

OliVe: Accelerating Large Language Models via Hardware-friendly Outlier-Victim Pair Quantization

Cong Guo*

guocong@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai Qi Zhi Institute
Shanghai, China

Jiaming Tang*

sakits_tjm@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai Qi Zhi Institute
Shanghai, China

Weiming Hu[†]

huweim1120@gmail.com
Shanghai Jiao Tong University
Shanghai Qi Zhi Institute
Shanghai, China

Jingwen Leng[‡]

leng-jw@cs.sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai Qi Zhi Institute
Shanghai, China

Chen Zhang

chzhang1990@gmail.com
Microsoft Research
Beijing, China

Fan Yang

fanyang@microsoft.com
Microsoft Research
Beijing, China

Yunxin Liu

liuyunxin@air.tsinghua.edu.cn
Institute for AI Industry Research
(AIR), Tsinghua University
Beijing, China
Shanghai Artificial Intelligence
Laboratory
Shanghai, China

Minyi Guo[‡]

guo-my@cs.sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai Qi Zhi Institute
Shanghai, China

Yuhao Zhu

yzhu@rochester.edu
University of Rochester
Rochester, New York, USA

ABSTRACT

Transformer-based large language models (LLMs) have achieved great success with the growing model size. LLMs' size grows by 240× every two years, which outpaces the hardware progress and makes model inference increasingly costly. Model quantization is a promising approach to mitigate the widening gap between LLM size and hardware capacity. However, the existence of outliers, values with significant magnitudes, in LLMs makes existing quantization methods less effective. Prior outlier-aware quantization schemes adopt sparsity encoding techniques to separate outliers from normal values where the process requires *global* coordination (e.g., a global sparsity coordination list). This incurs complex encoding/decoding hardware logics and an extra orchestration controller for the computation between outlier and normal values. As such, it is not hardware-efficient and hence only achieves sub-optimal quantization benefits.

*Contribute equally to this paper.

[†]Work done while affiliated with ShanghaiTech University.

[‡]Jingwen Leng and Minyi Guo are corresponding authors of this paper.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ISCA '23, June 17–21, 2023, Orlando, FL, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0095-8/23/06...\$15.00

<https://doi.org/10.1145/3579371.3589038>

We propose OliVe, an algorithm/architecture co-designed solution that adopts an outlier-victim pair (OVP) quantization and handles outlier values *locally* with low hardware overheads and high performance gains. The key insight of OliVe is that outliers are important while the normal values *next* to them are not. Thus those normal values (called victims) can be sacrificed to accommodate outliers. This enables a memory-aligned OVP encoding scheme, which can be efficiently integrated to the existing hardware accelerators like systolic array and tensor core. As a result, OliVe-based accelerator surpasses the existing outlier-aware accelerator, GOBO, by 4.5× speedup and 4.0× energy reduction, respectively, with a superior model accuracy.

CCS CONCEPTS

• **Computer systems organization** → **Neural networks; Data flow architectures; Single instruction, multiple data; Systolic arrays.**

KEYWORDS

Large Language Model, Outlier-Victim Pair, Quantization

ACM Reference Format:

Cong Guo, Jiaming Tang, Weiming Hu, Jingwen Leng, Chen Zhang, Fan Yang, Yunxin Liu, Minyi Guo, and Yuhao Zhu. 2023. OliVe: Accelerating Large Language Models via Hardware-friendly Outlier-Victim Pair Quantization. In *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23)*, June 17–21, 2023, Orlando, FL, USA. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3579371.3589038>

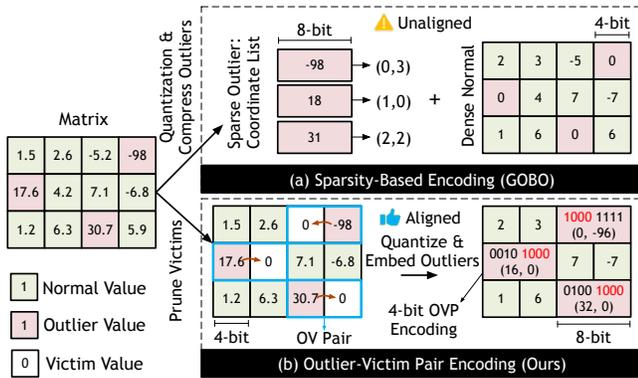


Figure 1: Outlier-aware encoding comparison. (a) Prior quantization works adopt sparsity-based encoding that store normal and outlier values separately. (b) Our proposed outlier-victim pair encoding stores normal and outlier values locally.

1 INTRODUCTION

Transformer-based large language models (LLMs) [77] have demonstrated great success in the past years. Such success is often achieved with the increasingly larger model size: the model size grows by $240\times$ every two years, significantly outpacing the hardware progress ($3.1\times$ per two years) [24]. As a result, the inference of LLMs becomes challenging and costly. For instance, OPT-175B [90], a recent Transformer-based LLM, has 175 billion parameters, which cannot fit in the latest high-end H100 GPU with 80GB memory.

Quantization [6, 7, 21, 22, 72, 74, 79, 93] is one of the most hardware-efficient ways to reduce inference costs for large models. It uses low-precision data types to compress models and accelerate the computation with practical hardware implementations, e.g., TPU [42] and GPU tensor core [60].

However, existing quantization schemes [18, 74, 86] are less effective in Transformer-based LLMs. Recent studies show when the model size exceeds a threshold (e.g., 6 billion), the model performance is vulnerable to only a tiny fraction ($< 0.1\%$) of outliers, whose values are much more significant than normal values [18]. Indiscriminately clipping both outlier and normal values will lead to significant drops in model accuracy [18, 82]. As a result, the common practice is to adopt a larger bit-width, e.g., 8-bit or 16-bit, to quantize Transform-based models, compared to convolutional networks (CNNs).

Researchers have proposed various quantization/architecture co-design works [39, 61, 75, 82, 85] to deal with the outliers in Transformer models. For example, outlier suppression [82] proposes to suppress the outliers. But it still has significant accuracy loss in the lower bit-width (4-bit), suggesting the difficulty in accommodating the effects of outliers. In addition, architecture researchers have designed sophisticated outlier-aware hardware architectures to store outliers with high precision to maintain model accuracy. These outlier-aware quantization frameworks divide the tensor into normal and outlier values, and encode them separately using different ways. For normal values, a dense matrix with low precision (e.g., 4-bit) quantization is adopted. And the sparse and high-precision (e.g.,

8-bit and 16-bit) outlier values can be compressed with sparsity-based encoding. Such encoding unfortunately leads to unaligned memory access. For example, GOBOs [85] and OLAccels [61] use the coordinate list to indicate the location of each outlier value in the matrix, as shown in Fig. 1a. BiScaled-DNNs [39] exploits block sparse indices format to store the outlier indices, and DRQ [75] uses the direct bitmap for outliers. These outlier-aware solutions require complex architectural designs with significant hardware overheads to accommodate outliers. Moreover, due to the random and unaligned memory access, the sparsity-based encoding is incompatible with the memory sub-systems of existing accelerators, such as GPU and TPU. Specifically, GOBO [85] can only de/compress weight tensors on the off-chip DRAM, it still relies on the original on-chip memory and computation architecture of GPU with high precision FP16/32.

The aforementioned outlier-aware architectures separate normal values from outliers in a *global* way. For instance, GOBO [85] involves a global sparse coordinate list in the quantization and computation, leading to a large hardware overhead and low performance benefits. In this work, we aim to design an architecture to handle outliers in a *localized* way with high hardware efficiency. To achieve that, we group two consecutive fixed-size values in a tensor and analyze their impact to model accuracy. There can be three kinds of pairs: i) a normal pair with two normal values, ii) one-outlier pair with one normal value and one outlier value, iii) two-outlier pair with two outlier values. We observe that the third two-outlier pair almost never shows up in well-trained LLMs. For the second one-outlier pair, we find that *only keeping its outlier value while pruning its normal value* (i.e., treating it as zero) is sufficient to maintain the model accuracy.

Based on the above observations, we propose a novel outlier-aware quantization architecture, called *Olive*, based on the outlier-victim pair (OVP) encoding. The salient feature of *Olive* is memory-aligned and therefore hardware-friendly. As illustrated in Fig. 1b, *Olive* first prunes normal values that are adjacent to the outliers as zero. These pruned normal values are called *victims*, which sacrifice themselves and make space for outliers. Then, we exploit the extra space provided by victims and embed the outliers into the low-precision matrix.

Olive is able to maintain a high accuracy for large Transformer models with a low hardware overhead due to the following reasons. First, *Olive* incorporates victims to tackle outliers in LLMs. The effects of victims resemble model pruning [36]. Although clipping a few (0.1%) outliers will lead to a disastrous accuracy drop [18, 82], pruning the same amount of “normal” values will only impact model accuracy slightly ($< 0.1\%$ drop). Therefore, *Olive* sacrifices (“prunes”) those insignificant values as victims for the outliers, allowing a more aggressive encoding scheme to accommodate extremely significant values. Second, the OVP encoding follows a specific outlier-victim (or victim-outlier) pattern to achieve memory alignment with little hardware overheads. Each victim is adjacent to an outlier, and the outlier-victim pair must align the memory access pattern. For example, in Fig. 1b, right outlier -98 in the OVP pair needs a left victim, and left outliers 17.6 and 30.7 require the right victims. That can align 8-bit (1-byte) memory accesses with high efficiency. This design enables a completely localized outlier decoding/encoding process.

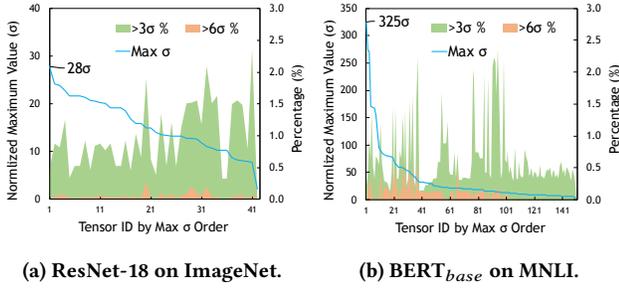


Figure 2: Outlier Comparison of CNN model and Transformer model. The σ is the standard deviation of the tensor. We normalize the maximum number by σ to plot the Max σ curve (left y-axis). The $> 3\sigma\%$ and $> 6\sigma\%$ (right y-axis) are the percentage of the values of $> 3\sigma$ and $> 6\sigma$, respectively.

To implement OliVe, different data types are employed for outliers and normal values, which have different dynamic ranges and representation formats, including int4 and FP4. As shown in Fig. 1b, we propose a novel encoding method (Sec. 3) for the 4-bit OV pair, which composes a 4-bit outlier and a 4-bit victim into a special 8-bit format and differs from the original int8 or FP8. Due to its hardware-friendly and compatible design, OliVe can be easily integrated into existing quantization frameworks and accelerator architectures such as systolic array in Google TPUs [41] and tensor core in NVIDIA GPUs [58, 60]. OliVe can also inherently support the mixed-precision and mixed-type architecture, showing its flexibility and practicality for larger-scale Transformer models.

