

如何构建低成本高效能的 视觉感知系统

潘争

驭势科技



极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新



扫一扫下载极客时间App

人工智能基础课

“通俗易懂的人工智能入门课”

王天一
博士 副教授



扫一扫，免费试读

AI技术内参

你的360度人工智能信息助理

洪亮劫
Etsy 数据科学主管



扫一扫，免费试读



关注落地技术，探寻AI应用场景

- 14万AI领域垂直用户
- 8000+社群技术交流人员，不乏行业内顶级技术专家
- 每周一节干货技术分享课
- AI一线领军人物的访谈
- AI大会的专家干货演讲整理
- 《AI前线》月刊
- AI技能图谱
- 线下沙龙



扫码关注带你涨姿势

QCon

全球软件开发大会

成为软件技术专家 的必经之路

[北京站] 2018

会议：2018年4月20-22日 / 培训：2018年4月18-19日

北京·国际会议中心

8折

购票中, 每张立减1360元

团购享受更多优惠



识别二维码了解更多

ArchSummit

全球架构师峰会

2018 · 深圳站

从2012年开始算起，InfoQ已经举办了9场ArchSummit全球架构师峰会，有来自Microsoft、Google、Facebook、Twitter、LinkedIn、阿里巴巴、腾讯、百度等技术专家分享过他们的实践经验，至今累计已经为中国技术人奉上了近千场精彩演讲。

