Layer 40_self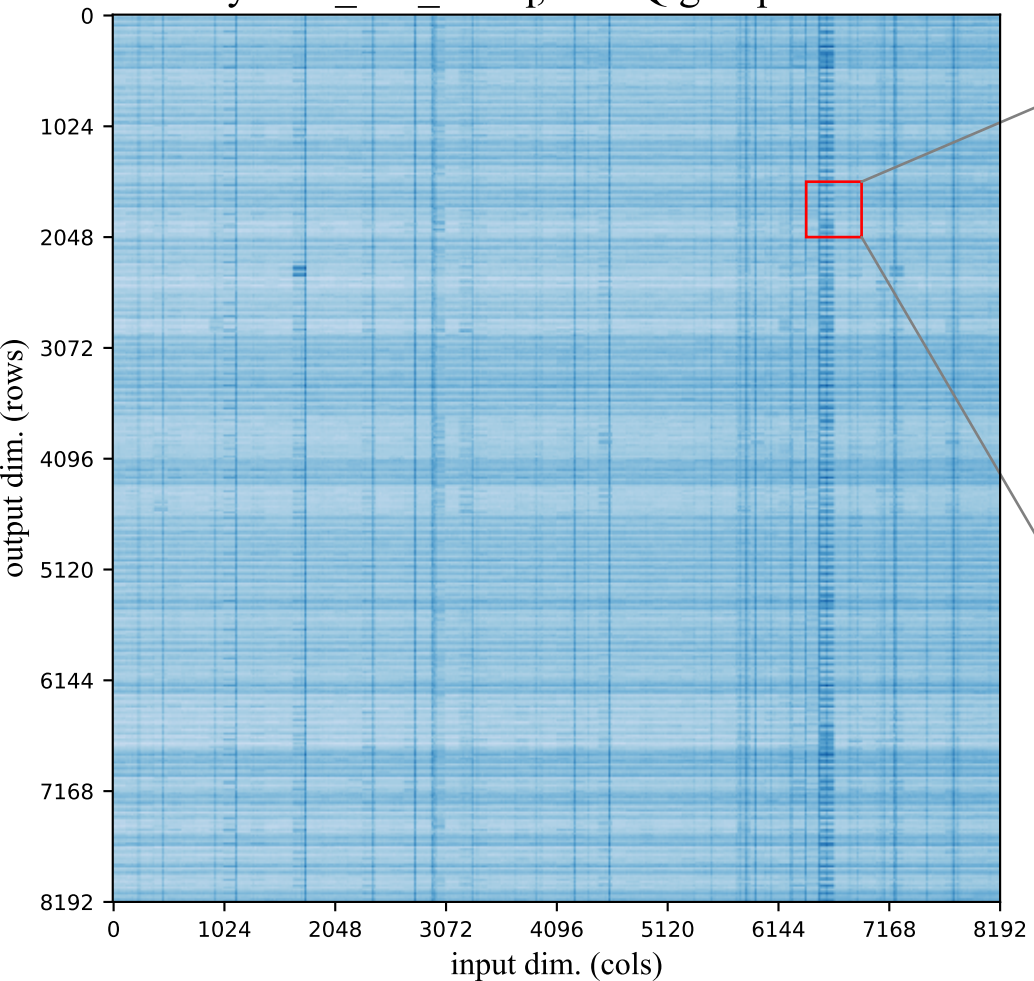_attn.q, GPTQ group size 128