To the best of our knowledge, OliVe is the first work pushing the limit of Transformer post-training quantization (PTQ) [4], which requires no retraining after quantization, to the 4-bit level for both the weight and activation tensors with the accuracy loss of $< 1\%$. Surprisingly, OliVe’s 4-bit PTQ accuracies for BERT [19] and BART [49] models outperform the 6-bit PTQ results of outlier suppression [82], a state-of-the-art Transformer quantization method. OliVe-based accelerator surpasses the existing outlier-aware accelerators OLAcel [61] and GOBO [85] by 3.8 \times and 4.5 \times performance improvement, and 2.1 \times and 4.0 \times energy reduction, respectively. More importantly, the OliVe-based accelerator has more comprehensive and practical applicability than other outlier-specific architectures.

We make the following contributions in this paper.

- We conduct the pair-wise importance analysis and show that outliers are important while their adjacent normal values are not, revealing the algorithmic opportunity of outlier-victim pair (OVP) that sacrifices the colocated normal values (called victims) to accommodate the outliers.
- We propose the OVP-based quantization framework, called OliVe, which includes an efficient hardware encoding and novel outlier representation data type.
- We propose the efficient architectural implementation and integration of OliVe quantization, and show that its efficiency and benefits outperform the existing outlier-aware quantization algorithms and hardware accelerators.

2 MOTIVATION: ALIGNED OUTLIER

In this section, we first show that the outlier of the Transformer model is much more significant and important compared to convolution neural networks (CNN). Previous works [74, 75, 85, 86] propose the outlier-aware quantization microarchitecture with adaptive bit length to accomplish the low-bit quantization but necessitate substantial hardware resources to deal with the variable-length data, which cause unaligned memory accesses and are incompatible with the memory sub-system of existing accelerators, e.g., GPU [60]. In contrast, we propose a memory-aligned and hardware-friendly method, called outlier-victim pair mechanism, which is inspired by DNN pruning and our outlier group location analysis for Transformers. We can prune some “victims” to make space to embed high-precision outliers into the memory-aligned low-bit tensor with ignorable accuracy loss.

2.1 Outlier Matters

We visually demonstrate how significant the Transformer’s outlier is in Fig. 2. We adopt the empirical 3σ rules [83] of the normal distribution to divide the values into outlier and normal values. We employ the ResNet-18 [37] as the representative for the CNN model and the BERT_{base} [19] for the Transformer model. We fit the DNN tensors with normal distribution, i.e., Equation 1, where x is the value, μ is the mean, and σ is the standard deviation. We convert the tensor into a standard normal distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (1)$$

We collect all tensors’ maximum values and normalize them by the σ (Max σ). We sort and plot the tensors by their Max σ in Fig. 2.

Most tensors can fit the normal distribution 3σ rules, i.e., about 99.7% of the values lie within three standard deviations of the mean. The outlier ($> 3\sigma$) ratio of most tensors is lower than 0.5%, and the values of $> 6\sigma$ are extremely few in tensors. Therefore, normal values are relatively concentrated, indicating that we can quantize the normal values with a narrow range to enhance the resolution of quantization.

The more obvious observation is that the Max σ of the Transformer is larger than that of CNN by one order of magnitude. Some research [14, 43] shows that although the outliers are clipped for CNN models, the accuracy can still be restored to the original value with the retraining algorithm under ultra-low-bit precision, e.g., 4-bit. However, it is challenging for Transformer models, which have much more significant outliers. The state-of-the-art quantization works [18, 82] also demonstrate a similar observation and only can achieve the original accuracy with higher-precision quantization for large-scale Transformer models due to the outliers. Therefore, keeping the outlier without clipping will significantly benefit quantizing Transformer models.

2.2 Outlier Is Unaligned

The importance of outliers has attracted many research interests, which sparked several outlier-aware architectures, as depicted in Tbl. 1. OLAcel [61] and GOBO [85] are similar and exploit the coordinate list to indicate the location of outliers, which use high-precision (8-bit or 16-bit) quantization. BiScaled-DNN [39] and DRQ [75] employ block sparse index and bitmap, respectively.

Accelerator	Encoding	Aligned Memory?	GPU Compatible?
OLAccel [61]	Coordinate list	No	No
BiScaled-DNN [39]	Block sparse index	Aligned data Unaligned index	No
DRQ [75]	Binary mask map	Unaligned data Aligned index	No
GOBO [85]	Coordinate list	No	DRAM-only
OliVe (Ours)	Outlier-victim pair	Yes	Yes

Table 1: Comparison between existing outlier-aware accelerators and our proposed method OliVe.

BiScaled-DNN quantizes all values with the same bit-width but different scale factors for normal values and outliers, which are aligned. However, the extra index compressed in the block sparsity method is unaligned. On the contrary, DRQ’s bitmap is aligned, but data is stored by mixed and thus unaligned 4- & 8-bit values.

In summary, prior works design the outlier-aware architecture based on the sparsity of outliers, which leads to unaligned memory storage and accesses. More seriously, the indices of sparsity-based encoding and the outliers are separate. As such, they need the extra outlier controller to parse indices for the outliers and orchestrate the computation between normal values and outlier values. For example, the extra outlier controllers of GOBO and OLAccel count up to 55% and 71% overhead to the total area of the processing element (PE) array [61, 85]. The sparsity-based encoding for outliers is also **incompatible** with the memory sub-system of existing accelerators. For the GOBO design [85], it can only compress and decompress the memory at the DRAM level for GPU. This greatly limits the applicability of its proposed outlier-aware architecture.

Therefore, a more hardware-friendly and applicable outlier decoding/encoding method should be proposed to fit the outlier-aware quantization. Our proposed OliVe architecture is able to align memory accesses and is also compatible with existing accelerators based on the OVP mechanism.

2.3 Outlier and Victim Analysis

Generally, the sparsity-based encoding borrowed from DNN pruning is a straightforward and effective solution for sparse outliers. However, these works ignored that quantization is different from pruning. For pruning, the pruned zero values do not participate in the computation. As such, the pruning method has to compress

Pair Type	Normal-Normal	Outlier-Normal	Outlier-Outlier
BERT _{base} [19]	99.12%	0.84%	0.04%
BERT _{large} [19]	99.24%	0.71%	0.05%
GPT2-XL [66]	98.80%	1.14%	0.06%
OPT-6.7B [90]	99.33%	0.64%	0.03%

Table 2: The percentage of three types of pair.

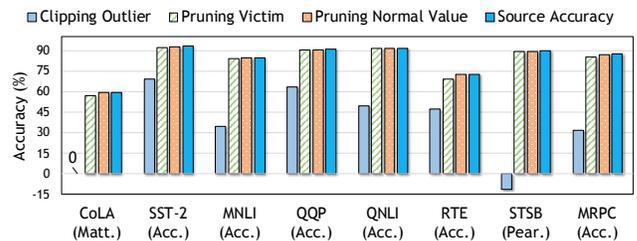


Figure 3: Accuracy comparison of multiple pruning methods.

the sparse values with sparsity-based encoding. For quantization, the quantized normal values are the majority and need computation. Naturally, the outlier values can exploit the normal values to achieve memory alignment instead of sparsity-based encoding.

As depicted in Fig. 1b in Sec. 1, we employ the insight of pruning but with a different perspective from prior works. The new method employs the **outlier-victim pair** (OVP) mechanism. We first prune some quantized low-precision normal values, which we call **victims**. These victims are adjacent to the outliers and make extra space for the high-precision outliers. Therefore, we can embed the outliers in their original location without explicit sparse indexing. That can avoid the complex indexing hardware and make it compatible with GPU. To align the memory, we distinguish the “right outlier” and “left outlier” according to their position in the pair. We assign a right victim for the left outlier (e.g., 17.6 in Fig. 1b) and a left victim for the right outlier (e.g., -98 in Fig. 1b).

The OVP mechanism is based on our observation of large Transformer models, including BERT-base [19], BERT-large [19], GPT2-XL [66], and OPT-6.7B [90]. We collect all tensors, calculate their standard variance σ , and divide the values into normal values ($< 3\sigma$) and outlier values ($> 3\sigma$) by the 3σ rule. We then pair every two adjacent values (no overlapping), which leads to three types: normal-normal pair, outlier-normal pair, and outlier-outlier pair, as shown in Tbl. 2. These three types have two normal values, one normal value and one outlier value, and two outlier values, respectively.

Tbl. 2 demonstrates that most (about 99%) pairs are normal-normal pairs, with only around 1% of outlier-normal pairs. Outlier-outlier pairs need to prune the smaller outlier in the pair. Fortunately, the outlier-outlier pairs only have an extremely low probability of less than 0.06% in all studied models. Therefore, the outlier distribution is extremely dispersed, and we can retain most outliers.

We also conducted the accuracy experiments with the BERT_{base} model [82] on the GLUE dataset [78], as depicted in Fig. 3. First, we clip the outliers to the 3σ , where clipping is the common method adopted by quantization. Then, we prune the victims and normal values to zero. The victims are adjacent to the outliers, and normal values are randomly pruned with the same amount as the outliers. We keep the rest values with full precision (FP32). Although such few outliers (about 1%) are clipped, as shown in Fig. 3 clipping outlier, the accuracy loss is unacceptable for the BERT model. The results emphasize the importance of outliers in Transformer-based model. For comparison, pruning random normal values has almost no accuracy loss than the source accuracy. The pruning of victim values only shows a negligible accuracy decrease than the pruning of normal values because the victims include some outliers due to

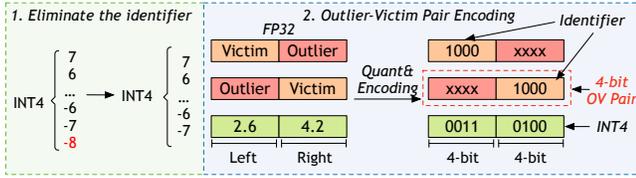


Figure 4: The 4-bit outlier-victim pair encoding.

the outlier-outlier pair and have specific locations corresponding to the adjacent outlier.

In summary, our analysis indicates that outliers are important while the victims are not, so that we can sacrifice victims to accommodate the outliers. This motivates us to design the hardware-friendly OVP mechanism that provides aligned outlier-aware quantization to accelerate the large Transformer models. In the next section, we will introduce the outlier-victim pair encoding design.