限时**7折**报名中，名额有限，速速报名吧！

● 2012.08.10-12 深圳站



2018.07.06-09 深圳站

会议：07.06-07.07

培训：07.08-07.09



TABLE OF CONTENTES

效率精度平衡的卷积网络

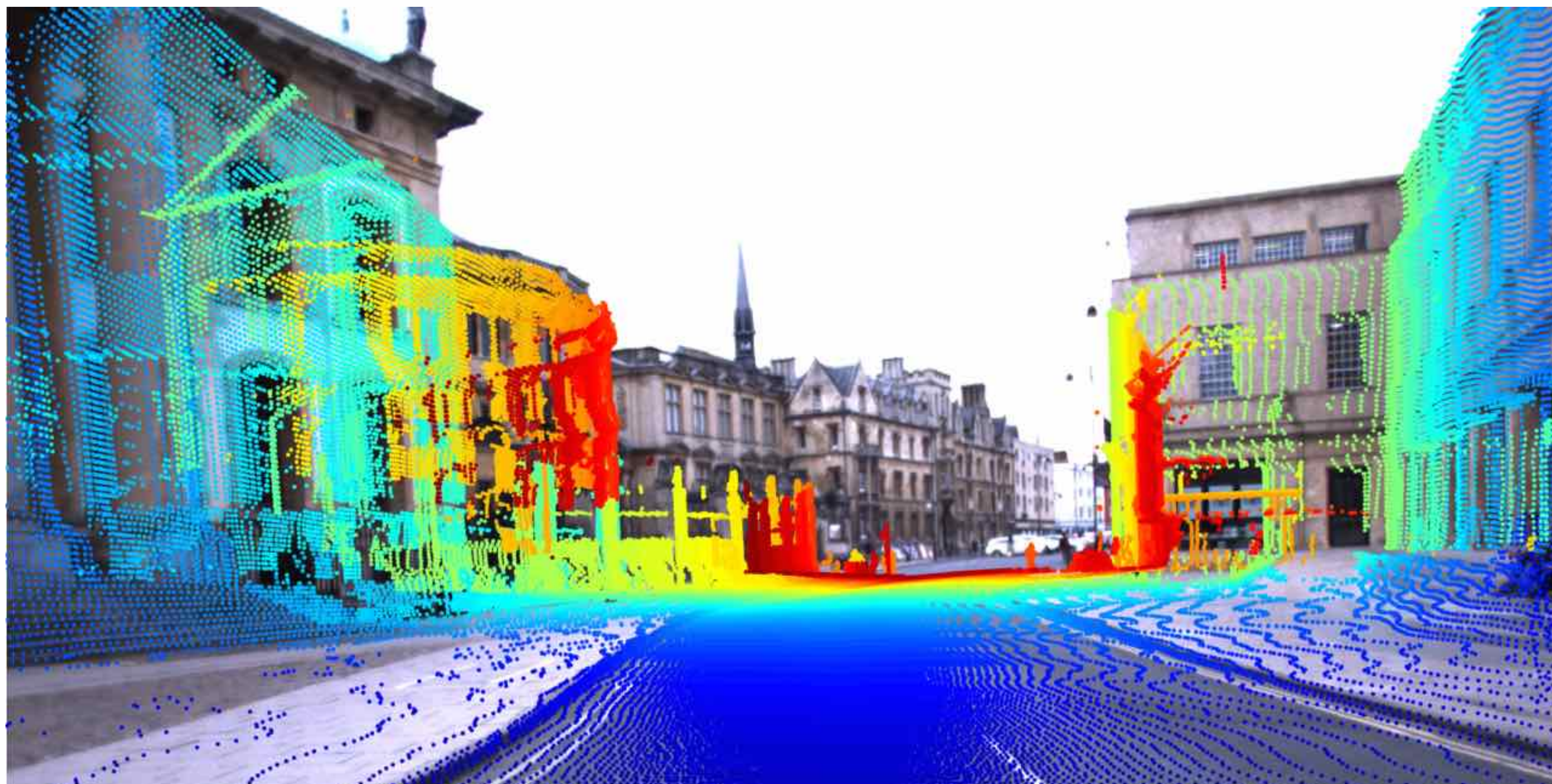
卷积网络的压缩

嵌入式GPU+CPU的加速

低成本FPGA的加速

