BinaryAttention: One-Bit QK-Attention for Vision and Diffusion Transformers

Supplementary Material

In this supplementary file, we provide following materials:

A. Proof of Theorem 1

Proof. Consider the element (i, j) of the matrix $\boldsymbol{s}\boldsymbol{t}^T$:

$$[\boldsymbol{s}\boldsymbol{t}^T]_{ij} = \boldsymbol{s}_i\boldsymbol{t}_j = \text{sign}(\boldsymbol{q}_i)\text{sign}(\boldsymbol{k}_j).$$

Since $\boldsymbol{q}$ and $\boldsymbol{k}$ are assumed to be jointly Gaussian with zero-mean, the pair $(\boldsymbol{q}_i, \boldsymbol{k}_j)$ is also jointly Gaussian. We have:

$$\text{Var}(\boldsymbol{q}_i) = \boldsymbol{\Sigma}_{qq}[i, i],\ \text{Var}(\boldsymbol{k}_j) = \boldsymbol{\Sigma}_{kk}[j, j],$$
$$\text{Cov}(\boldsymbol{q}_i, \boldsymbol{k}_j) = \boldsymbol{\Sigma}_{qk}[i, j].$$

Then the correlation of $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$ is given by:

$$\rho_{ij} = \frac{\boldsymbol{\Sigma}_{qk}[i, j]}{\sqrt{\boldsymbol{\Sigma}_{qq}[i, i]\boldsymbol{\Sigma}_{kk}[j, j]}} = \boldsymbol{C}_{ij}.$$

Let $x = \boldsymbol{\Sigma}_{qq}^{-\frac{1}{2}}[i, i]\boldsymbol{q}_i$ and $y = \boldsymbol{\Sigma}_{kk}^{-\frac{1}{2}}[j, j]\boldsymbol{k}_j$, then the two variables x, y are standard Gaussian with correlation ρ_{ij}, and the joint density can be expressed as:

$$p(x, y) = \frac{1}{2\pi\sqrt{1-\rho_{ij}}}\exp\left(-\frac{x^2 - 2\rho_{ij}xy + y^2}{2(1-\rho_{ij}^2)}\right).$$

We now calculate the expectation of $\text{sign}(x)\text{sign}(y)$, where

$$\text{sign}(x)\text{sign}(y) = \begin{cases} +1 & \text{if } x \geq 0, y \geq 0 \text{ or } x \leq 0, y \leq 0 \\ -1 & \text{if } x \geq 0, y < 0 \text{ or } x < 0, y \geq 0 \end{cases}.$$

By the symmetry of standard Gaussian, we have

$$\mathbb{E}[\text{sign}(x)\text{sign}(y)] = 4\mathbb{P}(x \geq 0, y \geq 0) - 1.$$

Since

$$\mathbb{P}(x \geq 0, y \geq 0) = \int_0^\infty \int_0^\infty p(x, y)dxdy \xrightarrow[y=R\sin\theta]{x=R\cos\theta}$$
$$\frac{1}{2\pi\sqrt{1-\rho_{ij}}} \int_0^{\frac{\pi}{2}} \int_0^\infty \exp\left(-\frac{R^2(1-\rho_{ij}\sin 2\theta)}{2(1-\rho_{ij}^2)}\right) RdRd\theta$$
$$= \frac{1}{2\pi}\arcsin\rho_{ij} + \frac{1}{4},$$

we have:

$$\mathbb{E}[\text{sign}(x)\text{sign}(y)] = \frac{2}{\pi}\arcsin\rho_{ij} = \frac{2}{\pi}\arcsin\boldsymbol{C}_{ij}.$$

Note that the $\text{sign}(\cdot)$ function is scale-invariant for any strictly positive scales, we can ensure that

$$\mathbb{E}[\text{sign}(\boldsymbol{q}_i)\text{sign}(\boldsymbol{k}_j)] = \mathbb{E}[\text{sign}(x)\text{sign}(y)] = \frac{2}{\pi}\arcsin\boldsymbol{C}_{ij}.$$

Since it holds for all $i, j = 1, \ldots, d$, there is:

$$\mathbb{E}[\boldsymbol{s}\boldsymbol{t}^T] = \frac{2}{\pi}\arcsin\boldsymbol{C}.$$

□

B. Algorithm of BinaryAttention

Our implementation of BinaryAttention is built upon the fundamental principles of FlashAttention2 [2] and SageAttention [7] while introducing specialized optimizations for binary and low-precision computations. The complete algorithm is presented in Algorithm 1.