3 OUTLIER-VICTIM PAIR ENCODING

In this section, we present the details of outlier-victim pair (OVP) encoding that is *globally identical but locally distinguishable* for outlier and normal values. The OVP encoding can maintain globally aligned memory access and distinguish the outliers locally with ignorable overhead. For normal values, we can support multiple data types to fit the adaptive data type. For encoding outliers, we design an outlier-specific data type, adaptive bias float, abfloat, which can avoid range overlapping between normal values and outliers, thus improving the utilization ratio of the numerical representation space of outlier encoding. Finally, based on the OVP encoding, we propose a framework that can automatically select the outlier threshold for OVP encoding to determine a suitable ratio of the outlier-victim pair.

3.1 OVP Encoding Algorithm

Based on the previous pair-wise tensor value analysis, there are three pair types: normal-normal, outlier-normal, and outlier-outlier. For outlier-normal, the normal value in the pair will be pruned and turned into a victim. For outlier-outlier, we remain the large one

Algorithm 1: The 4-bit OVP encoding algorithm.

Input: Values, val_1, val_2 ; Outlier threshold, T .
Output: OVP encoding, out_1, out_2 .

```

1 def OVPairEncoding( $val_1, val_2, T$ ):
2   if  $val_1 > T$  and  $val_1 > val_2$  then
3      $out_1 = \text{OutlierQuantization}(val_1)$ ;
4      $out_2 = 1000_2$ ; // Outlier identifier.
5   else if  $val_2 > T$  then
6      $out_1 = 1000_2$ 
7      $out_2 = \text{OutlierQuantization}(val_2)$ ;
8   else
9      $out_1 = \text{NormalQuantization}(val_1)$ ;
10     $out_2 = \text{NormalQuantization}(val_2)$ ;
11  return  $out_1, out_2$ 

```

and prune the other. Then, we get the normal-normal pairs and outlier-victim pairs in the DNN tensors.

Outlier Identifier. To distinguish from the normal-normal pair, we need a special identifier for the outlier-victim pair. And this distinct identifier cannot appear in the normal-normal pair, which means we need to eliminate one number in the representation of normal values. For example, as shown in Fig. 4, we employ the signed int4 (4-bit integer) for the normal value quantization. The original int4 can represent the integers in the range of $[-8, 7]$, where 1000_2 represents the value of -8 . First, we make 1000_2 the outlier identifier and remove the value of 1000_2 from int4, whose encoding range becomes $[-7, 7]$. Second, we quantize the outlier-victim pairs with 4-bit OVP encoding. We set the victims with the outlier identifier 1000_2 and quantize the outlier with the outlier-specific data type (Sec. 3.3). Naturally, there are two types of OV pair, i.e., left outlier (O-V) and right outlier (V-O) pair. Due to the distinct outlier identifier design, we can implicitly distinguish them without using an extra index bit (Sec. 4.2).

Algo. 1 shows the 4-bit OVP encoding algorithm, which needs to read two values simultaneously, where the requirement is very easy to meet. For the hardware implementation, we can add a buffer for the encoder. Also, the OVP encoder can be implemented by embedding in the quantization unit with ignorable overheads. For the software implementation, we can make a thread handle two values simultaneously. As a result, the encoding algorithm can be implemented efficiently in both hardware and software, which we describe more details later.

3.2 Data Type for Normal Values

For normal values, we build upon prior work [32], which can support multiple data types, including int4, flint4 (4-bit flint), and int8, as shown in Tbl. 3. The int4 type is one of the most widely used data types for 4-bit quantization with integers in the value range of $[-7, 7]$. The flint4 type is proposed by prior work ANT [32], which has shown that selecting the data type according to a tensor’s distribution achieves the state-of-the-art performance and accuracy.

Based on the above insights, we also adopt the mixed data types to quantize normal values in our OVP pair encoding. For flint4, we use the same binary value of 1000_2 as the outlier identifier. Specifically, 1000_2 of flint4 corresponds to -0 , which is not used in the original design. In other words, our OVP encoding seamlessly works for flint4 without wasting any number representations. We use the original flint4 encoding algorithm [32] to quantize normal values.

Moreover, the OVP encoding can be generally extended to higher-precision quantization, such as the 8-bit. Similarly, the 8-bit normal

Data Type	Values	Outlier Identifier
int4	$0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \pm 7$	1000_2 (-8)
flint4 [32]	$0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 8, \pm 16$	1000_2 (-0)
int8	$0, \pm 1, \pm 2, \dots, \pm 126, \pm 127$	10000000_2 (-128)

Table 3: Data types for normal values of OVP encoding.

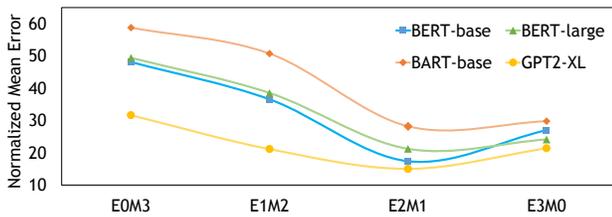


Figure 5: The rounding error of the largest outliers quantized with different data types. Experiments were conducted on BERT-base, BERT-large, BART-base, and GPT2-XL.

value also needs to eliminate one number. For instance, `int8` can represent $[-128, 127]$ integers, and we can make 10000000_2 the outlier identifier for `int8` and narrow its range to $[-127, 127]$. Similarly, the encoding algorithm can easily extend to read two 8-bit elements simultaneously.

3.3 Data Type for Outliers: Abfloat

Next, we quantize outliers using the outlier-specific data type. The large outliers usually have a wide range, for which we use float-based data to quantize. We propose a data type called **adaptive biased float**, `abfloat` in short. The key idea is that by adding a proper bias to the exponent, all encoded values can skip the interval where normal values lie and provide more range for outliers.

Float-to-Fixed Conversion. To accommodate the normal values and avoid fractions, we first convert the floating-point encoding to the fixed point with an exponent. Also, the fixed point is friendly to the hardware implementation and has a lower overhead than the floating point. We transform the the floating point to fixed point with the following equation,

$$\text{sign} \times (1 \ll \text{mb} + \text{mantissa}) \ll (\text{exponent} + \text{bias}), \quad (2)$$

where `mb` is the mantissa bit-width. Therefore, this fixed-point encoding scheme is more friendly and efficient for hardware implementation, as it only involves shift operations. Tbl. 4 shows the example of fixed-point `E2M1` data type.

Adaptive Bias. Obviously, Tbl. 3 and Tbl. 4 show that the range of fixed-point `abfloat` overlaps with the normal values. For example, `int4` and `E2M1` contain the same numbers, 3, 4, and 6. Another example is that `flint4` and `E2M1` have almost the same number range except for 24. Therefore, we need the adaptive bias to adjust the range of `abfloat`. For example, we set `bias = 2` for `E2M1`, whose real values will be extended to $\{12, \dots, 96\}$, which is complementary with the `int4` normal value. Similarly, we set `bias = 3`

Binary	Exponent	Integer	Real Value
<u>000</u>	0	0	0
<u>001</u>	0	3	$3 \times 2^0 = 3$
<u>01x</u>	1	2, 3	$2 \times 2^1 = 4, 3 \times 2^1 = 6$
<u>10x</u>	2	2, 3	$2 \times 2^2 = 8, 3 \times 2^2 = 12$
<u>11x</u>	3	2, 3	$2 \times 2^3 = 16, 3 \times 2^3 = 24$

Table 4: The 3-bit unsigned `E2M1`, which means two bits for exponent and one bit for mantissa, with `bias = 0`.

Algorithm 2: The `abfloat` encoding algorithm.

Input: Element e ; Bias, b ;
Output: Quantized Element q ;

```

1 def AbfloatQuant( $e, b$ ):
    // Get exponent and base integer.
2      $exp = \lfloor \log_2(\text{abs}(e)) \rfloor - 1$ ;
3      $base\_int = \text{Round}[e/2^{exp}]$ ;
4     if  $base\_int == 4$  then
5          $exp = exp + 1$ ;
6          $base\_int = base\_int - 2$ ;
    // Encoded as abfloat data type.
7      $exp = exp - b$ ;
8      $base\_int = base\_int \& 1$ ;
9      $unsigned\_q = \text{Concat}(exp, base\_int)$ ;
10     $q = \text{Concat}(e < 0, unsigned\_q)$ 
11    return  $q$ 

```

and extend range to $\{24, \dots, 192\}$ for `flint4` data type. We design a new decoder and instruction to implement adaptive bias in accelerators for the `abfloat` (Sec. 4.2).

E2M1 Abfloat. The 4-bit signed float has four possible configurations of exponent and mantissa: `E0M3`, `E1M2`, `E2M1`, and `E3M0`. They have different ranges and precisions. We conduct the following experiments to choose the most appropriate configuration as the final outlier-specific data type. To accommodate the broad range of outlier values, we quantize the largest outlier values (i.e., $\text{Max } \sigma$ in Fig. 2) in Transformer models using all `abfloat` types. Then, we collect the average absolute error, as shown in Fig. 5. We found that `E2M1` gives the least error in all tests, which provides both a large enough range and a certain degree of precision, and it also presents the best results in our subsequent evaluations. Similarly, we adopt signed `E4M3` for 8-bit `abfloat`.

Algo. 2 shows in detail how an element is encoded as `abfloat`. The outlier encoding is an element-wise function, which can be implemented on software and hardware efficiently. Outlier encoding should also eliminate the outlier identifier. Otherwise, the decoder cannot distinguish the outlier-victim pair. `Abfloat` has two zero numbers: `1000` (-0) and `0000` (0). Therefore, we disable the `1000` and `0000` for outlier values to avoid conflict with the outlier identifier.

3.4 Quantization Framework

We now apply OVP (outlier-victim pair) encoding for quantizing Transformer models. To decide the scale factor (i.e., outlier-victim threshold), we embed the OVP encoding with the existing mean squared error (MSE) minimization algorithm, which is commonly used by many quantization works [4, 6, 88]. The OVP-based quantization algorithm determines the threshold for distinguishing outliers and normal values. On one hand, a small threshold would lead to more outlier-victim pairs, which could potentially minimize the quantization error (i.e., MSE). On the other hand, it also increases the ratio of outlier-outlier pairs, where both values are outliers in the pair. If there are too many such outlier-outlier pairs, the MSE would increase owing to the pruning of outliers. Thus, we need to control the ratio of outlier-outlier pairs for better accuracy.

