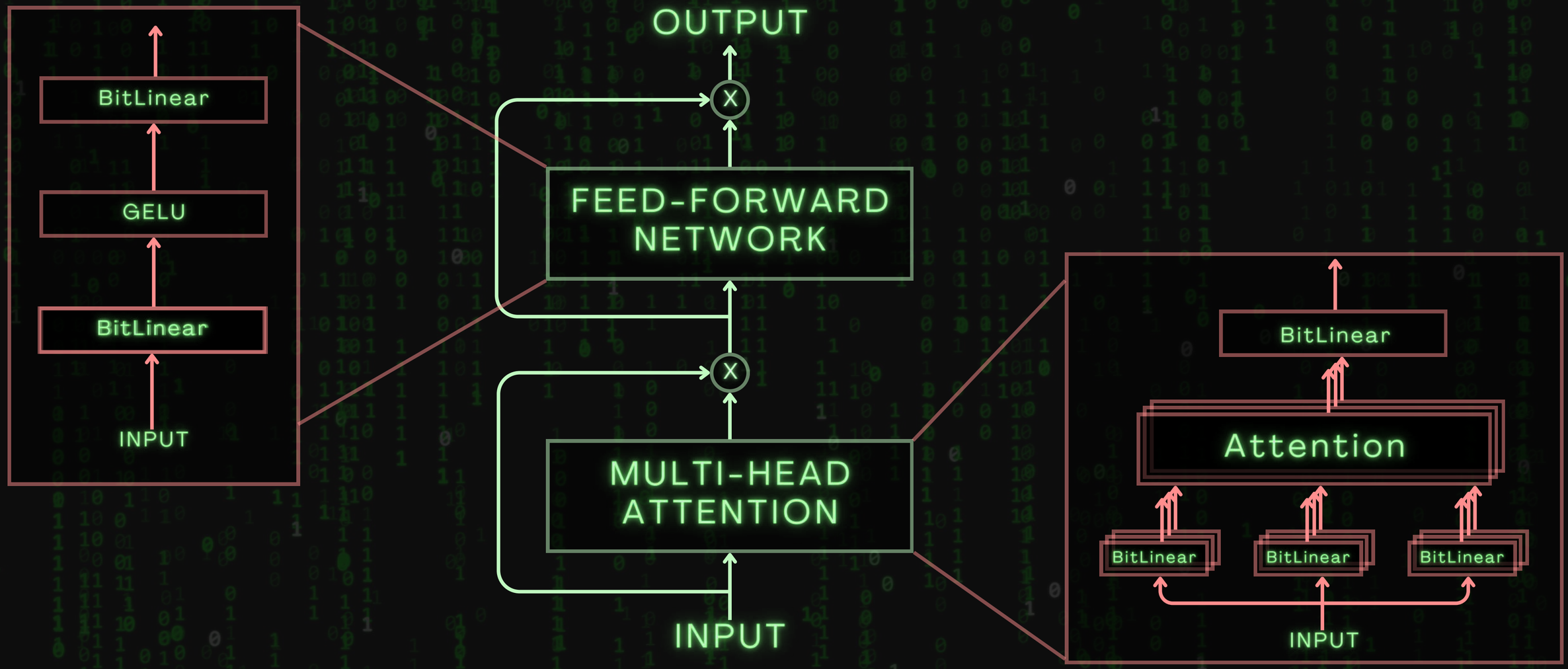
Apply the Clip Function to the Scaled Weights

RoundClip(0.63, -1, 1) = 1
RoundClip(-1.01, -1, 1) = -1
RoundClip(1.51, -1, 1) = 1
RoundClip(-1.89, -1, 1) = -1
RoundClip(0.38, -1, 1) = 0
RoundClip(-0.50, -1, 1) = 0

Quantized Weight Matrix

$$W = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 0 & 0 \end{bmatrix}$$

LLM TRANSFORMER



RESULTS

BitNet b1.58 Results

Models	Size	Memory (GB)↓	Latency (ms)↓	PPL↓
LLaMA LLM	700M	2.08 (1.00x)	1.18 (1.00x)	12.33
BitNet b1.58	700M	0.80 (2.60x)	0.96 (1.23x)	12.87
LLaMA LLM	1.3B	3.34 (1.00x)	1.62 (1.00x)	11.25
BitNet b1.58	1.3B	1.14 (2.93x)	0.97 (1.67x)	11.29
LLaMA LLM	3B	7.89 (1.00x)	5.07 (1.00x)	10.04
BitNet b1.58	3B	**2.22 (3.55x)**	**1.87 (2.71x)**	**9.91**
BitNet b1.58	3.9B	**2.38 (3.32x)**	**2.11 (2.40x)**	**9.62**

