* Inference in fp16: (1) Add normalization if it overflows （2）If it is integer RGB input (0-255) ,normalize it to be float (0-1)
* ~~Scale or Scale+ Shift: Determined by the range of real values being quantized~~
* Scale or Scale+ Shift: Use scale quantization
* Do not use the minimum negative value (e.g. [-127, 127] rather than [-128, 127]) to guarantee the range to be symmetric
* Use the minimum negative value may bring with bias and a small bias may be introduced towards negative infinity
* Using full range may not be the best choice of quantization
* Minimize the information loss between the original tensor and quantized tensor by KL-divergence
* Per tensor activation scale: Run calibration to choose best scaling factor
* Fine grained weight scale: Use maximum absolute value (full range) to compute scaling factor for 8-bit quantization
* Manually add clip at the right place can help quantization
* STE: Back propagate with (witch is actually not differentiable)
* STE is a method to add quantization to training