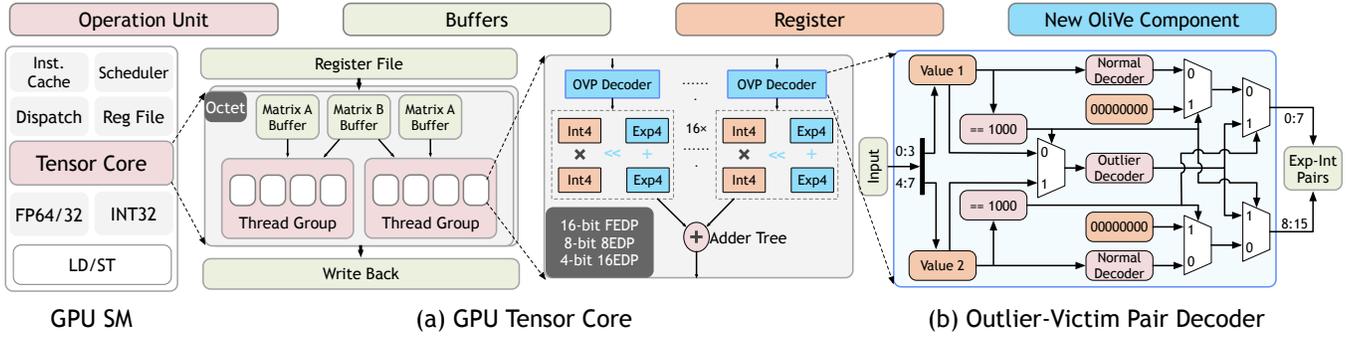


Figure 6: OliVe integration on GPU tensor cores (a), which only requires a set of lightweight OVP decoder (b).

In our work, we target the post-training quantization (PTQ) [57], which does not require retraining and hence is best suitable for large models as their trainings are expensive. However, we still need to use one batch of data from the **training set** for the scale factor selection. Intuitively, inspired by the 3σ rule, we take 3σ as the initial scale factor. Then the algorithm will search for the best scale factor with the smallest MSE within a specific range of this baseline, which shows good results in our evaluations. For quantization-aware training (QAT) [57], we can get a suitable scale factor by retraining it with the straight-through estimator (STE) [5].

4 OLIVE ARCHITECTURE

This section presents how to integrate OliVe in GPU and output-stationary systolic array architecture. We then present the hardware decoder for the aforementioned outlier-victim pair encoding and outlier data type. On these architectures, our proposed OliVe architecture can directly support the mixed precision [60, 72] and mixed data type [60, 72], which are efficient for quantizing DNN tensors that have different importance and distribution.

4.1 GPU Tensor Core

We first describe how to integrate the OliVe design into the tensor core architecture of GPU in the Fig. 6a. We employ Turing architecture [59] as our baseline GPU, which has 68 streaming multiprocessors (SMs), and each SM has eight tensor cores (544 in total), as shown in Tbl. 5. According to the modeling of prior work [67], each tensor core has two octets, which have eight FEDPs (four-element dot product). As such, there are $68 \times 8 \times 2 \times 8 \times 4 = 34,816$ 16-bit float multipliers. The Turing architecture can originally support mixed-precision computation. For example, the RTX 2080Ti GPU with Turing architecture [59] provides 107.6, 215.2, and 430.3 TOPS (tera operations per second) for 16-bit float, 8-bit int, and 4-bit int, respectively. Therefore, we assume that the tensor core can simultaneously support 8-bit 8EDP (eight-element dot product) and 4-bit 16EDP (16-element dot product), as shown in Fig. 6a.

Architecture	SM	TC	16-bit Unit	8-bit Unit	4-bit Unit
Turing [59]	68	544	34,816	69,632	139,264

Table 5: The Turing GPU architecture.

We can easily embed our proposed OliVe architecture in GPU, which adopts the SIMD architecture. We first put the 4-bit outlier-victim pair decoders (Fig. 6b) for each 16EDP. To support the new OliVe data types, we add an adder and a shifter for each 16EDP. Similarly, we also design the 8-bit decoder for the 8EDP units.

4.2 Decoders

Outlier-Victim Pair Decoder. To support outlier-victim pair decoding, we design a new decoder that can be easily embedded in existing accelerators. As shown in Fig. 6b, the decoder reads 1 byte, which is the smallest addressable memory unit in many architectures, and exactly one value pair. Then, the decoder transforms the outlier identifier 1000_2 to 0 and decodes the outlier value with the outlier decoder. To accommodate the computation of the outlier abfloat values, the decoder will generate an exponent-integer pair. Therefore, the decoder needs to append a 0000_2 as the exponent number for the normal int4 data type. For flint4, we exploit its original decoder [32] to get the exponent-integer pair.

Outlier Decoder. The above OVP decoder contains an outlier decoder for outlier values with the E2M1 abfloat data type. Fig. 7 shows the details of the 4-bit abfloat decoder design. For a 4-bit E2M1 abfloat number $x = (b_2b_1b_0)_2$, following equations decode exponent and integer:

$$\text{exponent} = \text{bias} + (b_2b_1)_2$$

$$\text{integer} = \begin{cases} 0 & \text{if } x = 000_2 \\ (1b_0)_2 & \text{otherwise} \end{cases}$$

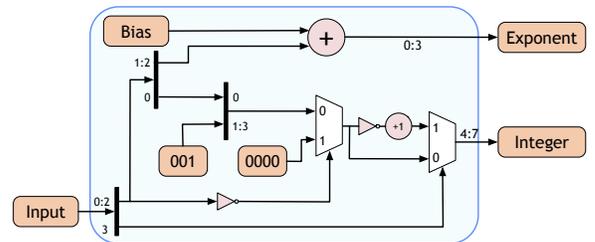


Figure 7: The 4-bit abfloat decoder for outlier values.

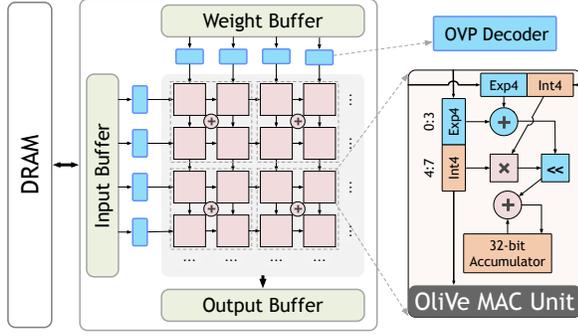


Figure 8: OliVe integration on systolic array.

For example, when the bias is 2, a number 0101_2 is 48_{10} , since its exponent is $2_{10} + 10_2 = 4_{10}$ and base integer is $11_2 = 3_{10}$. Therefore, its real value is $3 \ll 4 = 48$.

Similarly, we also design and implement the 8-bit outlier-victim pair decoder and the E4M3 abf10at outlier decoder, which are straightforward extensions of 4-bit instances. As such, we do not present their details due to the limited space.

4.3 Systolic Array

The systolic array (SA) integration is shown in Fig. 8. SA uses the same outlier-victim pair decoder design (Fig. 6b) as GPU, which shows the wide applicability of our design. But, unlike GPU, we only place the decoders along the borderlines, which can save most decoders. For example, if the array size is $n \times m$, we only need $n + m$ instead of $n \times m$ decoders. That is one advantage of SA over the GPU's SIMD architecture. Our proposed OliVe-based data type can also support the systolic array processing element (PE) with an extra adder and shifter. We add an extra adder for every four PEs to support high-precision quantization, e.g., int8.

4.4 OliVe MAC unit

After decoding for outlier and normal values, they are all transformed into unified exponent-integer pairs. To support the decoded exponent-integer pair computation, we need to add a shifter and an adder for the fixed-point MAC (multiply and accumulation) unit, as shown in Fig. 8 and the unit of Fig. 6 4-bit 16EDP. For example, we have two exponent-integer pairs $\langle a, b \rangle$ and $\langle c, d \rangle$, where a and c are exponents, b and d are integers, and $\langle a, b \rangle$ represents:

$$\langle a, b \rangle = b \ll a$$

Then, we can get the result:

$$\begin{aligned} & \langle a, b \rangle \times \langle c, d \rangle \\ &= (b \times d) \ll (a + c) \\ &= \langle a + c, b \times d \rangle \end{aligned}$$

Note that the final result can store with a 32-bit int.

4.5 Mixed Precision

As mentioned in Sec. 3, OliVe quantization can support the int8 for normal values and E4M3 abf10at for outlier values. Therefore,

we propose the mixed-precision processing element (PE) for the higher precision data types.

8-bit Int. For the GPU tensor core architecture, it is originally designed with mixed-precision computation. For the systolic array, our architecture naturally supports 8-bit computation with four 4-bit PEs [72]. For an int8 number x , the higher 4 bits and the lower 4 bits can be split into two 4-bit numbers h and l , and the x can be represented by:

$$x = (h_x \ll 4) + l_x = \langle \langle 4, h_x \rangle, \langle +, 0, l_x \rangle \rangle$$

We then can multiply two int8 numbers of x and y :

$$\begin{aligned} x \times y &= \langle \langle 4, h_x \rangle \times \langle 4, h_y \rangle + \langle 4, h_x \rangle \times \langle 0, l_y \rangle \\ &\quad + \langle 0, l_x \rangle \times \langle 4, h_y \rangle + \langle 0, l_x \rangle \times \langle 0, l_y \rangle \end{aligned}$$

$\underbrace{\hspace{10em}}_{PE0} \quad \underbrace{\hspace{10em}}_{PE1}$
 $\underbrace{\hspace{10em}}_{PE2} \quad \underbrace{\hspace{10em}}_{PE3}$

Therefore, we can use four 4-bit PEs to calculate the above four multiplications and accumulate the products to get the final product value of $x \times y$.

8-bit Abfloat Similarly, multiplication of 8-bit abfloat can be supported using the same approach. For an 8-bit abfloat number z , it is first decoded into an exponent e_z and an integer i_z . For i_z , we similarly split it into $i_z = (h_z \ll 4) + l_z$, then $z = \langle \langle 4 + e_z, h_z \rangle + \langle +, e_z, l_z \rangle \rangle$. Hence the same method can be used to perform 8-bit abfloat multiplication with four 4-bit PEs, where the abfloat has an extra e_z than int8.

In the most extreme case, two outliers with abfloat may be multiplied together. Because we adopt the 32-bit int as the accumulator, the maximum multiplicand should not be over $\sqrt{2^{31} - 1}$. Therefore, for the outlier value with the abfloat type, we will clip the absolute value of the outlier within $2^{15} < \sqrt{2^{31} - 1}$ to avoid the overflow for the int32 accumulators. Our experiments show that the outlier values of the Transformer models are much smaller than 2^{15} . Specifically, 2^{15} is about 768σ after normalization and quantization. As shown in Fig. 2, the maximum value of outliers does not exceed 325σ . Thus, we observe that no outlier is truncated in practice.