Computational Cost and Perplexity of BitNet b1.58 and LLaMA LLM

BitNet b1.58 Results

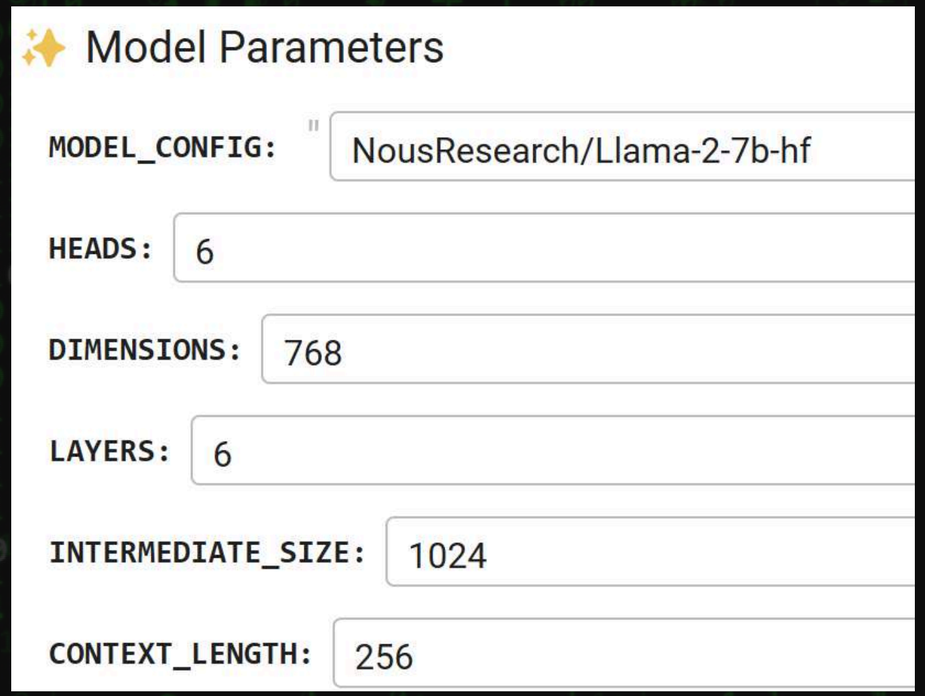
Models	Size	ARCe	ARCc	HS	BQ	OQ	PQ	WGe	Avg.
LLaMA LLM	700M	54.7	23.0	37.0	60.0	20.2	68.9	54.8	45.5
BitNet b1.58	700M	51.8	21.4	35.1	58.2	20.0	68.1	55.2	44.3
LLaMA LLM	1.3B	56.9	23.5	38.5	59.1	21.6	70.0	53.9	46.2
BitNet b1.58	1.3B	54.9	24.2	37.7	56.7	19.6	68.8	55.8	45.4
LLaMA LLM	3B	62.1	25.6	43.3	61.8	24.6	72.1	58.2	49.7
BitNet b1.58	3B	**61.4**	**28.3**	**42.9**	**61.5**	**26.6**	**71.5**	**59.3**	**50.2**
BitNet b1.58	3.9B	**64.2**	**28.7**	**44.2**	**63.5**	**24.2**	**73.2**	**60.5**	**51.2**