Algorithm 1: Implementation of BinaryAttention

Input: Matrices $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{N\times d}$, bias $\boldsymbol{B} \in \mathbb{R}^{N\times N}$, block size B_r, B_v .
Output: Matrices $\boldsymbol{O} \in \mathbb{R}^{N\times d}$.

Processing:
$(\mu_q, \boldsymbol{S}) \leftarrow \phi(\boldsymbol{Q}), (\mu_k, \boldsymbol{T}) \leftarrow \phi(\boldsymbol{K}), (\delta_v, \tilde{\boldsymbol{V}}) \leftarrow \psi(\boldsymbol{V})$;
`// Quantization by Eq.(5) and Eq.(7)`
Divide $\boldsymbol{S}, \boldsymbol{O}$ into $T_r := \lceil N/B_r \rceil$ blocks $\{\boldsymbol{S}_i\}$ and $\{\boldsymbol{O}_i\}$;
Divide $\boldsymbol{T}, \tilde{\boldsymbol{V}}$ into $T_v := \lceil N/B_v \rceil$ blocks $\{\boldsymbol{T}_j\}$ and $\{\tilde{\boldsymbol{V}}_j\}$;
Divide $\boldsymbol{B}$ into $T_r \times T_v$ blocks $\{\boldsymbol{B}_{ij}\}$; `// if bias`
for $i = 1$ **to** T_r **do**
  Load block $\boldsymbol{S}_i$ from HBM to SRAM ;
  Initialize $\boldsymbol{O}_{i,0} = (0)_{B_r\times d}, l_{i,0} = (0)_{B_r}, m_{i,0} = (-\infty)_{B_r}$;
  for $j = 1$ **to** T_v **do**
    Load blocks $\boldsymbol{T}_j, \tilde{\boldsymbol{V}}_j, \boldsymbol{B}_{ij}$ from HBM to SRAM ;
    $\boldsymbol{S}_{ij} \leftarrow \text{BinaryMatmul}(\boldsymbol{S}_i, \boldsymbol{T}_j) \times \mu_q \times \mu_k$;
    $\boldsymbol{S}_{ij} \leftarrow \boldsymbol{S}_{ij} + \boldsymbol{B}_{ij}$; `// if bias`
    $m_{ij} \leftarrow \max(m_{i,j-1}, \text{rowmax}(\boldsymbol{S}_{ij}))$;
    $\hat{\boldsymbol{P}}_{ij} \leftarrow \exp(\boldsymbol{S}_{ij} - m_{ij})$;
    $l_{ij} \leftarrow e^{m_{i,j-1}-m_{ij}} l_{i,j-1} + \text{rowsum}(\hat{\boldsymbol{P}}_{ij})$;
    $\boldsymbol{O}_{ij} \leftarrow \text{IntMatmul}(\hat{\boldsymbol{P}}_{ij} \times 255, \tilde{\boldsymbol{V}}_j)$;
    $\boldsymbol{O}_{ij} \leftarrow \text{diag}(e^{m_{i,j-1}-m_{ij}})^{-1}\boldsymbol{O}_{i,j-1} + \boldsymbol{O}_{ij}$;
  end
  $\boldsymbol{O}_i \leftarrow \text{diag}(l_{i,T_v})^{-1}\boldsymbol{O}_{i,T_v}/255 \times \delta_v$;
  Write $\boldsymbol{O}_i$;
end
return $\boldsymbol{O} = \{\boldsymbol{O}_i\}$.

Table 1. Ablation studies on BinaryAttention for the scaled binary representations, bias enhancement and self-distillation strategy on ImageNet-1K benchmark, using DeiT architectures.