4.6 Instruction Set

For 4-bit tensor cores, the Turing GPU architecture adopts the instruction `mma.s32.s4.s4.s32`. These four operands are matrices D (int32), A (int4), B (int4), and C (int32), and $D = A \times B + C$. To support the OVP-based computation on GPU, we design a new instruction called `mmaovp`:

$$\underbrace{\text{mmaovp}}_{\text{OVP-MMA}} . \text{s32} . \underbrace{\text{ovpi4}}_{\text{int4}} . \underbrace{\text{ovpf4}}_{\text{flint4}} . \text{s32} . \underbrace{\text{s4}}_{\text{bias}} .$$

Moreover, because of the memory-aligned design of the data type, OliVe maintains the original programming interface for GPUs. We can replace the original int-based instruction with OVP-based instruction (e.g., `mmaovp`) to easily construct the OVP-supported DNN quantization framework. Therefore, our OliVe framework has comprehensive and practical applicability, which is the most significant advantage of OliVe.

5 EVALUATION

In this section, we evaluate the LLM’s accuracy with OliVe quantization. We also demonstrate OliVe’s area overhead, speedup, and energy efficiency on GPU and systolic array, respectively.

5.1 Methodology

Framework and Evaluation Models. To evaluate our OliVe quantization framework, we implement it in Pytorch [62]. We evaluate BERT-base [19], BERT-large [19], and BART-base [49], the three most commonly used language models, on eight datasets of the GLUE benchmark [78]. In addition, we evaluate BERT-base [19] and BART-base [49] on the summarization tasks SQuAD v1.1 and SQuAD v2.0 [68]. To valid our quantization framework on large language models, we also evaluate GPT2-XL [66], BLOOM-7B1 [70], and OPT-6.7B [90] on Wikitext103 [83] and C4 [20] datasets. For all models mentioned above, we use state-of-the-art checkpoints from the huggingface repositories [55].

Quantization Baselines. We compare OliVe with existing quantization works, including GOBO [85], Outlier Suppression [82], Q8BERT [86], and ANT [32]. Outlier suppression [82] is the state-of-the-art Transformer quantization work. GOBO [85] is also an outlier-aware quantization work. Q8BERT [86] is a method for quantizing GEMM operations to 8-bit. ANT [32] is a hardware-friendly quantization framework that achieves state-of-the-art results in both performance and accuracy.

Accelerator Baselines. We compare the performance and energy of OliVe against five DNN quantization accelerators, including OLAccel [61], AdaptivFloat [76] (shorted as AdaFloat), GOBO [61], ANT [32], and original int8 tensor cores in GPU [59]. OLAccel [61] first proposed the outlier-aware quantization architecture for CNNs. We extend OLAccel to the Transformer-based models with element-wise mixed-precision weight and activation quantization. AdaFloat [76] extends the float type with a tensor-wise exponent bias. GOBO [85] is similar to OLAccel, but only supports the weight quantization for Transformer-based networks.

OliVe Implementation. We implement our decoder in Verilog RTL and synthesize it with Synopsys design compiler [47] with a 22 nm TSMC technology library to estimate its area, latency, and power. We use CACTI [56] to estimate the area and power of on-chip memories. We integrate OliVe into GPU and hardware accelerator for the end-to-end performance and energy evaluation.

For the GPU integration and evaluation, we modify and extend GPGPU-Sim 4.0 [3] and AccelSim [45] with the configuration of NVIDIA 2080 Ti architecture. We use AccelWattch [46], GPUWattch [48], and CACTI [56] for the energy estimation. The majority of Transformer layers are matrix multiplication operations. For GEMM implementation on the tensor core, we use CUTLASS [44], which is NVIDIA’s open-source implementation.

For the accelerator evaluation, we compare AdaFloat, OLAccel and ANT with OliVe. We develop a cycle-level simulator to estimate the overall performance of OliVe based on DnnWeaver [71]. Although DnnWeaver [71] is a FPGA tool set, prior DNN quantization accelerators, which include the BitFusion [72], and ANT [32], have extended its frontend to add the ASIC performance and energy simulation. As OliVe does not redesign the baseline accelerator

Method	Algorithm	CoLA	SST-2	MNLI	QQP	MRPC
BERT _{base}	32-bit	59.60	93.35	84.94	90.91	87.75
Ours	4-bit PTQ	59.30	92.43	84.10	90.36	87.99
ANT	4-bit QAT	53.91	92.43	83.45	-	-
ANT	4-bit PTQ	42.90	90.48	73.36	78.04	68.87
OS	4-bit QAT	50.56	91.86	83.05	90.33	84.31
OS	6-bit PTQ	54.40	91.86	82.02	88.94	83.33
Q8	8-bit QAT	58.48	92.24	-	-	-
BERT _{large}	32-bit	63.35	93.46	86.65	91.07	87.99
Ours	4-bit PTQ	63.99	92.89	84.89	90.14	86.52
BART _{base}	32-bit	56.32	93.35	86.45	91.34	87.50
Ours	4-bit PTQ	54.30	92.89	85.33	91.23	86.76
OS	4-bit QAT	50.83	92.43	84.57	90.93	87.01
OS	6-bit PTQ	44.51	90.94	82.98	88.45	80.88

Table 6: Results on GLUE datasets. Q8 and OS are Q8BERT [86] and outlier suppression [82] for short, respectively. Prior works do not report results in BERT_{large} so we only compare against the original full-precision model.

architecture, we can directly embed new OliVe-related instructions and data format in the simulator without breaking the original simulation flow. In other words, we have used and modified the open-sourced implementaions of BitFusion [72, 73], and ANT [32, 33].

5.2 Accuracy Results

We first evaluate the accuracy of OliVe quantization framework on different tasks and datasets, which is the prerequisite for applying it to reduce the inference cost of large language models (LLMs).

GLUE Dataset. We evaluate BERT-base [19], BERT-large [19] and BART-base [49] on eight datasets of GLUE benchmark, but due to space limitation, we only show the results on CoLA, SST-2, MNLI, QQP and MRPC datasets in Fig. 6. For the BERT-base model, our 4-bit PTQ method accuracy drop less than 1% compared to the original full precision model on all eight datasets and outperforms all studied methods including 4-bit, 6-bit, and 8-bit PTQ and QAT methods. Since GOBO [85] only quantizes weights, we use the same method to compare with it and the result is shown in Tbl. 7. Our method also outperforms the GOBO under the weight-only quantization setting. In addition, we evaluate the BERT-large model, which is evaluated by few prior quantization works due to the larger number of parameters and hence much more challenging compared to BERT-base. The results in Tbl. 6 show the accuracy loss for BERT-large is around 1% on the five presented datasets and similar results are found on other datasets. For the BART-base model, our 4-bit PTQ results in Tbl. 6 show around 2% accuracy loss compared to the accuracy of original full-precision in all datasets. In the above evaluation, our 4-bit PTQ results are better than all the PTQ and most of the QAT results reported by prior works.

SQuAD Dataset. We also evaluate the accuracy of OliVe quantization on summarization task SQuAD [68], which is more challenging than the previous GLUE dataset. Tbl. 8 shows the results on SQuAD v1.1 and SQuAD v2.0 datasets. On both datasets, our

4-bit PTQ method obtains a less than 2% accuracy loss on the BERT-base model and around 3% accuracy loss on the BART-base model, which is better than the 6-bit PTQ method of the state-of-the-art quantization work outlier suppression.

Large Language Models. We evaluate the accuracy of Olive for LLMs under the PTQ setting. LLMs’ inference is challenging as it requires significant memory, which makes their retraining even more resource-consuming. Thus, the PTQ method without retraining is more desirable than the QAT method for LLMs.

The recent work [18] has shown that the int8 quantization has a significant accuracy drop when the number of parameters of the OPT model grows to 6.7B. As shown in Tbl. 9, our 8-bit PTQ method has only a negligible perplexity increase on OPT-6.7B (lower is better), while the accuracy of the int8-based quantization method has a significant degradation and is worse than our 4-bit PTQ method on the C4 dataset. On GPT2-XL and BLOOM-7B1 models, our 8-bit PTQ method essentially achieves the original perplexity, and the 4-bit PTQ method achieves the performance close to int8. For comparison, the accuracy results of int4 and 4-bit ANT are unacceptable (10-1000× worse than FP32 model).

To summarize, our Olive quantization framework pushes the limit of 4-bit quantization to a new state-of-the-art, as it is able to achieve nearly original accuracy for the commonly used language models including BERT-base, BERT-large, and BART-base on most datasets. Moreover, Olive also gives the state-of-the-art results of 4-bit and 8-bit quantization on large language models like GPT2-XL, BLOOM-7B1, and OPT-6.7B.

5.3 GPU Performance and Energy

We evaluate LLMs on the GPU simulator, where the batch size is set to 2 for GPT-like models and 16 for BERT-like models. For Olive, 4-bit quantization can limit the loss to a relatively small error range. GOBO [85] can achieve the original accuracy of all models but has a significant overhead on compressing weight in DRAM. Note that GOBO only quantizes the weight tensors and computes with FP16. We implemented GOBO’s memory organization in the GPU. For ANT [32], we make all models close to the original accuracy or perplexity by mixed precision (BERT-like models [19, 49] with < 1% loss and GPT-like models [66, 70, 90] with < 3 perplexity) with the PTQ setting. In addition, we also compare the original int8 of GPU, which has unacceptable accuracy loss, just for performance and

Method	Bits	MNLI	STSB(Pear.)
BERT _{base}	32	84.94	89.70
Ours (weights only)	4	84.75	89.62
GOBO*(weights only)	4	84.45	88.33

Table 7: Comparison with GOBO on the MNLI and STSB dataset. *The accuracy of our GOBO implementation slightly differs from the number reported in the original paper [85].

Method	Bits	SQuAD v1.1	SQuAD v2.0
BERT _{base}	32	88.28/80.82	77.34/73.60
Ours	4	86.38/78.16	75.90/72.08
Outlier Suppression	6	84.48/75.53	74.69/70.55
BART _{base}	32	91.63/84.79	80.82/77.41
Ours	4	88.15/79.87	77.37/73.69
Outlier Suppression	6	83.68/75.34	74.44/70.36

Table 8: PTQ results on SQuAD datasets.

energy comparison to GPU baseline. We compare the GPU architecture integrated with our Olive design against various baselines. The performance and energy results are shown in Fig. 9.

Performance. Fig. 9a compares the speedup values of different quantization methods on GPUs. Olive achieves the best performance and has higher speedups on the larger language models than GOBO. Due to the FP16 computation and weight-only quantization, GOBO [85] achieves the lowest performance among all studied designs. In contrast, Olive quantizes both activation and weight to low bits and does not increase the memory access overhead. This avoids performance degradation when the number of parameters increases. The PTQ seriously degrades the accuracy of ANT [32] as it cannot handle outliers. In ANT, 80% of layers ends up using int8 quantization so the performance results between ANT and int8 are close. On average, Olive achieves 4.5×, 2.7×, and 2.4× speedup values over GOBO, int8, and ANT, respectively.

