AiCon
全球人工智能与机器学习技术大会

如何构建低成本高效能的视觉感知系统

潘争

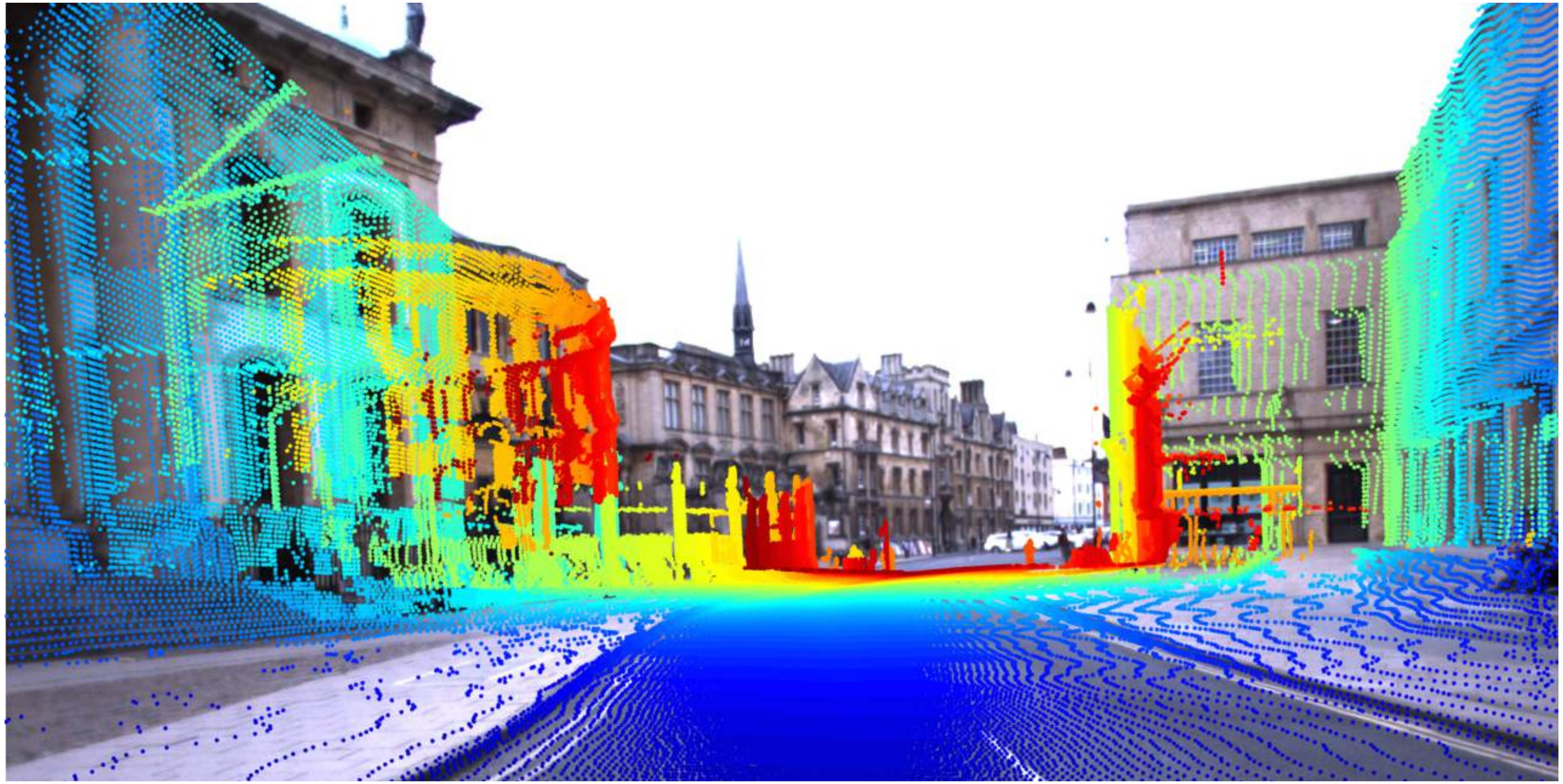
驭势科技

主办方 Geekbang> 极客邦科技 InfoQ

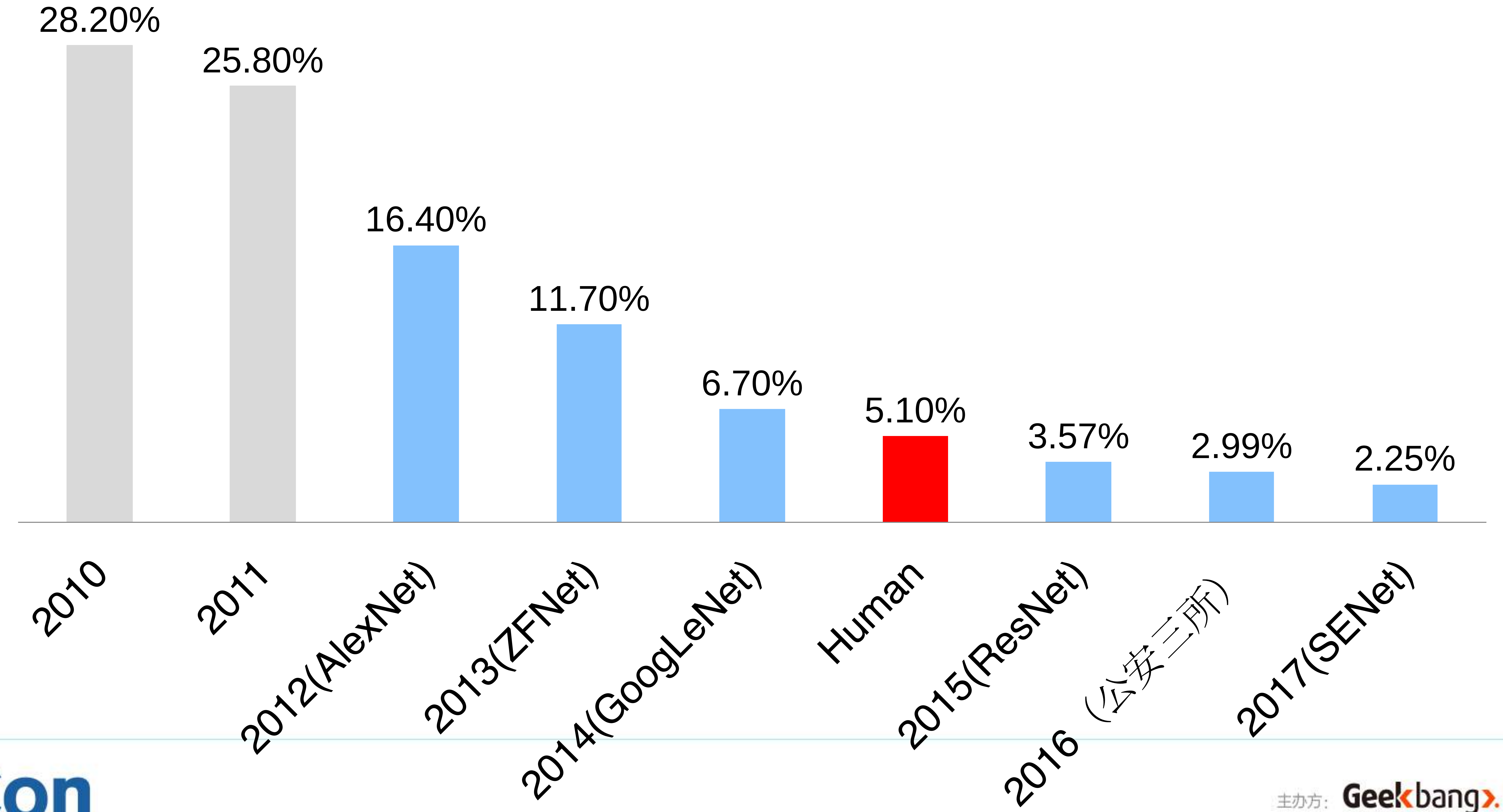
TABLE OF CONTENTES

视觉感知的优势

- 信息更丰富
- 视野更宽阔
- 基建更配合
- 硬件更便宜



视觉识别算法飞速进步
ImageNet Top-5错误率

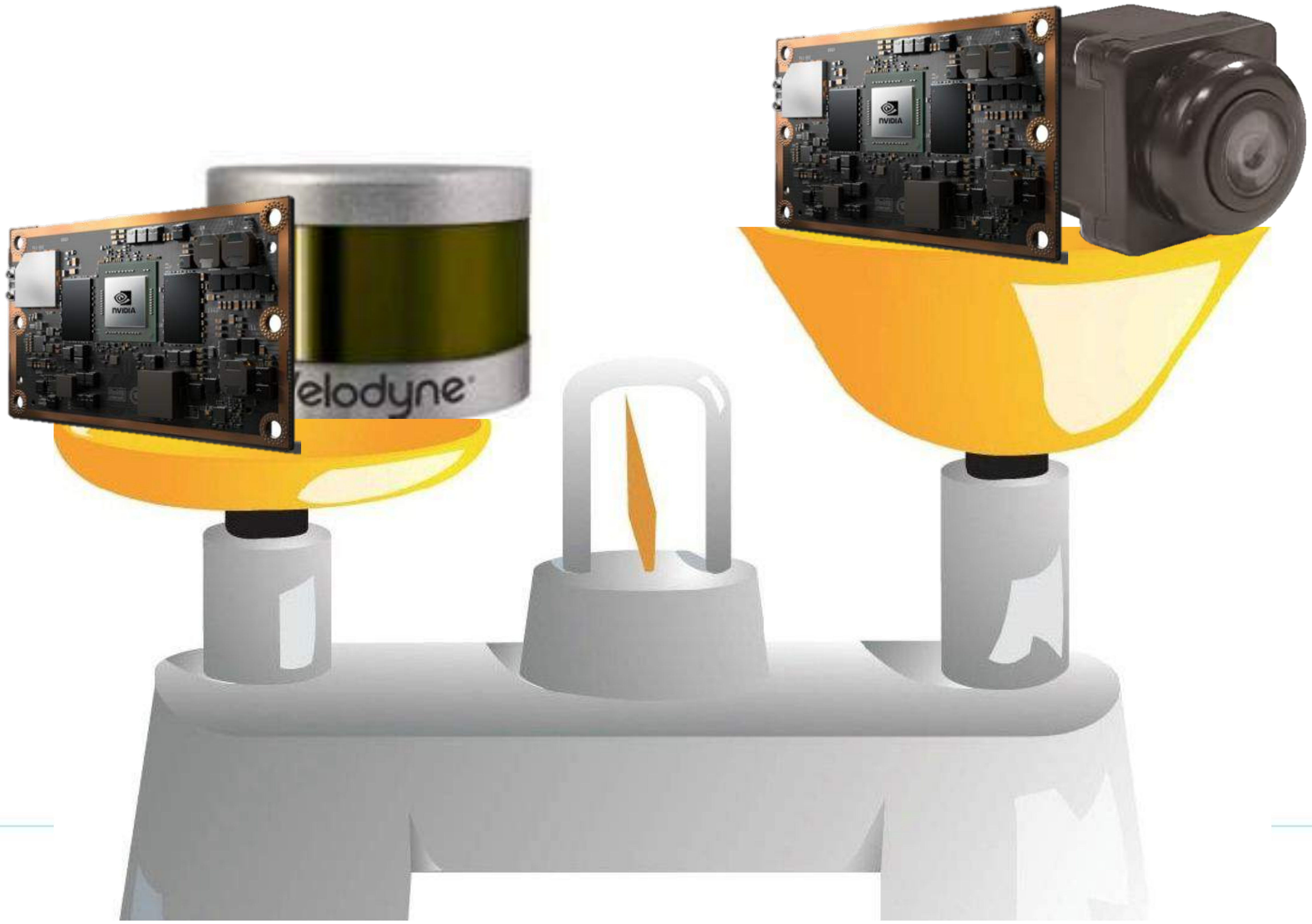


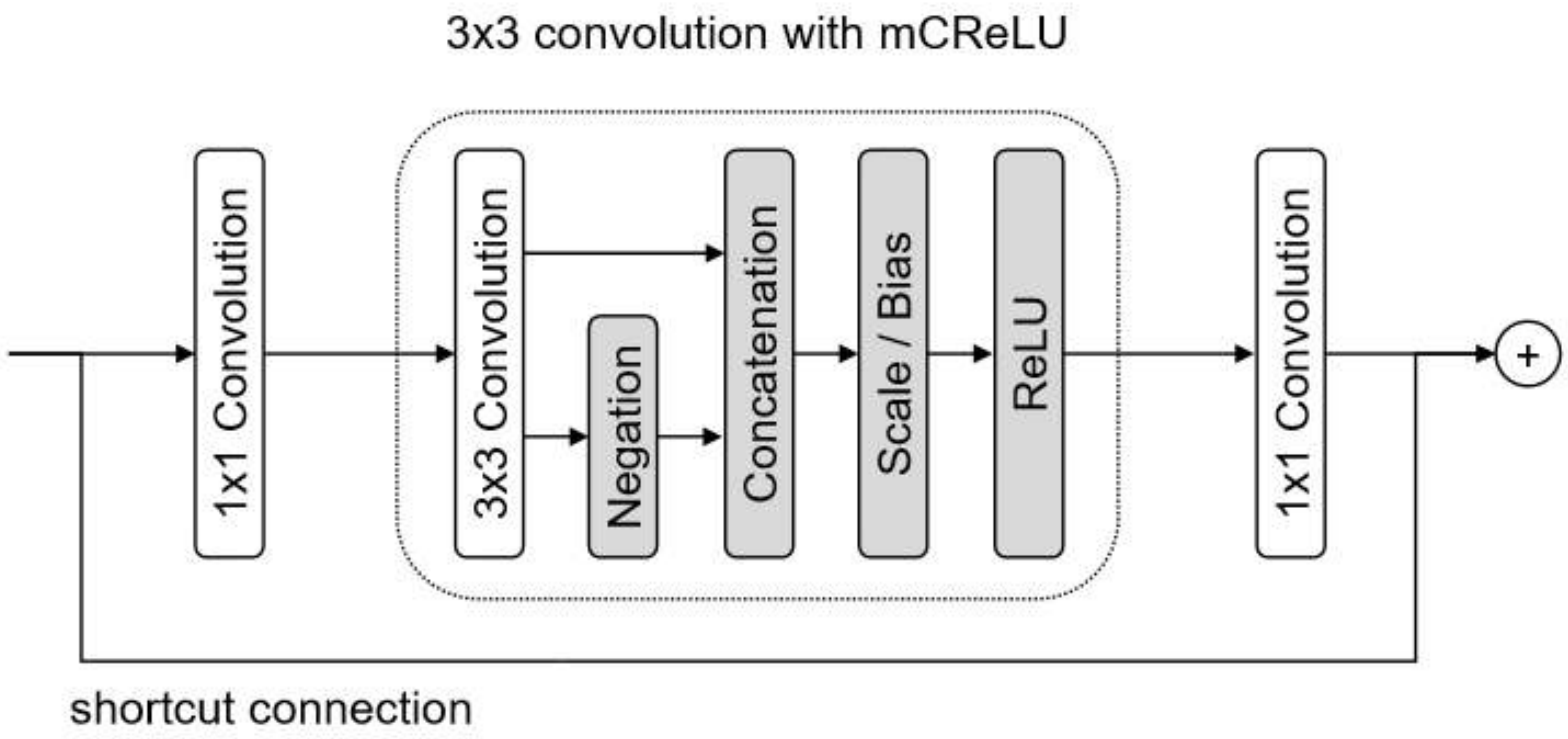
