Faster Inference of Integer SWIN Transformer by Removing the GELU Activation

Mohammadreza Tayaranian Seyyed Hasan Mozafari
James J. Clark Brett Meyer Warren Gross
Department of Electrical and Computer Engineering, McGill University

McGill

Introduction

- SWIN Transformer is an enhanced vision transformer architecture [4].
- It uses windowed attention to accommodate larger input images.
- Window attention operations slow down the inference of SWIN transformer: $\text{SWIN}_{\text{SMALL}}$ is **55%** slower than $\text{ViT}_{\text{SMALL}}$ [5].
- We use **integer quantization** for faster inference of SWIN transformer.

A non-linear operation f is not easily quantizable due to $f(s\hat{x}) \neq sf(\hat{x})$

Previous Work

The goal of these works is to mitigate the overhead of using non-integer operations in an integer model:

1. Replace non-linear operations with approximation functions [1, 2, 3]:
   - ✓ Linear or piece-wise linear approximations that are easy to quantize.
   - X Complicated implementations with higher cost than the original overhead!
2. Keep non-integer operations, but fuse them together (Figure 1):
   - ✓ One data conversion for multiple non-integer operations.
   - X Non-integer operations still (sometimes unnecessarily) exist.

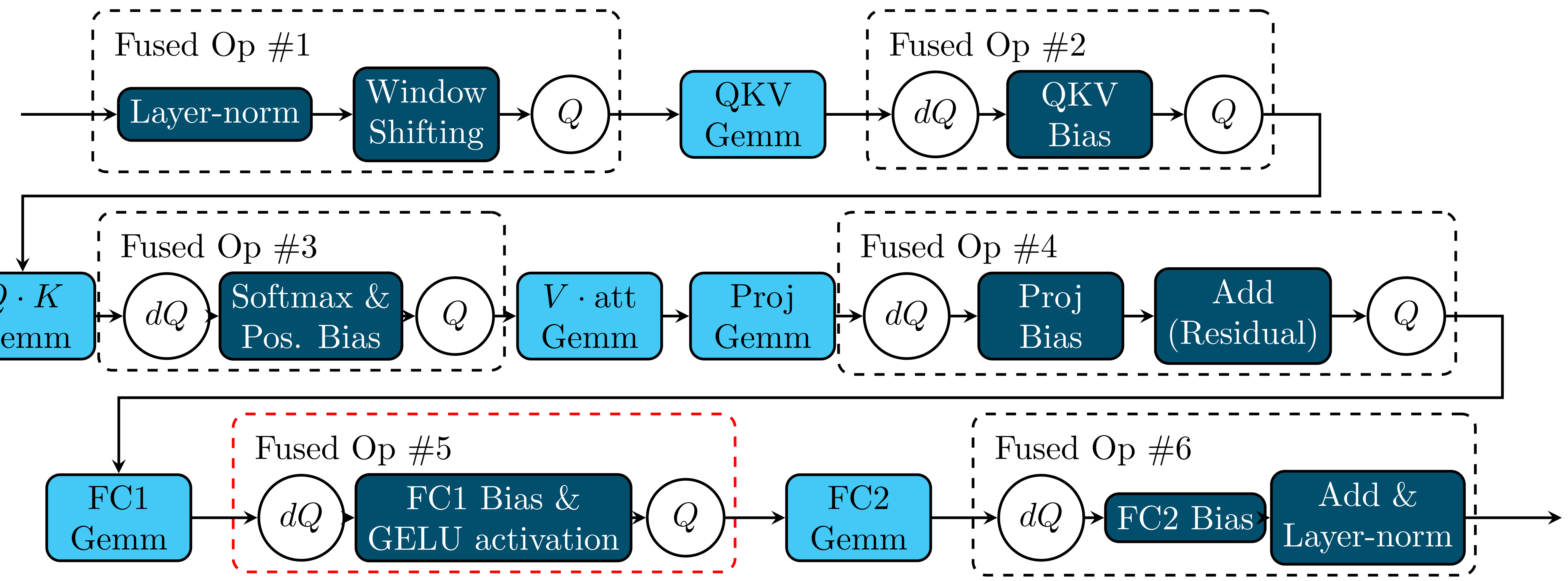

Figure 1. High level schematics of the integer SWIN Transformer. Based on the FasterTransformer framework.