Energy. Fig. 9b shows the normalized energy comparison of different designs, including constant, static, and dynamic power. And the dynamic power includes DRAM, L2 cache, L1 data cache, shared memory, register file, and processing elements (CUDA core and tensor core). The L1 contains the sum of the L1 cache and shared memory energy. Olive has the lowest energy due to the aligned 4-bit design and GPU compatibility. Due to the worse accuracy result of the mixed precision, ANT is also close to int8 on the energy. Overall, 4-bit Olive is very hardware-friendly so that it can take full advantage of the energy savings with lower bits. Olive achieves average 4.0×, 2.3×, and 2.0× energy reduction over GOBO, int8, and ANT, respectively.

Area. To measure the overhead of Olive decoder on the GPU, we scale the Olive decoder to 12 nm, which is the same manufacturing process as RTX 2080 Ti [59] and calculate the tile area. According to Tbl. 5, there are 139,264 4-bit decoders and 69,632 8-bit decoders

Method	GPT2-XL		BLOOM-7B1		OPT-6.7B	
	Wiki	C4	Wiki	C4	Wiki	C4
FP32	17.48	16.30	13.05	14.94	22.14	10.63
int8	18.29	17.35	14.04	16.18	37.45	74.30
8-bit Olive	17.49	16.37	13.13	15.04	22.34	10.73
int4	1E+4	9E+3	3E+6	9E+6	5E+2	1E+2
4-bit ANT	27.79	27.35	23.22	27.36	4E+4	4E+4
4-bit Olive	19.11	18.08	15.16	17.18	55.44	32.41

Table 9: PTQ results on large language models. The accuracy metric is perplexity, and lower is better.

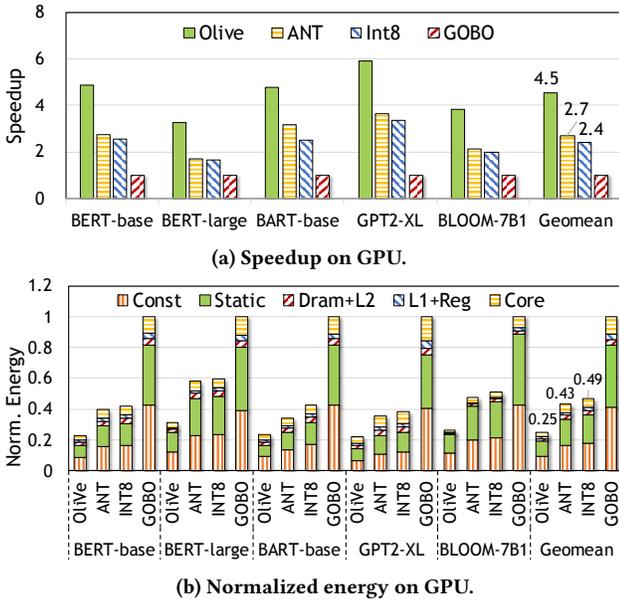


Figure 9: Comparison of four different designs on GPU.

on the GPU die and their area is shown in Tbl. 10. Since the GPU die size of RTX 2080 Ti is 754 mm^2 , the 4-bit decoder and 8-bit decoder only account for 0.250% and 0.166% of the entire GPU area respectively, which we believe is a tiny and worthy overhead.

5.4 Accelerator Performance and Energy

As explained in Sec. 5.1, we also integrate OliVe to the systolic-array-based hardware accelerator and compare its performance and energy against existing designs of ANT [32], OLAcel [61], and AdaFloat [76]. Similar to its GPU implementation, ANT is a mixed-precision design. Since AdaFloat does not support mixed precision, we only provide the 8-bit quantization results. All accelerators can achieve close to original accuracy for all Transformer models.

Performance. As shown in Fig. 10a, OliVe has the most significant advantage in latency speedup. Owing to its inability to deal with outliers, the performance of ANT is similar to OLAcel on most models. The speedup values of OliVe are very similar on all models, and they do not change with the increasing number of model parameters. On average, OliVe achieves 4.8x, 3.8x, and 3.7x speedup value over AdaFloat, OLAcel, and ANT, respectively.

Energy. Fig. 10b shows the normalized energy consumption of different designs composed of static and dynamic energy (DRAM, on-chip buffer, and core). OliVe has the lowest energy consumption. Compared to OLAcel, OliVe has a significant advantage in terms

Component	Number	Area (mm^2)	Area Ratio
4-bit Decoder ($13.53\mu\text{m}^2$)	139,264	1.88	0.250%
8-bit Decoder ($18.00\mu\text{m}^2$)	69,632	1.25	0.166%

Table 10: The area of OliVe decoder on RTX 2080 Ti.

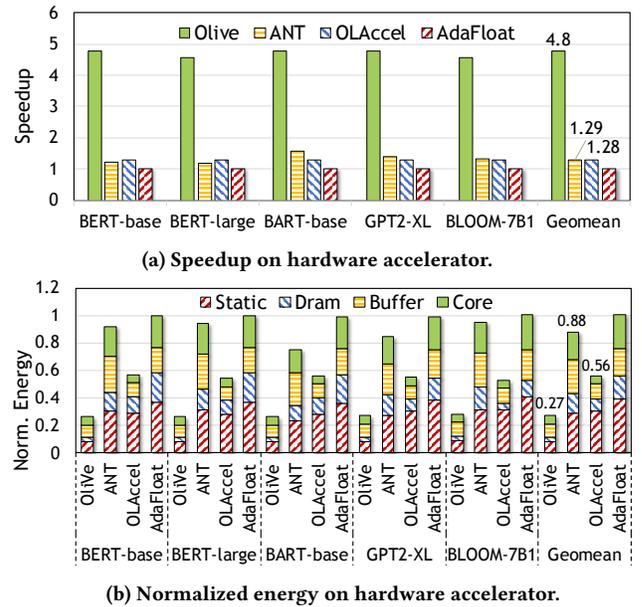


Figure 10: Comparison of different designs on accelerators.

of static and DRAM. Worse mixed-precision results increase ANT energy consumption, which is even close to AdaFloat in BLOOM-7B1 model. On average, OliVe achieves 3.7x, 2.1x, and 3.3x energy reduction over AdaFloat, OLAcel, and ANT, respectively.

Area. Tbl. 11 shows the area breakdown of OliVe-based systolic array architecture under 22 nm process. In this scenario, the 4-bit and 8-bit decoders introduce about 2.2% and 1.5% overhead of the core area, respectively, which is inconsiderable compared to the area of PEs in the array. Considering on-chip memory structures, the overall area overhead would be even smaller. In addition, we also scale other accelerators to 22 nm using DeepScaleTool [69] and get similar results to those numbers. Note that we implement all accelerators with a similar area size. The small area overhead of our OliVe directly benefits from the carefully-designed outlier-victim pair (OVP) encoding.

6 RELATED WORK AND DISCUSSION

This section presents and discusses research on DNN acceleration and compression. With the growing computation requirements of DNN models, it is crucial to design the algorithms and architecture to accelerate DNN models. Various compression methods, such

Component	Number	Area (mm^2)	Area Ratio
4-bit Decoder ($37.22\mu\text{m}^2$)	128	0.00476	2.2%
8-bit Decoder ($49.50\mu\text{m}^2$)	64	0.00317	1.5%
4-bit PE ($50.01\mu\text{m}^2$)	4096	0.205	96.3%

Table 11: Area breakdown of OliVe under 22 nm process.

as pruning and quantization, have been proposed to exploit the redundancy property of DNNs.

DNN Acceleration. In the past few years, various architectures [10, 12, 13, 23, 25, 34, 35, 51, 63, 64, 87, 96] have been proposed to match the computation characteristics of DNN models. To accelerate the DNN system, most optimizations focus on compilation [11, 40, 91, 92, 95, 97] and scheduling [8, 9, 15–17, 31, 52–54, 84].

The DNN acceleration highly relies on the performance of matrix multiplication. Therefore, several works focus on improving data reuse and simplifying control logic through a tailored dataflow architecture for matrix multiplication [10, 12, 25, 34, 35, 41, 63, 64, 87, 96]. TPU [41] introduces a highly optimized dataflow architecture that efficiently reuses data across multiple computation stages. Modern GPUs [60] now incorporate matrix multiplication accelerators, such as tensor core, optimized for SIMD operations to enhance DNN workload acceleration further.

Pruning. Pruning means removing a portion of weight, input, or output of DNN layers, resulting in a sparse model with reduced model size. However, a significant reduction leads to irregular memory accesses, which are negative for the acceleration of inference and training. To address this issue, researchers propose several sparse optimizations in algorithms and hardware architectures to reduce inefficient computation [2, 26–29, 36, 64, 65, 80, 89, 94, 98]. In addition, a sparse tensor core is introduced in NVIDIA Ampere GPU architecture [1] to support the 2:4 structured sparsity.

Quantization. Quantization is another effective and efficient way to reduce the DNN model size and computation burden. There are two popular quantization methods, i.e., quantization-aware training (QAT) [38, 50, 81, 99] and post-training quantization (PTQ) [30, 35, 38, 81]. QAT allows the model to adapt to quantization noise by retraining. PTQ is very effective to implement since it converts the original FP32 model directly into a lower-bit model without the training data and pipeline. Thus, PTQ is more feasible for language models at billion scales.

By quantizing data to low bit-width, quantization accelerators can significantly reduce memory bandwidth requirements and increase the computation speed. BitFusion [72] combines the low-bit PEs to support different bit-width quantization. OLAcel [61] utilizes 16-bit MAC to the first layer and 4-bit MAC to other layers. DRQ [75] quantizes data in sensitive and insensitive areas with different precision, which is similar to outlier-aware quantization. GOBO [85] is an accelerator that takes advantage of outlier-aware quantization, which quantizes the outliers of weights with higher precision. However, the outlier-aware quantization accelerators mentioned above have unaligned memory accesses, resulting in additional overhead and a limited computing speed. ANT [32] proposes a fixed-length adaptive quantization framework but only takes the distribution of tensors into account and ignores the importance of outliers. In contrast, our proposed novel Olive quantization framework can handle outlier values in a memory-aligned and hardware-friendly way.

AdaptivFloat [76] is similar to abfloat in adding a bias to the exponent, but the motivations and how the bias is determined are different. AdaptivFloat is to adapt to the dynamic ranges of different layers and calculates the optimal bias at a layer granularity using its

algorithm. Our abfloat is to make full use of the encoding range, so it simply adds a uniform bias to all encoding values to skip the range of normal values, which is simpler to implement.

GPU Architecture. NVIDIA has been updating its new generations of GPUs, e.g., Ampere architecture [1], which adds the sparse tensor core for structured sparsity in DNNs and compute data compression to increase the memory access bandwidth. The structured sparsity for tensor cores is orthogonal to our proposed quantization as our element-wise quantization does not affect (sparse) tensor core dataflow. Ampere GPU's compute data compression can compress zero values and similar bytes in DRAM and L2 cache. As such, it is lossless and therefore general-purpose. It is also transparent and orthogonal to Olive, which does not modify the memory system. In contrast, prior quantization work [85] perform compression at the DRAM-level, which could be impacted by the data compression in Ampere GPUs.