Accuracy of BitNet b1.58 and LLaMA LLM on the End Tasks

SIMULATION

AutoBitnet

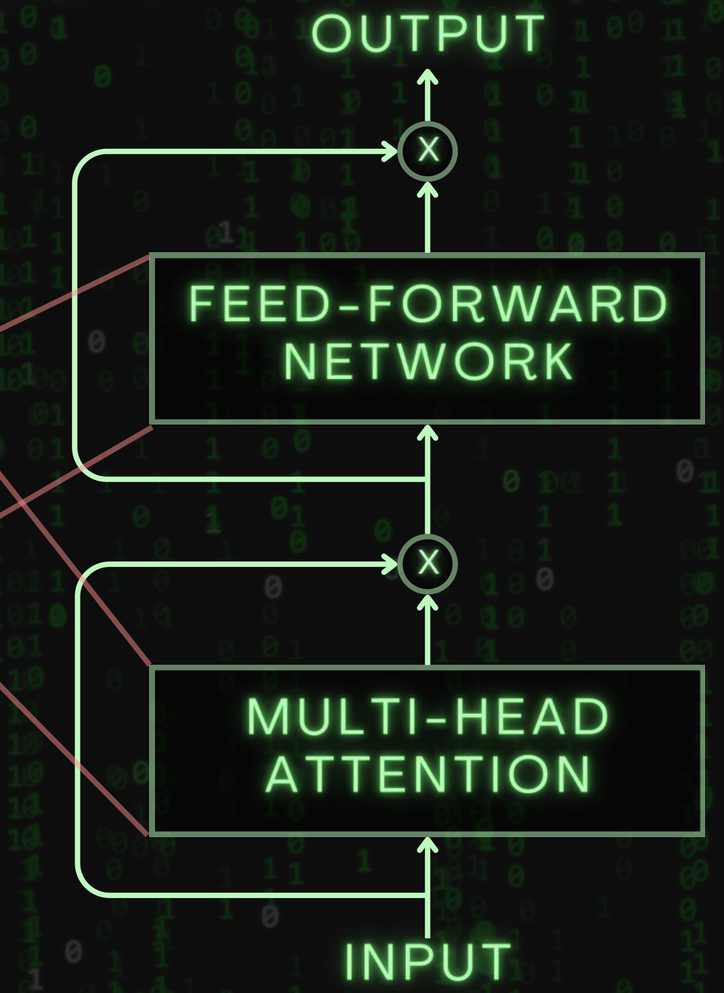
Automated tool that allows to train a BitNet b1.58 on the baselines of any LLaMA architecture on a Google Colab T4 GPU.



Abideen, Z. ul. (2024, April 4). Llama-Bitnet | Training a 1.58 bit LLM. Medium. https://medium.com/@zaiinn440/llama-bitnet-training-a-1-58-bit-llm-3831e517430a

AutoBitnet

```
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(32000, 768, padding_idx=0)
    (layers): ModuleList(
      (0-5): 6 x LlamaDecoderLayer(
        (self_attn): LlamaSdpaAttention(
          (q_proj): BitLinear(in_features=768, out_features=768, bias=False)
          (k_proj): BitLinear(in_features=768, out_features=768, bias=False)
          (v_proj): BitLinear(in_features=768, out_features=768, bias=False)
          (o_proj): BitLinear(in_features=768, out_features=768, bias=False)
          (rotary_emb): LlamaRotaryEmbedding()
        )
        (mlp): LlamaMLP(
          (gate_proj): BitLinear(in_features=768, out_features=1024, bias=False)
          (up_proj): BitLinear(in_features=768, out_features=1024, bias=False)
          (down_proj): BitLinear(in_features=1024, out_features=768, bias=False)
          (act_fn): SiLU()
        )
        (input_layernorm): Identity()
        (post_attention_layernorm): LlamaRMSNorm()
      )
    )
    (norm): LlamaRMSNorm()
  )
  (lm_head): Linear(in_features=768, out_features=32000, bias=False)
)
```



AutoBitnet

Training Parameters

Decoder Layers = 6
Input Size = 768
Output Size = 768

Output Features = 32,000

Retrained Model

NousResearch/Llama-2-7b-hf

Dataset

Cosmopedia-100k-pretrain

Training Time

5:55:24

```
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(32000, 768, padding_idx=0)
    (layers): ModuleList(
      (0-5): 6 x LlamaDecoderLayer(
        (self_attn): LlamaSdpaAttention(
          (q_proj): BitLinear(in_features=768, out_features=768, bias=False)
          (k_proj): BitLinear(in_features=768, out_features=768, bias=False)
          (v_proj): BitLinear(in_features=768, out_features=768, bias=False)
          (o_proj): BitLinear(in_features=768, out_features=768, bias=False)
          (rotary_emb): LlamaRotaryEmbedding()
        )
        (mlp): LlamaMLP(
          (gate_proj): BitLinear(in_features=768, out_features=1024, bias=False)
          (up_proj): BitLinear(in_features=768, out_features=1024, bias=False)
          (down_proj): BitLinear(in_features=1024, out_features=768, bias=False)
          (act_fn): SiLU()
        )
        (input_layernorm): Identity()
        (post_attention_layernorm): LlamaRMSNorm()
      )
    )
    (norm): LlamaRMSNorm()
  )
  (lm_head): Linear(in_features=768, out_features=32000, bias=False)
)
```