视觉感知从demo到deploy

- Demo：精度要高，不计成本，不管标准，不算功耗
- Deploy：低成本，低功耗，合车规，实时性，精度用户满意

硬件更便宜？

硬件更便宜



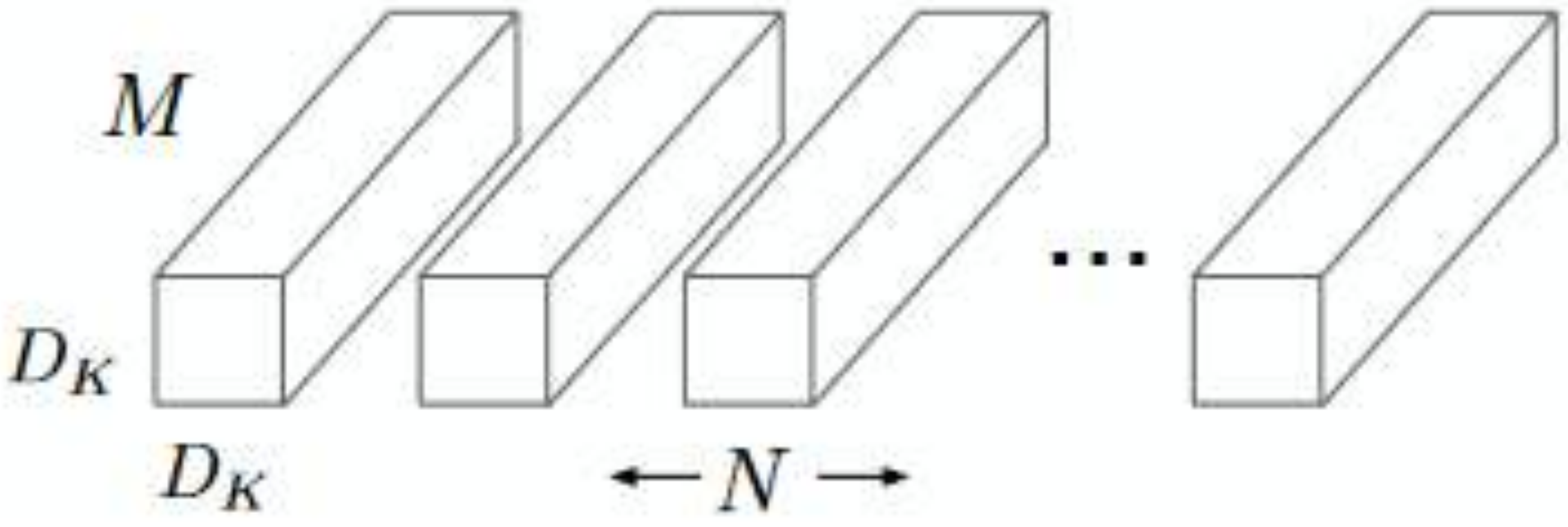


PVANet

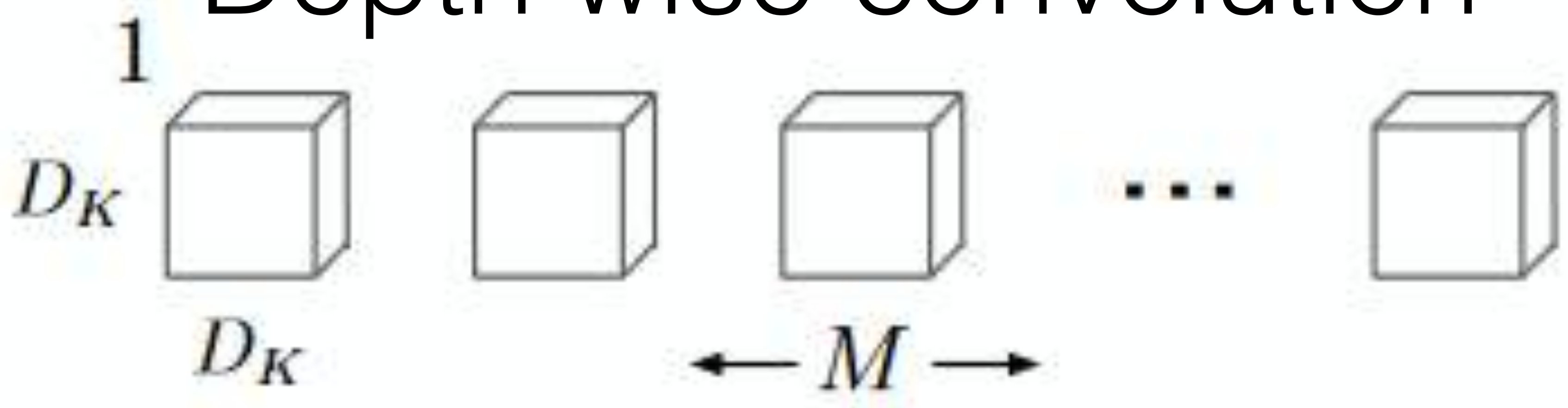
Hong S, Roh B, Kim K H, et al. PVANet: Lightweight Deep Neural Networks for Real-time Object Detection. arXiv, 2016.