Scale	Bias	Distillation	DeiT-T Top-1 Accuracy	DeiT-S Top-1 Accuracy	DeiT-B Top-1 Accuracy
Baseline (full-precision)			72.2	79.8	81.8
✗	✗	✗	71.95	79.59	81.10
✓	✗	✗	72.42	79.81	81.33
✓	✗	✓	72.44	79.97	81.99
✓	✓	✓	72.88	80.24	82.04

Table 2. Attention pattern comparison between FlashAttention2 and BinaryAttention by DeiT-B on ImageNet-1K validation set.

	CosSim	Relative L1	RMSE	Precision
Layer (0)	0.9186	0.4840	0.4165	0.7716
Layer (6)	0.8740	0.7353	0.5084	0.7301

C. Ablation Studies

We first conduct a series of ablation studies to analyze the core components of BinaryAttention, including scaled binary representations, bias enhancement, and the self-distillation strategy. The experiments are performed on the ImageNet-1K [3] benchmark by using DeiT [6] architectures. The results are summarized in Tab. 1.

Scaled Binary Representations. We evaluate the role of scaling factors in binary representations. When scaling is not applied, BinaryAttention shows performance drop across all models, with top-1 accuracy decreased by -0.25%, -0.21%, and -0.70% for DeiT-T, -S, and -B, respectively. Introducing scaling factors effectively solves this issue by minimizing the quantization error, with DeiT-T even exceeding its full-precision baseline (72.42% vs. 72.2%), demonstrating that proper scaling is essential for preserving representational capabilities in binary space.

Bias Enhancement. We simply employ a learnable relative position bias [5] as the bias term, which exhibits distinct effects across model scales. It provides an accuracy gain of 0.44% and 0.27% for DeiT-T and -S, respectively, while offering a slight improvement for DeiT-B, from 81.99% to 82.04%. This discrepancy stems from the relationship between model capacity and the expressive power of binary representations. For smaller models, the limited dimension constrains the diversity of attention patterns, making them more susceptible to distribution collapse. The bias term effectively mitigates this by introducing additional contextual or structural information. For larger models, higher-dimensional binary representations naturally preserve richer similarity structures, yielding more modest gains.

Self-distillation Strategy. We investigate the role of self-distillation, which slightly improves DeiT-T and -S models but significantly boosts the accuracy of DeiT-B by 0.66%. This improvement suggests that self-distillation effectively counteracts the distribution shift introduced by quantization errors while encouraging sign-aligned similarity between binary representations and its full-precision counterparts.

Table 3. Memory comparison by DeiT-T using FlashAttention2, SageAttention and BinaryAttention at resolutions of 512 and 1024.

Method	Top-1	Mem. (512)	Mem. (1024)
FlashAttention2	72.2	1705M	5304M
SageAttention	72.11	1705M	5304M
BinaryAttention (den.)	72.88	3246M	29904M
BinaryAttention (dec.)	72.97	1706M	5307M

Table 4. Latency of attention kernels and quantization components measured on A100 GPUs.

FlashAttention2	SageAttention	BinaryAttention	Quant Q&K	Quant V
175.3ms	124.6ms	88.2ms	2.8ms	1.9ms

Attention Pattern Fidelity. We further analyze whether BinaryAttention preserves the original attention dynamics. We use Cosine Similarity, Relative L1 Distance, RMSE, and Precision as evaluation metrics, where Precision measures the accuracy of matching the top 100 most attended tokens. As shown in Tab. 2, BinaryAttention maintains high consistency with full-precision attention on ImageNet-1K validation set, with cosine similarity above 0.87 and precision around 0.75, demonstrating that BinaryAttention effectively preserves key relational patterns and structural relationships.

Memory and Quantization Overhead. Finally, we report the memory footprint in Tab. 3 and the quantization overhead in Tab. 4. BinaryAttention incurs extra memory primarily from the bias term. With a dense bias, memory grows rapidly with resolution, and with a decomposable bias, *e.g.*, a sum over spatial directions, the overhead becomes almost negligible. Meanwhile, the quantization cost is modest, requiring 2.8ms for query and key and 1.9ms for value (4.7ms in total), which accounts for about 5% of the BinaryAttention kernel.