On the other hand, DNN quantization is a lossy compression. We believe the strictly lossless compression would have limited benefits for DNN quantization. Thus, our work could complement Ampere's current compute data compression as a special-purpose solution. Since existing GPU simulators [3, 45] cannot support data compression, we will continue to follow up and study this problem in the future work.

7 CONCLUSION

In this work, we propose a novel outlier-victim pair (OVP) quantization, which can handle outlier values with low hardware overhead and achieve high performance gains. The key insight is to sacrifice the normal values next to those essential outliers (called victims) to accommodate them. The OVP encoding designed based on this idea is able to make outliers and normal values globally identical but locally distinguishable. To the best of our knowledge, Olive pushes the limit of 4-bit quantization to a new state-of-the-art, as it is able to achieve nearly original accuracy for commonly used language models. Moreover, our architecture design can be efficiently integrated into existing hardware accelerators such as tensor core and systolic array. Finally, Olive-based accelerator surpasses the existing outlier-aware accelerator, GOBO, by 4.5× speedup and 4.0× energy reduction, respectively.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant 2022YFB4501401, the National Natural Science Foundation of China (NSFC) grant (62222210, and 62072297, and 61832006). The authors would like to thank the anonymous reviewers for their constructive feedback for improving the work. We also thank Tailong Wangliu, Shuangjie Ruan for their continuous support.

REFERENCES

- [1] 2020. Nvidia ampere architecture whitepaper. <https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf>.
- [2] Jorge Albericio, Patrick Judd, Tayler Hetherington, Tor Aamodt, Natalie Enright Jerger, and Andreas Moshovos. 2016. Cnvlutin: Ineffectual-neuron-free deep neural network computing. *ACM SIGARCH Computer Architecture News* 44, 3 (2016), 1–13.
- [3] Ali Bakhoda, George Yuan, Wilson Fung, Henry Wong, and Tor Aamodt. 2009. Analyzing CUDA workloads using a detailed GPU simulator. *ISPASS 2009* -

- International Symposium on Performance Analysis of Systems and Software*, 163–174. <https://doi.org/10.1109/ISPASS.2009.4919648>
- [4] Ron Banner, Yury Nahshan, and Daniel Soudry. 2019. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in Neural Information Processing Systems* 32 (2019).
- [5] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* (2013).
- [6] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13169–13178.
- [7] Zhaowei Cai and Nuno Vasconcelos. 2020. Rethinking differentiable search for mixed-precision neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2349–2358.
- [8] Quan Chen, Hailong Yang, Minyi Guo, Ram Srivatsa Kannan, Jason Mars, and Lingjia Tang. 2017. Prophet: Precise QoS Prediction on Non-Preemptive Accelerators to Improve Utilization in Warehouse-Scale Computers. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS 2017, Xi'an, China, April 8-12, 2017*. ACM, 17–32. <https://doi.org/10.1145/3037697.3037700>
- [9] Quan Chen, Hailong Yang, Jason Mars, and Lingjia Tang. 2016. Baymax: QoS Awareness and Increased Utilization for Non-Preemptive Accelerators in Warehouse Scale Computers. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS 2016, Atlanta, GA, USA, April 2-6, 2016*. ACM, 681–696. <https://doi.org/10.1145/2872362.2872368>
- [10] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. 2014. Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. *ACM SIGARCH Computer Architecture News* 42, 1 (2014), 269–284.
- [11] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Q. Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning. In *13th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2018, Carlsbad, CA, USA, October 8-10, 2018*. USENIX Association, 578–594. <https://doi.org/10.5555/3291168.3291211>
- [12] Yunji Chen, Tao Luo, Shaoli Liu, Shijun Zhang, Liqiang He, Jia Wang, Ling Li, Tianshi Chen, Zhiwei Xu, Ninghui Sun, et al. 2014. Dadiannao: A machine-learning supercomputer. In *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 609–622.
- [13] Yu-Hsin Chen, Tushar Krishna, Joel S Emer, and Vivienne Sze. 2016. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE journal of solid-state circuits* 52, 1 (2016), 127–138.
- [14] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. 2018. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085* (2018).
- [15] Yujeong Choi, Yunseong Kim, and Minsoo Rhu. 2021. Lazy Batching: An SLA-aware Batching System for Cloud Machine Learning Inference. In *IEEE International Symposium on High-Performance Computer Architecture, HPCA 2021, Seoul, South Korea, February 27 - March 3, 2021*. IEEE, 493–506. <https://doi.org/10.1109/HPCA51647.2021.00049>
- [16] Weihao Cui, Mengze Wei, Quan Chen, Xiaoxin Tang, Jingwen Leng, Li Li, and Mingyi Guo. 2019. Ebird: Elastic Batch for Improving Responsiveness and Throughput of Deep Learning Services. In *37th IEEE International Conference on Computer Design, ICCD 2019, Abu Dhabi, United Arab Emirates, November 17-20, 2019*. IEEE, 497–505. <https://doi.org/10.1109/ICCD46524.2019.00075>
- [17] Weihao Cui, Han Zhao, Quan Chen, Hao Wei, Zirui Li, Deze Zeng, Chao Li, and Minyi Guo. 2022. DVABatch: Diversity-aware Multi-Entry Multi-Exit Batching for Efficient Processing of DNN Services on GPUs. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)*. 183–198.
- [18] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339* (2022).
- [19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [20] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *arXiv:arXiv:2104.08758*
- [21] Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems* 33 (2020), 18518–18529.
- [22] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2019. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 293–302.
- [23] Zidong Du, Robert Fasthuber, Tianshi Chen, Paolo Ienne, Ling Li, Tao Luo, Xiaobing Feng, Yunji Chen, and Olivier Temam. 2015. ShiDianNao: Shifting vision processing closer to the sensor. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture*. 92–104.
- [24] Amir Gholami, Zhewei Yao, Sehoon Kim, Michael W Mahoney, and Kurt Keutzer. 2021. AI and Memory Wall. *RiseLab Medium Post* (2021).
- [25] Vinayak Gokhale, Jonghoon Jin, Aysegül Dundar, Berin Martini, and Eugenio Culurciello. 2014. A 240 g-ops/s mobile coprocessor for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 682–687.
- [26] Yue Guan, Jingwen Leng, Chao Li, Quan Chen, and Minyi Guo. 2020. How Far Does BERT Look At: Distance-based Clustering and Analysis of BERT’s Attention. *arXiv preprint arXiv:2011.00943* (2020).
- [27] Yue Guan, Zhengyi Li, Jingwen Leng, Zhouhan Lin, and Minyi Guo. 2022. Trankimmer: Transformer Learns to Layer-wise Skim. *arXiv preprint arXiv:2205.07324* (2022).
- [28] Yue Guan, Zhengyi Li, Zhouhan Lin, Yuhao Zhu, Jingwen Leng, and Minyi Guo. 2022. Block-skim: Efficient question answering for transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10710–10719.
- [29] Cong Guo, Bo Yang Hsueh, Jingwen Leng, Yuxian Qiu, Yue Guan, Zehuan Wang, Xiaoying Jia, Xipeng Li, Minyi Guo, and Yuhao Zhu. 2020. Accelerating sparse dnn models without hardware-support via tile-wise sparsity. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–15.
- [30] Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, and Minyi Guo. 2022. SQuant: On-the-Fly Data-Free Quantization via Diagonal Hessian Approximation. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=JXhR0KNZz0c>
- [31] Cong Guo, Yuxian Qiu, Jingwen Leng, Chen Zhang, Ying Cao, Quanlu Zhang, Yunxin Liu, Fan Yang, and Minyi Guo. 2022. Nesting Forward Automatic Differentiation for Memory-Efficient Deep Neural Network Training. In *2022 IEEE 40th International Conference on Computer Design (ICCD)*. IEEE, 738–745.
- [32] Cong Guo, Chen Zhang, Jingwen Leng, Zihan Liu, Fan Yang, Yunxin Liu, Minyi Guo, and Yuhao Zhu. 2022. ANT: Exploiting Adaptive Numerical Data Type for Low-bit Deep Neural Network Quantization. In *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 1414–1433.
- [33] Cong Guo, Chen Zhang, Jingwen Leng, Zihan Liu, Fan Yang, Yunxin Liu, Minyi Guo, and Yuhao Zhu. 2022. ANT github repository. https://github.com/clevercool/ANT_Micro22.
- [34] Cong Guo, Yangjie Zhou, Jingwen Leng, Yuhao Zhu, Zidong Du, Quan Chen, Chao Li, Bin Yao, and Minyi Guo. 2020. Balancing Efficiency and Flexibility for DNN Acceleration via Temporal GPU-Systolic Array Integration. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*. 1–6.
- [35] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. 2015. Deep learning with limited numerical precision. In *International conference on machine learning*. PMLR, 1737–1746.
- [36] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv preprint arXiv:1510.00149* (2015).
- [37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [38] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2704–2713.
- [39] Shubham Jain, Swagath Venkataramani, Vijayalakshmi Srinivasan, Jungwook Choi, Kailash Gopalakrishnan, and Leland Chang. 2019. BiScaled-DNN: Quantizing long-tailed datastructures with two scale factors for deep neural networks. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.
- [40] Zhihao Jia, Oded Padon, James J. Thomas, Todd Warszawski, Matei Zaharia, and Alex Aiken. 2019. TASO: optimizing deep learning computation with automatic generation of graph substitutions. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles (SOSP)*. ACM, 47–62. <https://doi.org/10.1145/3341301.3359630>
- [41] Norman P. Jouppi, Doe Hyun Yoon, Matthew Ashcraft, Mark Gottscho, Thomas B. Jablin, George Kurian, James Laudon, Sheng Li, Peter Ma, Xiaoyu Ma, Thomas Norrie, Nishant Patil, Sushma Prasad, Cliff Young, Zongwei Zhou, and David Patterson. 2021. Ten Lessons From Three Generations Shaped Google’s TPUv4: Industrial Product. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*.
- [42] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. 2017.