PVANet+fasterRCNN物体检测

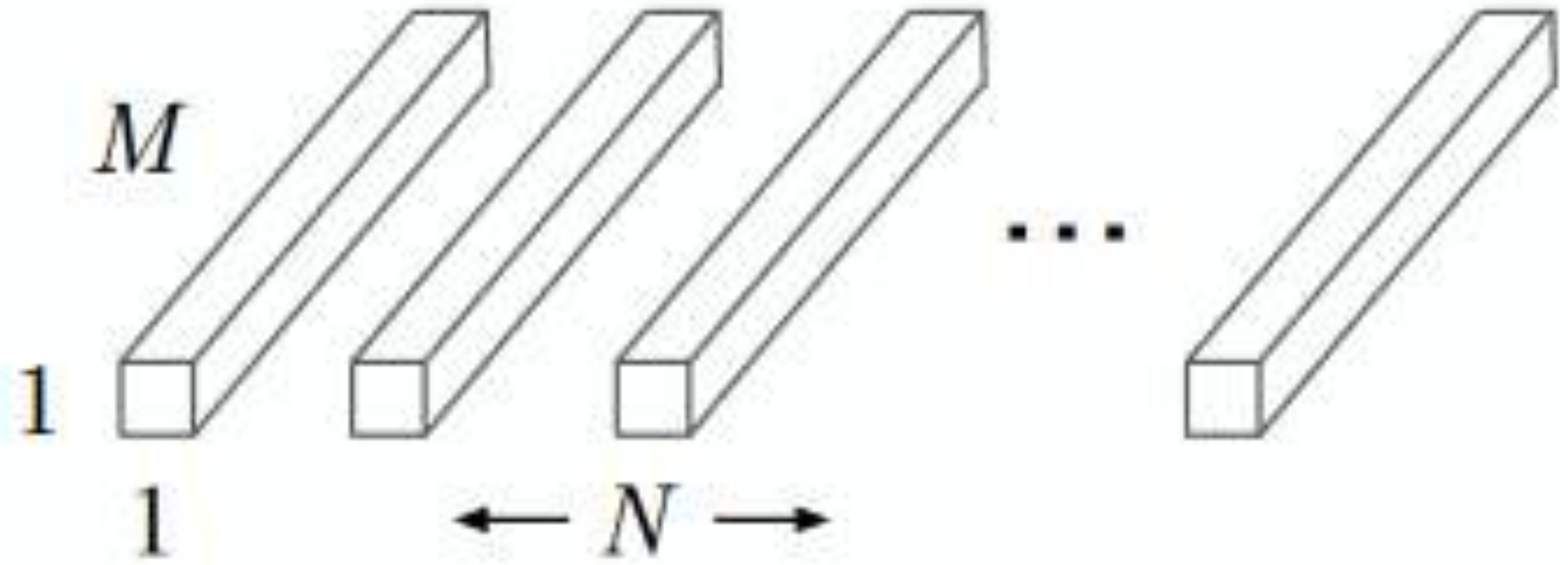
Standard convolution



Depth wise convolution



1x1 convolution

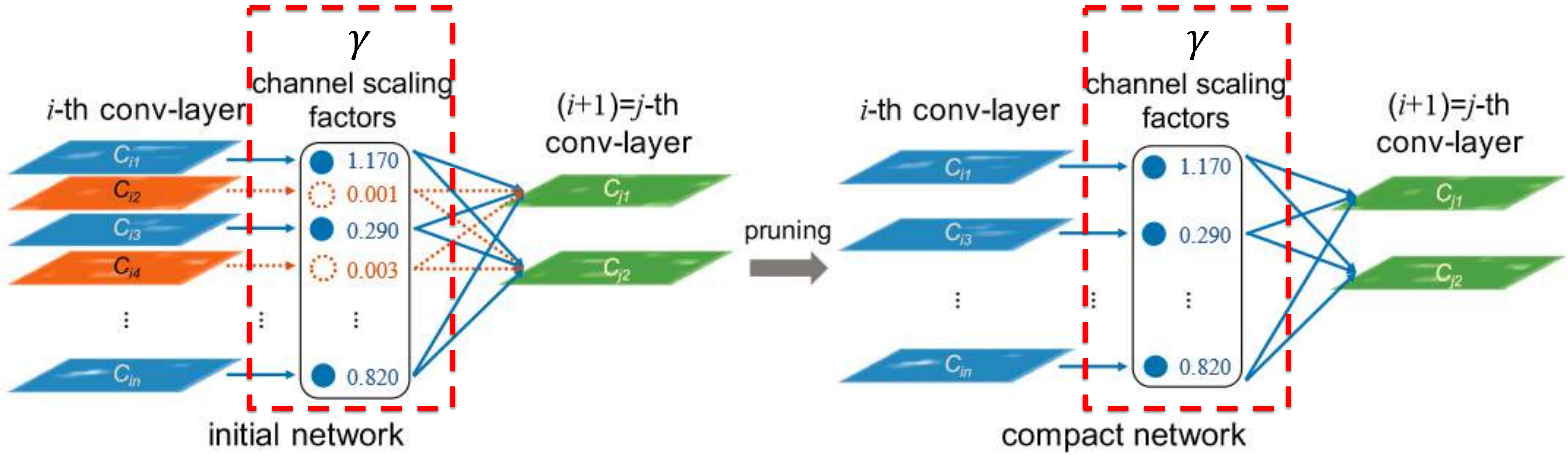


MobileNet

$$\frac{std.conv}{dw\ conv + 1x1\ conv} = \frac{1}{D_k^2} + \frac{1}{N}$$

Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv, 2017.

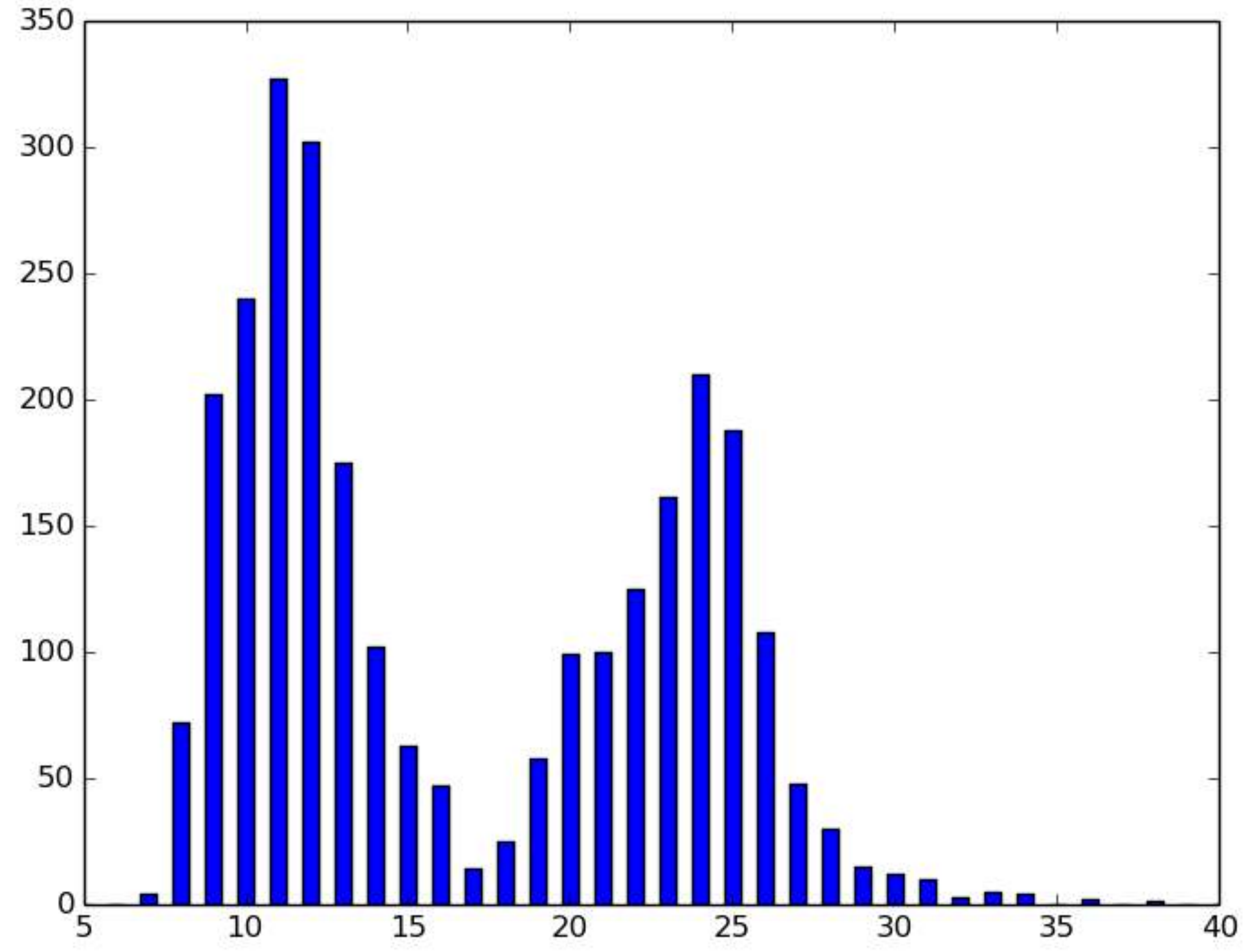
TABLE OF CONTENTES



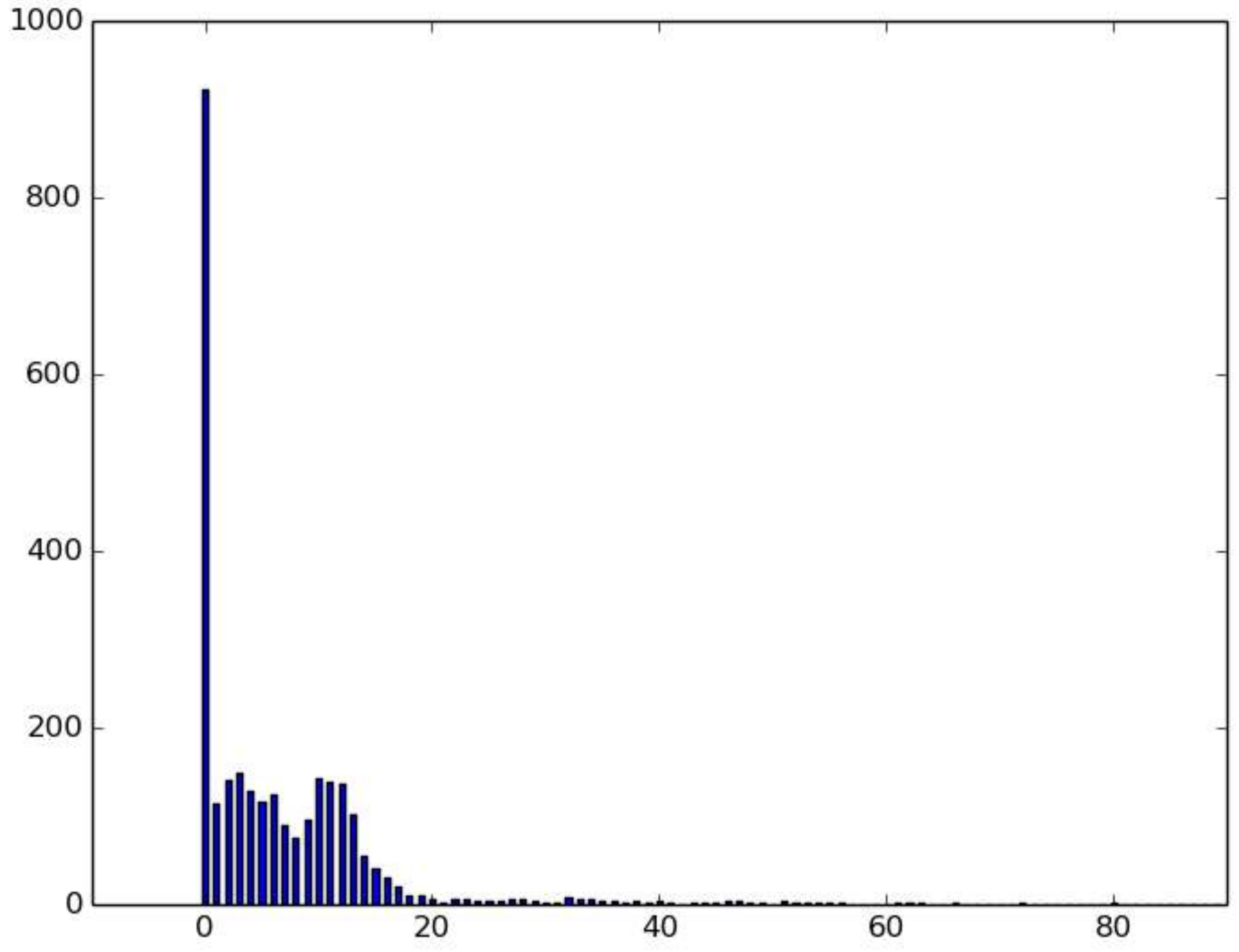
$$L(w) = \sum_{i} l(f(x_i, w), y_i) + \lambda \sum_{\gamma \in \Gamma} ||\gamma||_1$$