Total number of parameters: 77,468,928

AutoBitnet Inferences

```
prompt = "Complete the term: I am a Filipino and"
```

```
Generated text:
Complete the term: I am a Filipino and I'm not sure what you're doing.
```

```
prompt = "Make a concise answer to the question, What is the meaning of life?"
```

```
Generated text:
Make a concise answer to the question, What is the meaning of life?

I'm not sure what I mean, but I've been trying to find the answer. I think I have a lot of things that I can't help but feel a bit of a little bit.
```

```
prompt = "What is love?"
```

```
Generated text:
What is love?

I. Introduction
A. Definition of key terms
1. The role of a child in the world
2. Explanation of the importance of understanding the relationship between the two and the other
3. Aesthetic and psychological factors
B. Importance of studying the role and its role in shaping the future of healthcare
C. Overview of how the concept of personal development and empowerment
D. Practical example
```

CONCLUSION

JOURNAL CRITIQUE

The Era of 1-bit LLMs:
All Large Language Models are in 1.58 Bits

Shuming Ma* Hongyu Wang* Lingxiao Ma Lei Wang Wenhui Wang
Shaohan Huang Li Dong Ruiping Wang Jilong Xue Furu Wei◇
https://aka.ms/GeneralAI

Lacks Predecessor Context.

Narrow Focus.

BitNet1.58 can do...

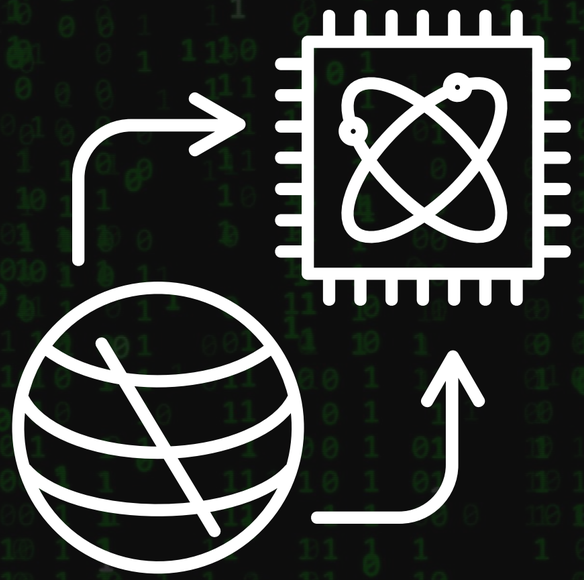
Energy Savings

Faster Run Time

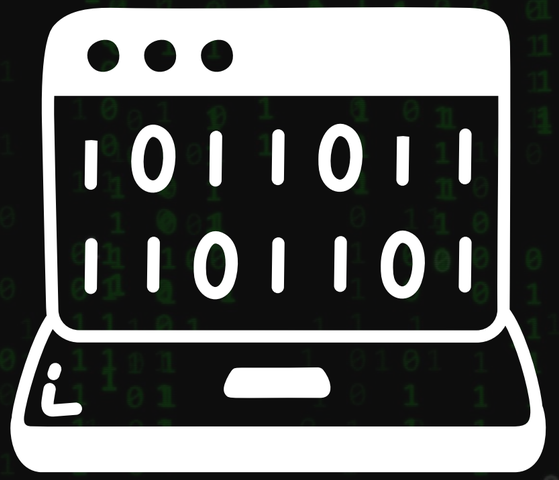
Cost Savings

APPLICATION

Mobile Devices



Binary-Ternary Emulation



Hardware Changes

Azhar, A. (2024, March 1). No more floating points, the era of 1.58-bit large language models. Azhar Labs. Medium. https://medium.com/ai-insights-cobet/no-more-floating-points-the-era-of-1-58-bit-large-language-models-b9805879ac0a

“Are we getting closer to the matrix?”

REFERENCES

References:

Abideen, Z. ul. (2024, April 4). Llama-Bitnet | Training a 1.58 bit LLM. Medium. https://medium.com/@zaiinn440/llama-bitnet-training-a-1-58-bit-llm-3831e517430a

Azhar, A. (2024, March 1). No more floating points, the era of 1.58-bit large language models. Azhar Labs. Medium. https://medium.com/ai-insights-cobet/no-more-floating-points-the-era-of-1-58-bit-large-language-models-b9805879ac0a

Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., & Wei, F. (2024). The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. arXiv. https://arxiv.org/pdf/2402.17764

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., Yang, F., Wang, R., Wu, Y., & Wei, F. (2023). BitNet: Scaling 1-bit Transformers for Large Language Models. arXiv. https://arxiv.org/abs/2310.11453