D. Experimental Details in Classification

Settings. We benchmark BinaryAttention on ImageNet-1K [3] dataset. Following the experimental configurations of DeiT [6], we employ the AdamW optimizer for 300 epochs with beta set to (0.9, 0.999), momentum of 0.9 and a batch size of 1024. An initial learning rate of 10^{-4}, a minimum learning rate of 10^{-5} and a weight decay of 0.02 are used. The learning rate follows a cosine annealing schedule with a warm-up of 5 epochs. We include commonly used augmentation and regularization strategies, consistent with the

training of DeiT. The drop path rate is set to 0.1 for all BinaryAttention variants. Before training, the models are initialized with full-precision pre-trained weights. We utilize the self-distillation strategy with the full-precision counterpart as teacher, and implement quantization-aware training with Straight-Through Estimators (STE) [1]. For the input resolution of 384×384, we continue to fine-tune the models for 30 epochs, with a batch size of 512, a constant learning rate of 10^{-5}, and a weight decay of 10^{-8}.

Table 5. Top-1 accuracy of DeiT models using BinaryAttention without bias at 100 and 300 fine-tuning epochs.

Epochs	DeiT-T	DeiT-S	DeiT-B
Baseline	72.2	79.8	81.8
100	71.98	79.44	81.80
300	72.44	79.97	81.99

Tuning Cost. In extreme low-bit quantization, fine-tuning is a standard and necessary step to bridge performance gaps. While training-free methods prioritize convenience, their performance is strictly limited by the baseline, whereas BinaryAttention raises the accuracy. For the best performance, we employ a fine-tuning schedule of 300 epochs, which is consistent with common practice in low-bit approaches such as BiViT [4]. We further report the performance at different fine-tuning epochs in Tab. 5. It can be seen that with 100 fine-tuning epochs, BinaryAttention already nearly matches the baseline, with Top-1 accuracy gaps of 0.22/0.36/0.00 on DeiT-T/S/B, respectively. Extending fine-tuning to 300 epochs not only recovers the baseline performance but yields a modest improvement.

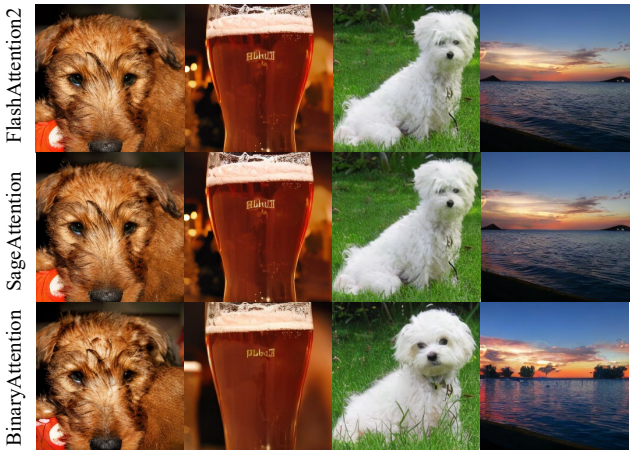

Figure 1. More qualitative comparisons of generated image by DiT-XL/2 (cfg=1.50) using FlashAttention2, SageAttention and BinaryAttention.

E. More Qualitative Comparisons

Fig. 1 provides more qualitative comparisons of FlashAttention2, SageAttention, and BinaryAttention, showing additional images generated by the DiT-XL/2 model (cfg=1.50). We can see that SageAttention and FlashAttention2 produce nearly identical images, while BinaryAttention produces slightly different content but maintains competitive generation quality with sufficient details.

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 3

[2] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 2

[4] Yefei He, Zhenyu Lou, Luoming Zhang, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. BiViT: Extremely compressed binary vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5651–5663, 2023. 3

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[6] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2

[7] Jintao Zhang, Jia Wei, Pengle Zhang, Jun Zhu, and Jianfei Chen. SageAttention: Accurate 8-bit attention for plug-and-play inference acceleration. In *International Conference on Learning Representations*, 2025. 1