Network slimming

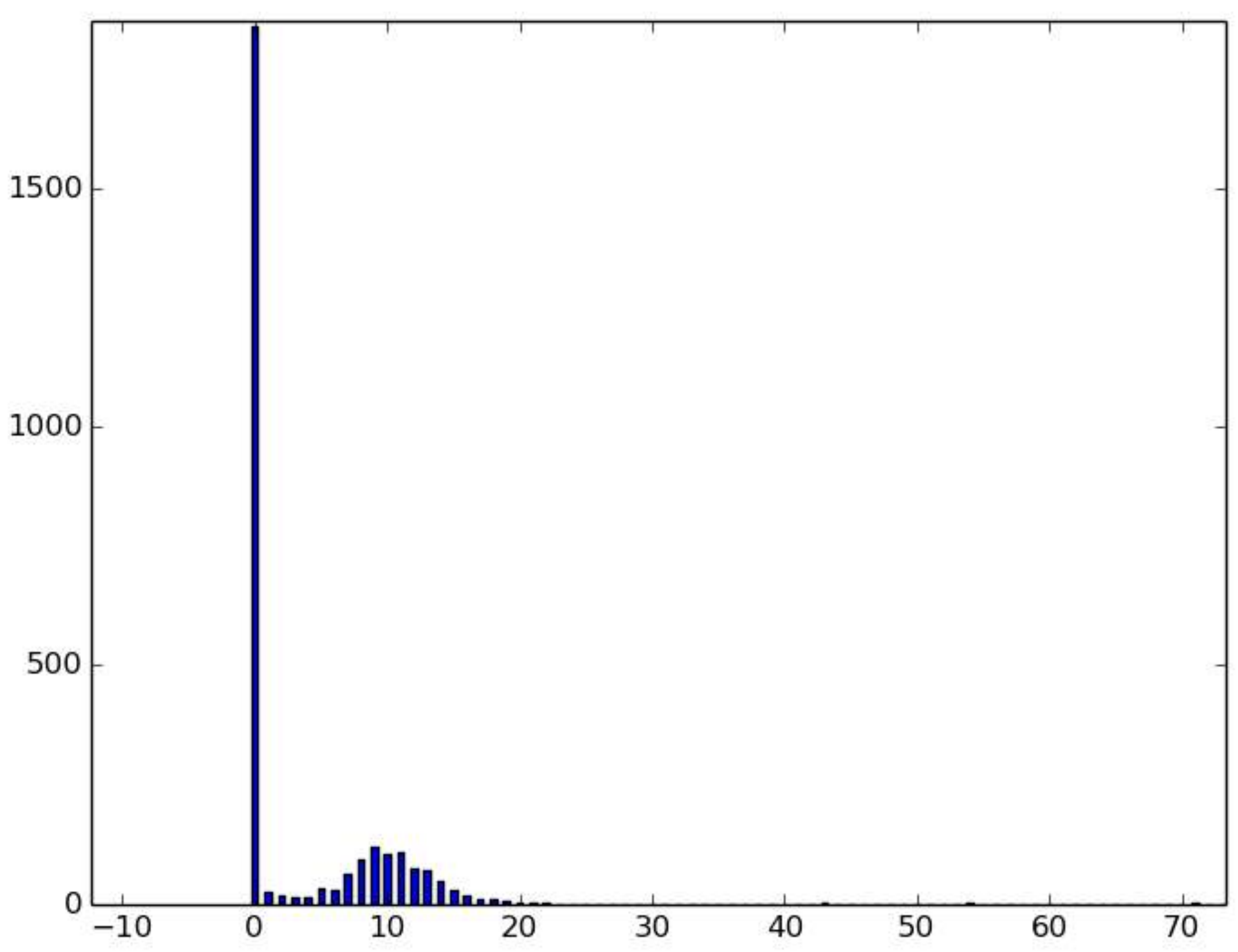
Liu Z, Li J, Shen Z, et al. Learning Efficient Convolutional Networks through Network Slimming. arXiv, 2017.