Proposed Method

Replace the non-linear GELU activation with the piece-wise linear ReLU.

- ✓ GELU is responsible for a big part of the inference latency (Table 2).
- ✓ Compared to approximation functions, ReLU is simpler to implement.
- ✓ No more non-integer operations. No need for data conversions.

We use an iterative algorithm to replace GELU with ReLU:

Input: SWIN: SWIN Transformer Model, Dataset: KD dataset;
Parameter: N: Number of transformer blocks;
student $\leftarrow$ clone(SWIN) ;
for $i \leftarrow 1$ *to* N **do**
  student.blocks[i].activation $\leftarrow$ ReLU ;
  student.blocks[i].bias $\leftarrow 0$;
  disable_gradient(student.blocks[i].bias) ;
  kd_epoch(student, SWIN, Dataset) ;
end

- The student model at the end is called **GELU-less SWIN**.
- We remove both the GELU and the bias inside the fused operation.
- We use knowledge distillation to avoid the accuracy drop.
- We use 10% of the ImageNet training dataset for KD.
- Number of epochs is 12 or 24.
- After KD, GELU-less SWIN is quantized using post-training quantization method of the FasterTransformer framework.

Experimental Results

We use our GELU-less integer SWIN for inference on the ImageNet dataset.

Compared to the integer SWIN from the FasterTransformer:

We gain at least 11% speedup, with less than 0.5% drop in accuracy.

Model	Method	Datatype	Top-1 Acc. (%)	Latency (ms)	Speedup
$\text{SWIN}_{\text{TINY}}$	Baseline	FP32	81.2	60.27	×1
	Half-precision	FP16	81.2	24.96	×2.41
	FasterTransformer	int8	80.1	17.04	×3.54
	Ours	int8	80.0	15.01	**×4.02**
$\text{SWIN}_{\text{SMALL}}$	Baseline	FP32	83.2	103.21	×1
	Half-precision	FP16	83.2	40.26	×2.56
	FasterTransformer	int8	83.0	25.05	×4.12
	Ours	int8	82.5	22.57	**×4.57**
$\text{SWIN}_{\text{BASE}}$	Baseline	FP32	83.4	157.31	×1
	Half-precision	FP16	83.4	58.39	×2.69
	FasterTransformer	int8	83.3	35.55	×4.43
	Ours	int8	84.6	31.66	**×4.97**
$\text{SWIN}_{\text{LARGE}}$	Baseline	FP32	86.2	284.41	×1
	Half-precision	FP16	86.2	104.76	×2.71
	FasterTransformer	int8	85.8	61.35	×4.64
	Ours	int8	85.5	53.22	**×5.34**

Table 1. ImageNet top-1 accuracy and inference latency of various SWIN models.

Model	Fused Op # 1	2	3	4	5	6
$\text{SWIN}_{\text{TINY}}$	0.32	1.51	2.55	0.88	2.03	1.16
$\text{SWIN}_{\text{SMALL}}$	0.3	2.01	4.1	1.17	2.48	1.36
$\text{SWIN}_{\text{BASE}}$	0.57	2.53	5.44	1.57	3.89	2.69
$\text{SWIN}_{\text{LARGE}}$	1.22	3.93	8.18	2.92	8.13	3.66
Average	0.6	2.5	5.07	1.64	4.13	2.22

Table 2. Latency of fused operations shown in Figure 1 for different SWIN models.

References

[1] Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5506–5518. PMLR, 18–24 Jul 2021.

[2] Zhikai Li and Qingyi Gu. I-vit: integer-only quantization for efficient vision transformer inference. *arXiv preprint arXiv:2207.01405*, 2022.

[3] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. *arXiv preprint arXiv:2111.13824*, 2021.

[4] 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.

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

mohammadreza.tayaranian@mail.mcgill.ca