- In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture*. 1–12.
- [43] Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang, and Changkyu Choi. 2019. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4350–4359.
- [44] Andrew Kerr, Haicheng Wu, Manish Gupta, Dustyn Blasig, Pradeep Ramini, Duane Merrill, Aniket Shivam, Piotr Majcher, Paul Springer, Markus Hohnerbach, Jin Wang, and Matt Nicely. 2022. CUTLASS. <https://github.com/NVIDIA/cutlass>
- [45] Mahmoud Khairy, Zhesheng Shen, Tor Aamodt, and Timothy Rogers. 2020. Accel-Sim: An Extensible Simulation Framework for Validated GPU Modeling. 473–486. <https://doi.org/10.1109/ISCA45697.2020.00047>
- [46] Mahmoud Khairy, Zhesheng Shen, Tor M Aamodt, and Timothy G Rogers. 2020. Accel-Sim: An extensible simulation framework for validated GPU modeling. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 473–486.
- [47] Pran Kurup and Taher Abbasi. 2012. *Logic synthesis using Synopsys®*. Springer Science & Business Media.
- [48] Jingwen Leng, Tayler Hetherington, Ahmed ElTantawy, Syed Gilani, Nam Kim, Tor Aamodt, and Vijay Janapa Reddi. 2013. GPUWatch: enabling energy optimizations in GPGPUs. *ACM SIGARCH Computer Architecture News* 41 (07 2013). <https://doi.org/10.1145/2508148.2485964>
- [49] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
- [50] Zhengyi Li, Cong Guo, Zhanda Zhu, Yangjie Zhou, Yuxian Qiu, Xiaotian Gao, Jingwen Leng, and Minyi Guo. 2022. Efficient Activation Quantization via Adaptive Rounding Border for Post-Training Quantization. *arXiv preprint arXiv:2208.11945* (2022).
- [51] Daofu Liu, Tianshi Chen, Shaoli Liu, Jinhong Zhou, Shengyuan Zhou, Olivier Teman, Xiaobing Feng, Xuehai Zhou, and Yunji Chen. 2015. PuDianNao: A Poly-valent Machine Learning Accelerator. In *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*. 369–381.
- [52] Zihan Liu, Jingwen Leng, Zhihui Zhang, Quan Chen, Chao Li, and Minyi Guo. 2022. VELTAIR: towards high-performance multi-tenant deep learning services via adaptive compilation and scheduling. In *ASPLOS '22: 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, 28 February 2022 - 4 March 2022*, Babak Falsafi, Michael Ferdman, Shan Lu, and Thomas F. Wenisch (Eds.). ACM, 388–401. <https://doi.org/10.1145/3503222.3507752>
- [53] David Lo, Liqun Cheng, Rama Govindaraju, Parthasarathy Ranganathan, and Christos Kozyrakis. 2015. Heracles: improving resource efficiency at scale. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture (ISCA)*. <https://doi.org/10.1145/2749469.2749475>
- [54] Jason Mars, Lingjia Tang, Robert Hundt, Kevin Skadron, and Mary Lou Soffa. 2011. Bubble-Up: increasing utilization in modern warehouse scale computers via sensible co-locations. In *IEEE/ACM International Symposium on Microarchitecture (MICRO)*. <https://doi.org/10.1145/2155620.2155650>
- [55] ModelTC. 2022. repositories. <https://huggingface.co/ModelTC>.
- [56] Naveen Muralimanohar, Rajeev Balasubramonian, and Norman P Jouppi. 2009. CACTI 6.0: A tool to model large caches. *HP laboratories* 27 (2009), 28.
- [57] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. 2021. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295* (2021).
- [58] Nvidia. 2017. NVIDIA Tesla V100 GPU Architecture. In *Technical report*. NVIDIA.
- [59] Nvidia. 2018. NVIDIA Turing GPU Architecture. In *Technical report*. NVIDIA.
- [60] Nvidia. 2020. NVIDIA A100 tensor core architecture. In *Technical report*. NVIDIA.
- [61] Eunhyeok Park, Dongyoung Kim, and Sungjoo Yoo. 2018. Energy-efficient neural network ‘accel’erators based on outlier-aware low-precision computation. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 688–698.
- [62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019), 8026–8037.
- [63] Maurice Peemen, Arnaud AA Setio, Bart Mesman, and Henk Corporaal. 2013. Memory-centric accelerator design for convolutional neural networks. In *2013 IEEE 31st International Conference on Computer Design (ICCD)*. IEEE, 13–19.
- [64] Eric Qin, Ananda Samajdar, Hyoukjun Kwon, Vineet Nadella, Sudarshan Srinivasan, Dipankar Das, Bharat Kaul, and Tushar Krishna. 2020. Sigma: A sparse and irregular gemm accelerator with flexible interconnects for dnn training. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 58–70.
- [65] Yuxian Qiu, Jingwen Leng, Cong Guo, Quan Chen, Chao Li, Minyi Guo, and Yuhao Zhu. 2019. Adversarial Defense Through Network Profiling Based Path Extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [66] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
- [67] Md Aamir Raihan, Negar Goli, and Tor M Aamodt. 2019. Modeling deep learning accelerator enabled gpus. In *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 79–92.
- [68] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- [69] Satyabrata Sarangi and Bevan Baas. 2021. DeepScaleTool: A Tool for the Accurate Estimation of Technology Scaling in the Deep-Submicron Era. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5.
- [70] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100* (2022).
- [71] Hardik Sharma, Jongse Park, Divya Mahajan, Emmanuel Amaro, Joon Kyung Kim, Chenkai Shao, Asit Mishra, and Hadi Esmaeilzadeh. 2016. From high-level deep neural models to FPGAs. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 1–12.
- [72] Hardik Sharma, Jongse Park, Naveen Suda, Liangzhen Lai, Benson Chau, Vikas Chandra, and Hadi Esmaeilzadeh. 2018. Bit fusion: Bit-level dynamically composable architecture for accelerating deep neural network. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 764–775.
- [73] Hardik Sharma, Jongse Park, Naveen Suda, Liangzhen Lai, Benson Chau, Vikas Chandra, and Hadi Esmaeilzadeh. 2018. Bitfusion github repository. <https://github.com/hsharma35/bitfusion>.
- [74] Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8815–8821.
- [75] Zhuoran Song, Bangqi Fu, Feiyang Wu, Zhaoming Jiang, Li Jiang, Naifeng Jing, and Xiaoyao Liang. 2020. Drq: dynamic region-based quantization for deep neural network acceleration. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 1010–1021.
- [76] Thierry Tambe, En-Yu Yang, Zishen Wan, Yuntian Deng, Vijay Janapa Reddi, Alexander Rush, David Brooks, and Gu-Yeon Wei. 2020. Algorithm-hardware co-design of adaptive floating-point encodings for resilient deep learning inference. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.
- [77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [78] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).
- [79] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. 2019. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8612–8620.
- [80] Yang Wang, Chen Zhang, Zhiqiang Xie, Cong Guo, Yunxin Liu, and Jingwen Leng. 2021. Dual-side sparse tensor core. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 1083–1095.
- [81] Ziwei Wang, Jiwen Lu, Chenxin Tao, Jie Zhou, and Qi Tian. 2019. Learning channel-wise interactions for binary convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 568–577.
- [82] Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. 2022. Outlier Suppression: Pushing the Limit of Low-bit Transformer Language Models. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). <https://openreview.net/forum?id=yW5zeRSFdz>
- [83] Wikipedia contributors. 2022. 68–95–99.7 rule – Wikipedia, The Free Encyclopedia. [Online].
- [84] Hailong Yang, Alex D. Breslow, Jason Mars, and Lingjia Tang. 2013. Bubble-flux: precise online QoS management for increased utilization in warehouse scale computers. In *The 40th Annual International Symposium on Computer Architecture (ISCA)*. <https://doi.org/10.1145/2485922.2485974>
- [85] Ali Hadi Zadeh, Isak Edo, Omar Mohamed Awad, and Andreas Moshovos. 2020. Gobo: Quantizing attention-based nlp models for low latency and energy efficient inference. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 811–824.
- [86] Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMCC2-NIPS)*. IEEE, 36–39.
- [87] Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, and Jason Cong. 2015. Optimizing fpga-based accelerator design for deep convolutional neural networks. In *Proceedings of the 2015 ACM/SIGDA international symposium on field-programmable gate arrays*. 161–170.

- [88] Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, and Gang Hua. 2018. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*. 365–382.
- [89] Shijin Zhang, Zidong Du, Lei Zhang, Huiying Lan, Shaoli Liu, Ling Li, Qi Guo, Tianshi Chen, and Yunji Chen. 2016. Cambricon-X: An accelerator for sparse neural networks. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 1–12.
- [90] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).
- [91] Lianmin Zheng, Chengfan Jia, Minmin Sun, Zhao Wu, Cody Hao Yu, Ameer Haj-Ali, Yida Wang, Jun Yang, Danyang Zhuo, Koushik Sen, Joseph E. Gonzalez, and Ion Stoica. 2020. Ansor: Generating High-Performance Tensor Programs for Deep Learning. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. <https://doi.org/10.5555/3488766.3488815>
- [92] Size Zheng, Yun Liang, Shuo Wang, Renze Chen, and Kaiwen Sheng. 2020. Flex-Tensor: An Automatic Schedule Exploration and Optimization Framework for Tensor Computation on Heterogeneous System. In *Architectural Support for Programming Languages and Operating Systems, Lausanne (ASPLOS)*. <https://doi.org/10.1145/3373376.3378508>
- [93] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. 2016. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160* (2016).
- [94] Xuda Zhou, Zidong Du, Qi Guo, Shaoli Liu, Chengsi Liu, Chao Wang, Xuehai Zhou, Ling Li, Tianshi Chen, and Yunji Chen. 2018. Cambricon-S: Addressing irregularity in sparse neural networks through a cooperative software/hardware approach. In *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 15–28.
- [95] Yangjie Zhou, Jingwen Leng, Yaoxu Song, Shuwen Lu, Mian Wang, Chao Li, Minyi Guo, Wenting Shen, Yong Li, Wei Lin, et al. 2023. uGrapher: High-Performance Graph Operator Computation via Unified Abstraction for Graph Neural Networks. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. 878–891.
- [96] Yangjie Zhou, Mengtian Yang, Cong Guo, Jingwen Leng, Yun Liang, Quan Chen, Minyi Guo, and Yuhao Zhu. 2021. Characterizing and demystifying the implicit convolution algorithm on commercial matrix-multiplication accelerators. In *2021 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 214–225.
- [97] Hongyu Zhu, Ruofan Wu, Yijia Diao, Shanbin Ke, Haoyu Li, Chen Zhang, Jilong Xue, Lingxiao Ma, Yuqing Xia, Wei Cui, Fan Yang, Mao Yang, Lidong Zhou, Asaf Cidon, and Gennady Pekhimenko. 2022. ROLLER: Fast and Efficient Tensor Compilation for Deep Learning. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*. 233–248.
- [98] Maohua Zhu, Tao Zhang, Zhenyu Gu, and Yuan Xie. 2019. Sparse tensor core: Algorithm and hardware co-design for vector-wise sparse neural networks on modern gpus. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. 359–371.
- [99] Bohan Zhuang, Mingkui Tan, Jing Liu, Lingqiao Liu, Ian Reid, and Chunhua Shen. 2021. Effective training of convolutional neural networks with low-bitwidth weights and activations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).