$\gamma = 0$

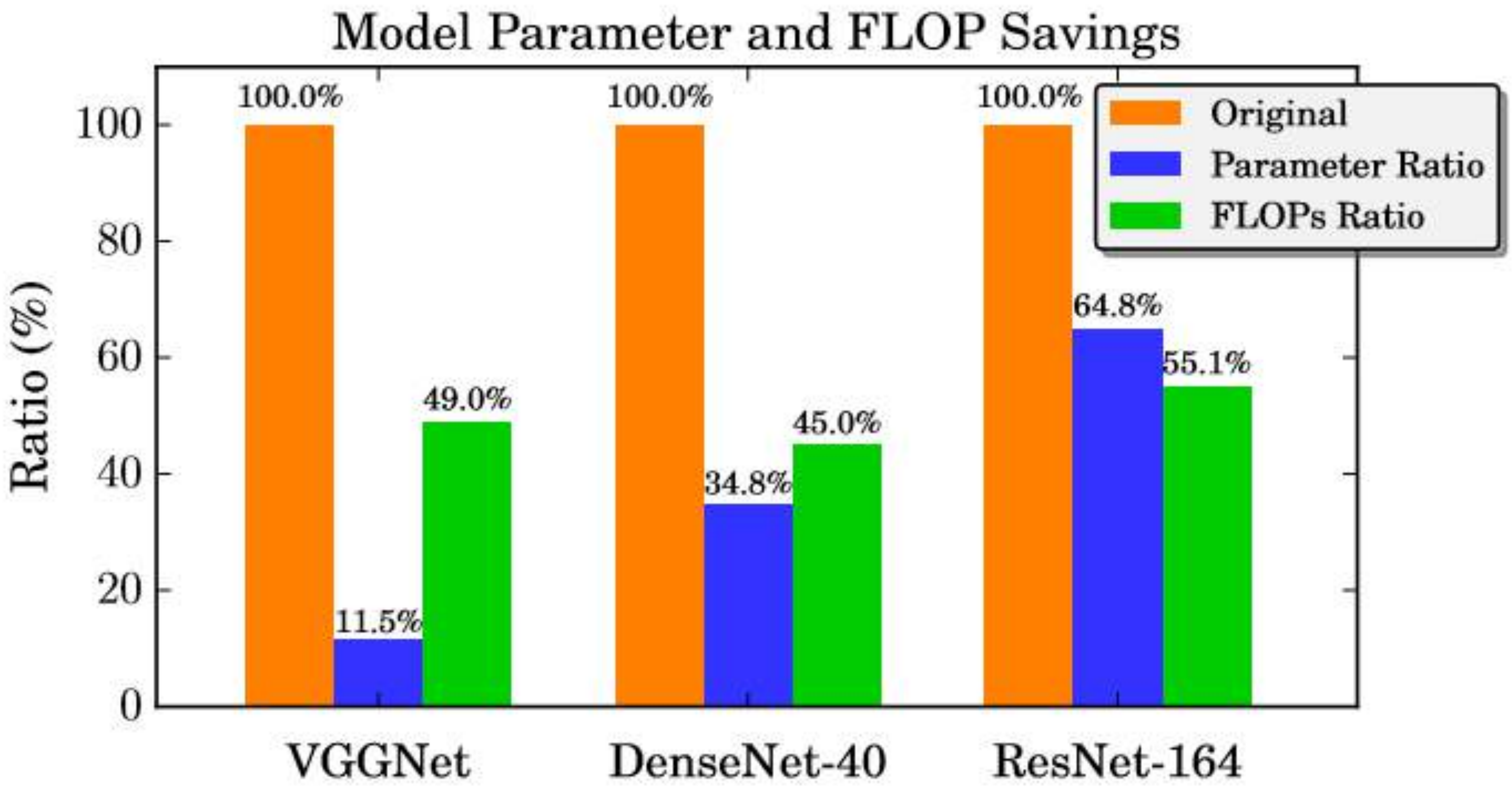


$\gamma = 0.0001$



$\gamma = 0.001$

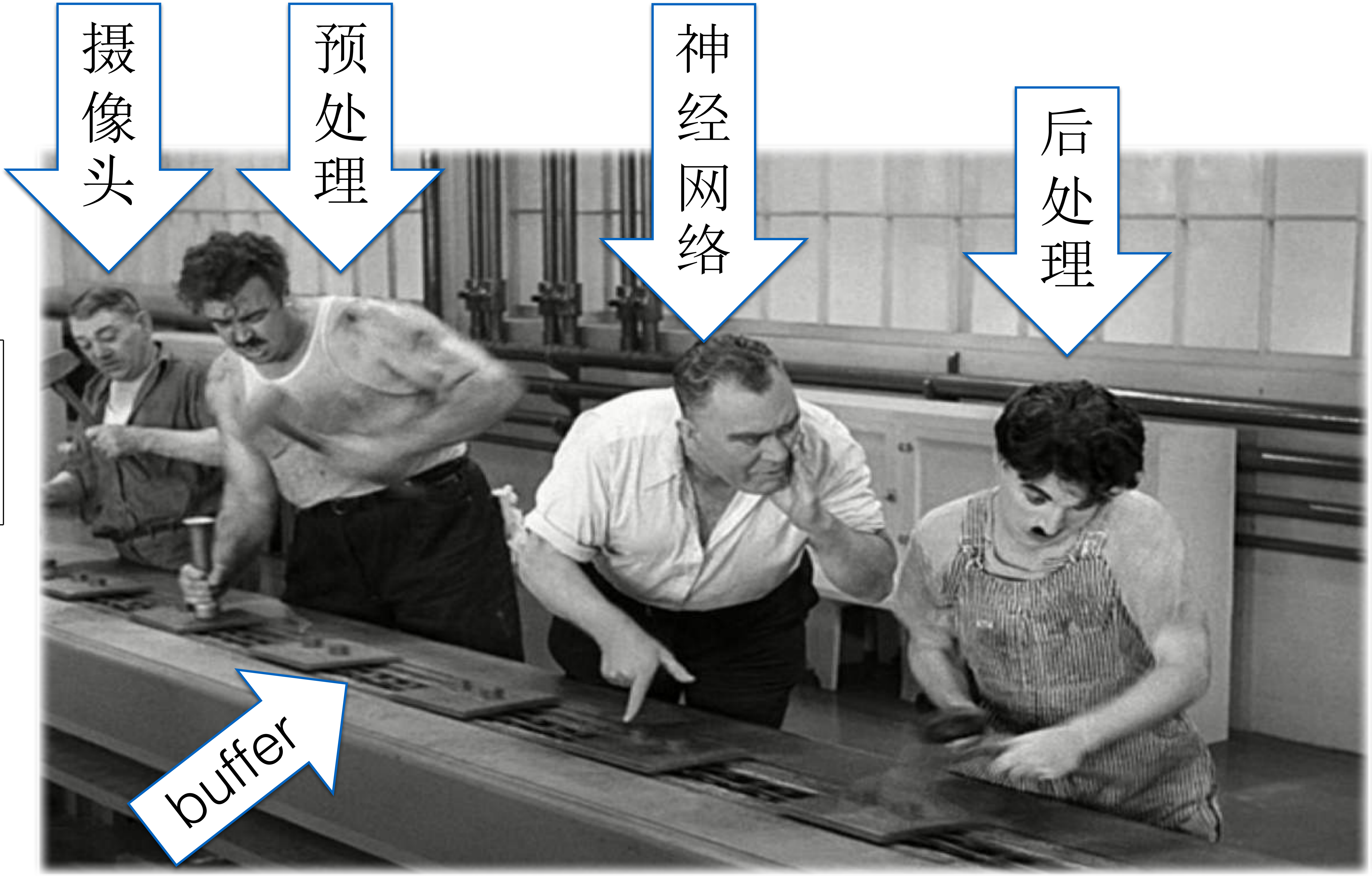
网络参数稀疏化效果


Model Parameter and FLOP Savings
Ratio (%)
100
80
60
40
20
0
100.0%
11.5%
49.0%
100.0%
34.8%
45.0%
100.0%
64.8%
55.1%
Original
Parameter Ratio
FLOPs Ratio
VGGNet
DenseNet-40
ResNet-164

