

# Supporting Sparse Inference in XNNPACK with RISC-V Vector Extension

Gary Yi-Hung Chen, Eric Hung-Yuan Chang, and Alan Quey-Liang Kao

## Summary

- Building on previous works [1][2], we contribute our implementation of sparse kernels [3] — including CONV2D-HWC2CHW, SPMM, and DWCONV2D-CHW — to the XNNPACK project.
- We present design considerations for RISC-V RVV microkernels in this work.
- Experiments
  - Our internal FPGA test board: AndeShape AE350, with AX45MPV CPU supporting VLEN = 128 and 256.
  - On MobileNetV2 with 85% sparsity, our implementation achieves **2.10× speedup** compared to the unpruned model.

## CONV2D-HWC2CHW

- This is the **third most time-consuming** kernel in sparse inference.
- Due to its **low compute intensity**, this kernel achieves only **0.77×** the performance of dense IGEMM.

## SPMM

- This is the **core computation** in sparse inference.
- By leveraging **RVV's elegant leftover handling**, we significantly simplified the kernel compared to [2].
- We observed that **zero-stride loads** on our testbed perform **worse** than a vector add after zero-reset.

## DWCONV2D-CHW

- This is the **second most time-consuming** kernel.
- Compared to dense DWCONV, our version shows **1.61× speedup** — although this might be due to suboptimal configuration in the baseline dense kernel. [4]

### Reference

- [1] Elsen, Erich, et al. Fast sparse convnets.
- [2] XNNPACK #7116
- [3] XNNPACK #8081
- [4] See QR codes for algorithm animation.

