

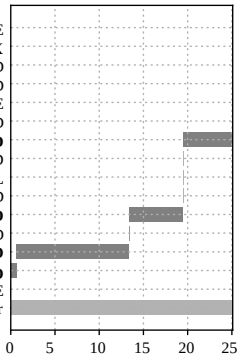
# Model A (fixed-point)

Tensor operation

CPU

CPU + TP  
(fixed-point)

DEQUANTIZE  
SOFTMAX  
FULLY\_CONNECTED  
FULLY\_CONNECTED  
RESHAPE  
MAX\_POOL\_2D  
**(4A) CONV\_2D**  
ADD  
MUL  
MAX\_POOL\_2D  
**(3A) CONV\_2D**  
MAX\_POOL\_2D  
**(2A) CONV\_2D**  
**(1A) CONV\_2D**  
QUANTIZE  
Interpreter



Time (s)