网络压缩结果

TABLE OF CONTENTES

当前无法显示该图像。



Pipeline

TensorRT

- FP16和INT8自动量化
- 多层合并
- 自动选择并行算法
- 显存动态优化
- 多任务并发

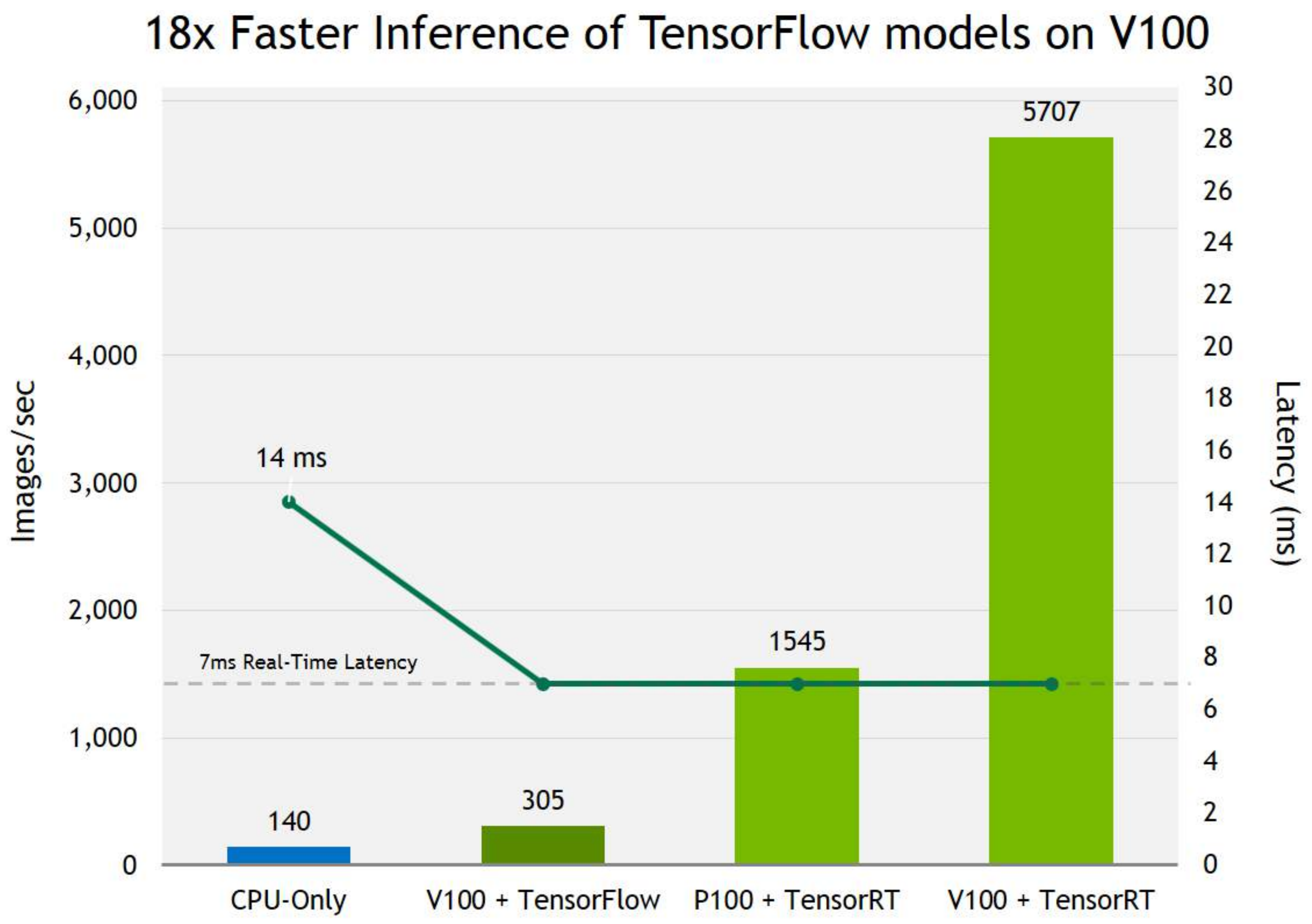


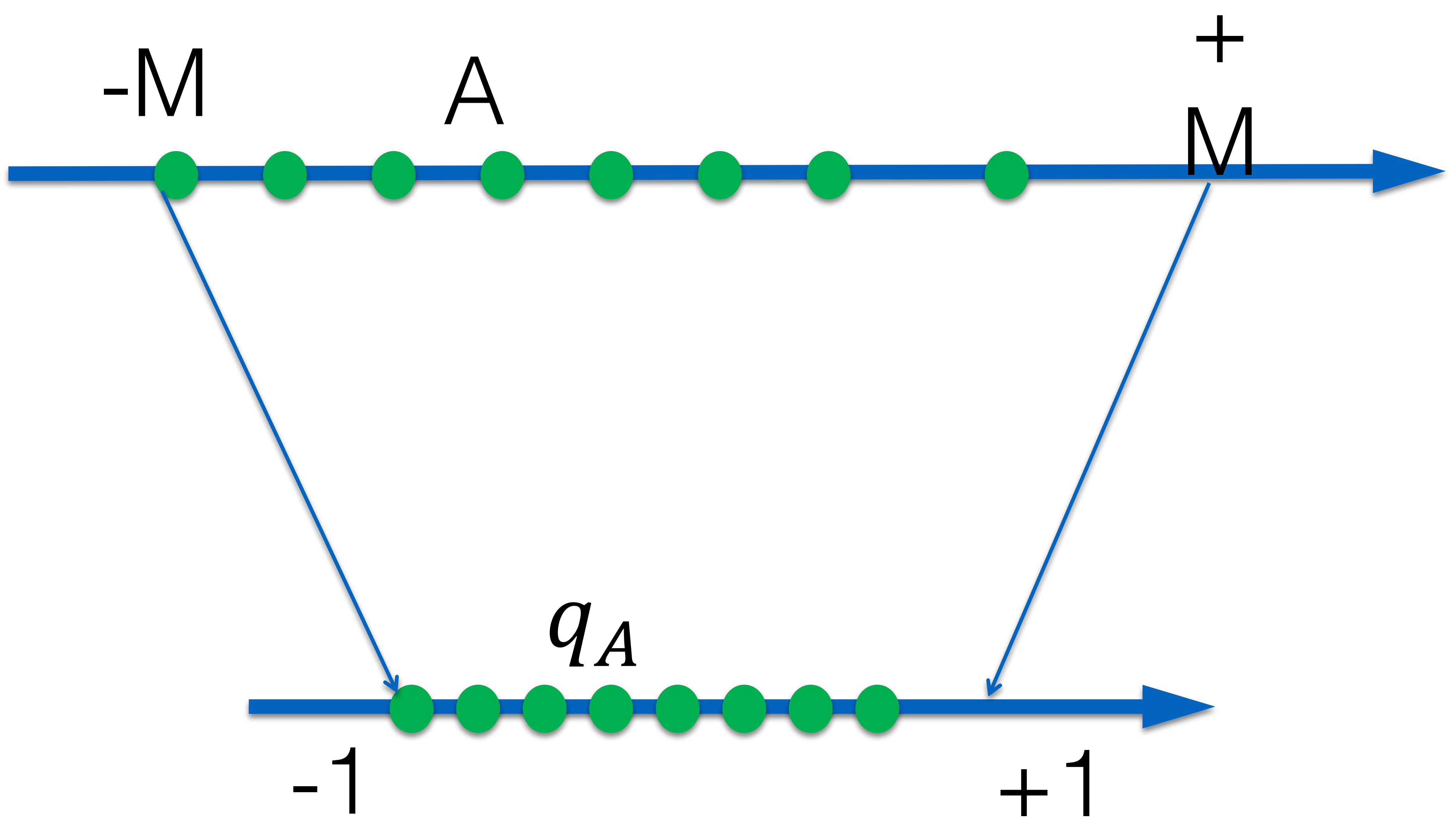
TABLE OF CONTENTES

FPGA定点化

- 表示范围FP32：-3.4×10^{38}~3.4×10^{38}，INT8：-128~127
- 定点小数$A=(A_0)A_1\cdots A_k.\cdots A_n$ (A_0为符号位，A_i为0/1)代表的小数为
$$(-1)^{A_0}[A_1*2^{k-1}+A_2*2^{k-2}+\cdots+A_k*2^0+\cdots+A_n*2^{k-n}]$$
- 定点小数表示范围在$\pm2^k(1-0.5^n)$之间，精度（最小单位）为2^{k-n}
- 用INT8定点表示FP32： $A=2^k*(-1)^{X_0}*0.A_1A_2\cdots A_n$

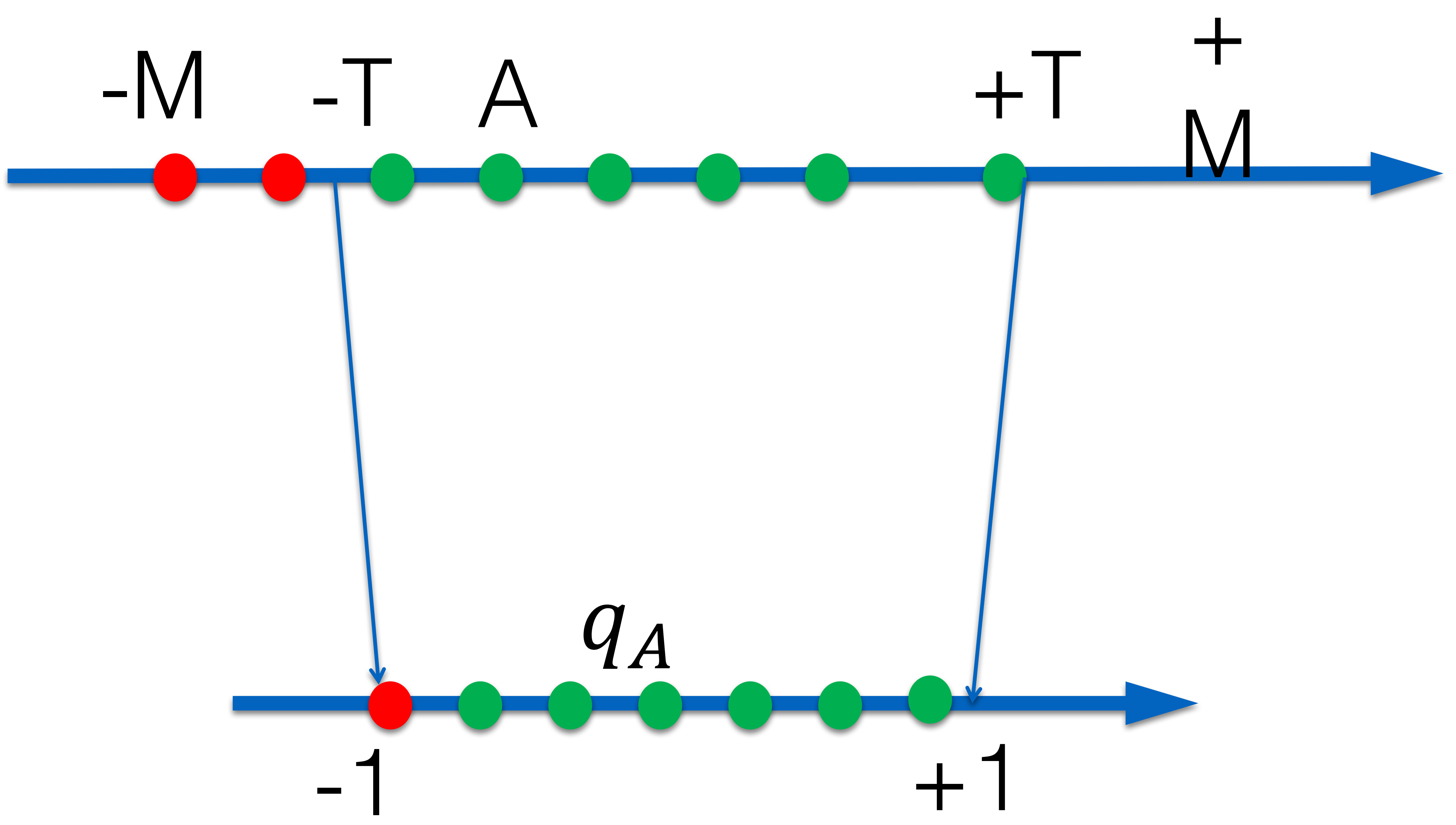
FP32 Value A= FP32 scale factor s_A * INT8 Value q_A

- $\sum_j A_jB_j=s_As_B\sum_j q_{Aj}q_{Bj}$



$M = \max|A| \leq 2^k$

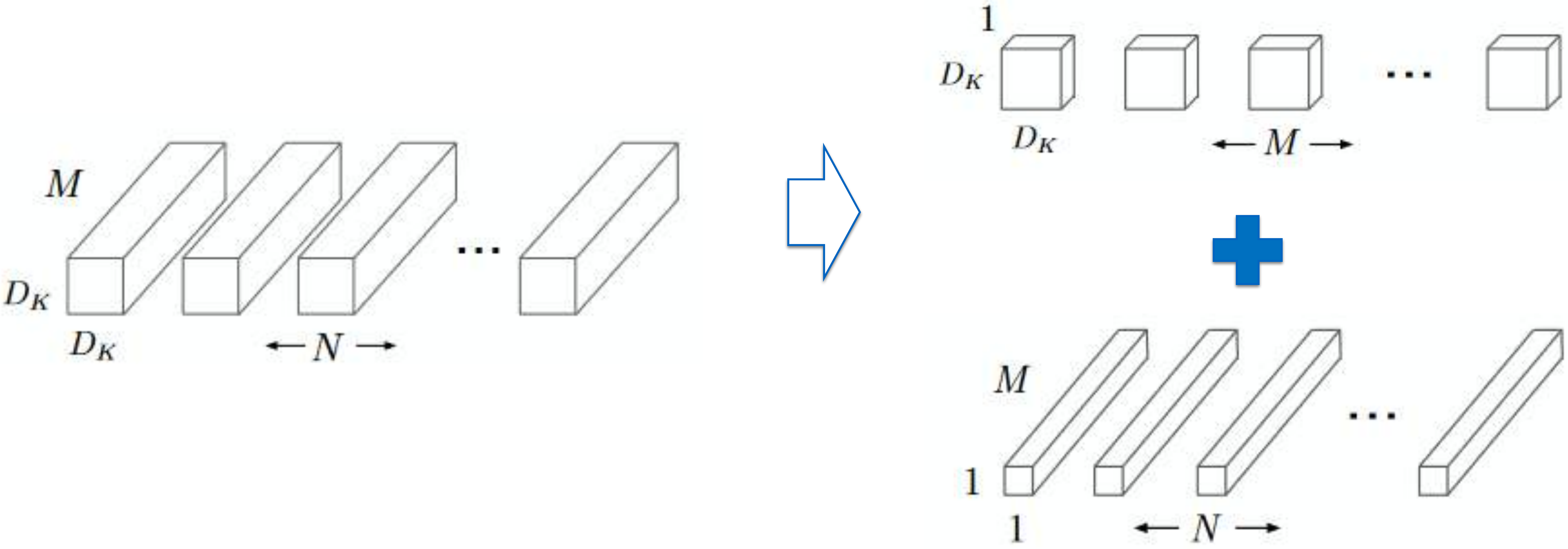
表示范围大，精度差



$M > T = 2^k$

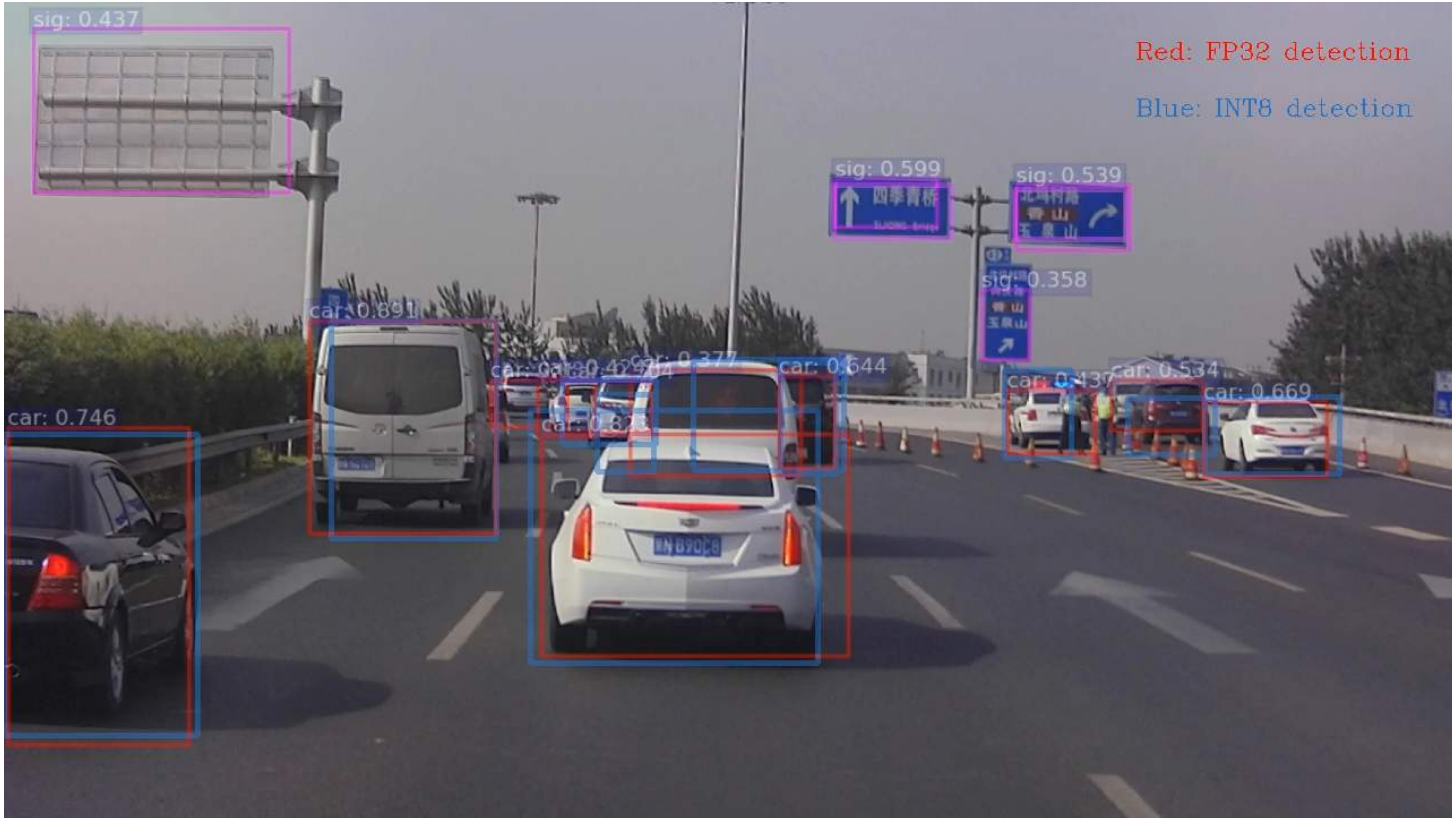
表示范围小，精度好

表示范围与精度的取舍



FPGA的网络选择

- MobileNet使用depth wise convolution+1x1 convolution
- 理论计算量低，同时精度很高
- GPU加速比比较差，但适合CPU和定制计算设备



FPGA+MobileNet物体检测

UISEE Visual Perception demo video

Thanks!