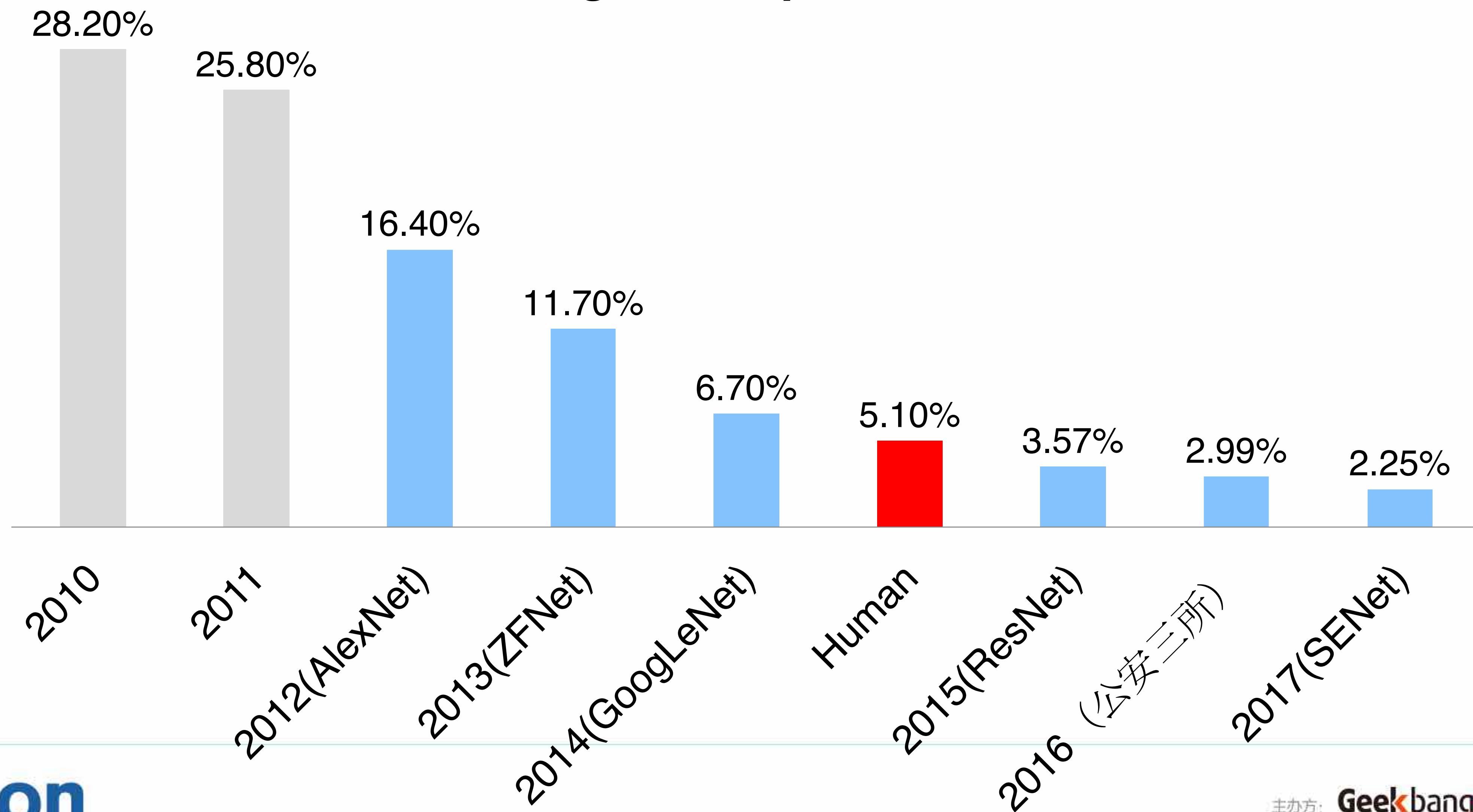
视觉感知的优势

- 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

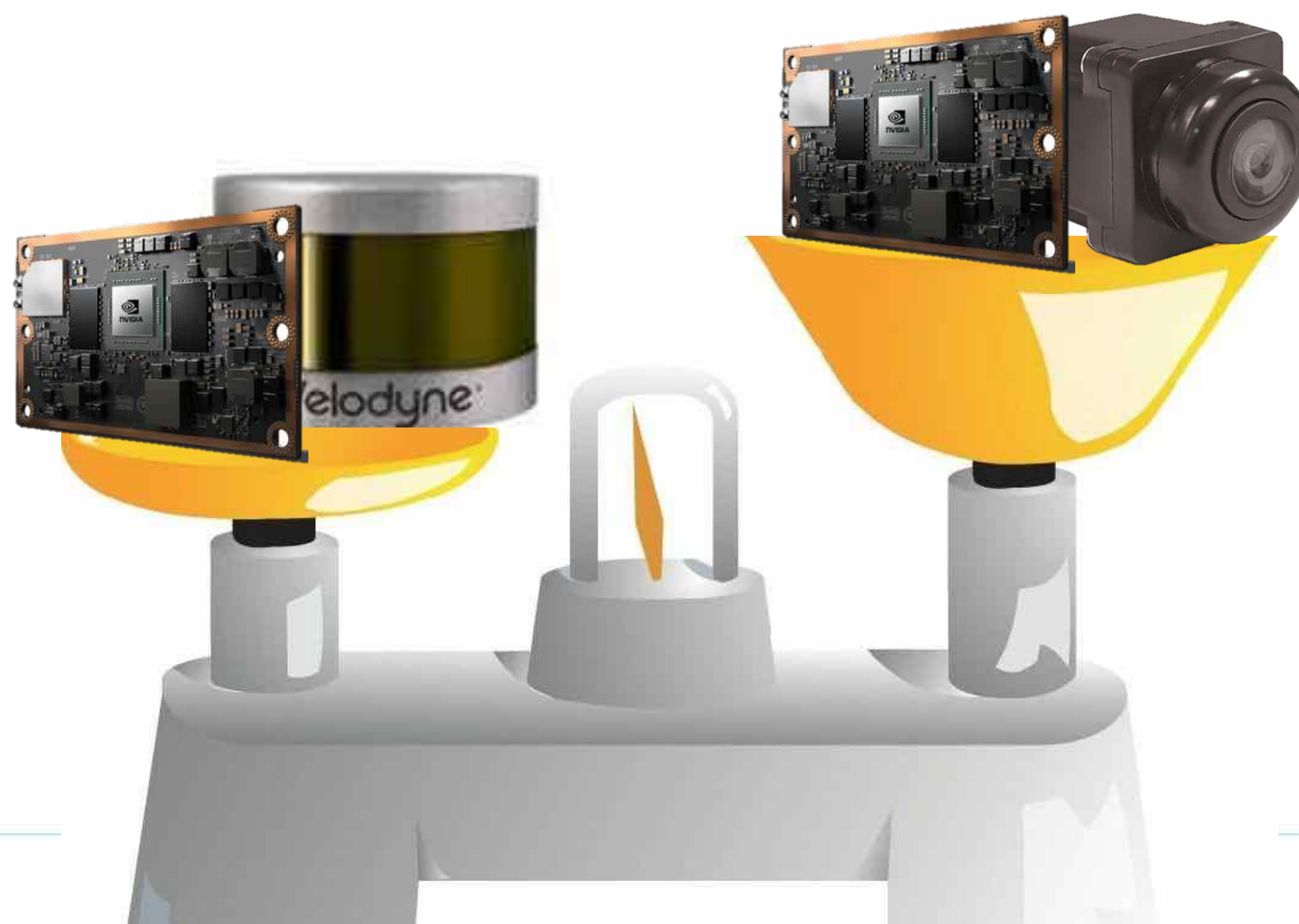
- Demo：精度要高，不计成本，不管标准，不算功耗
- Deploy：低成本，低功耗，合车规，实时性，精度用户满意

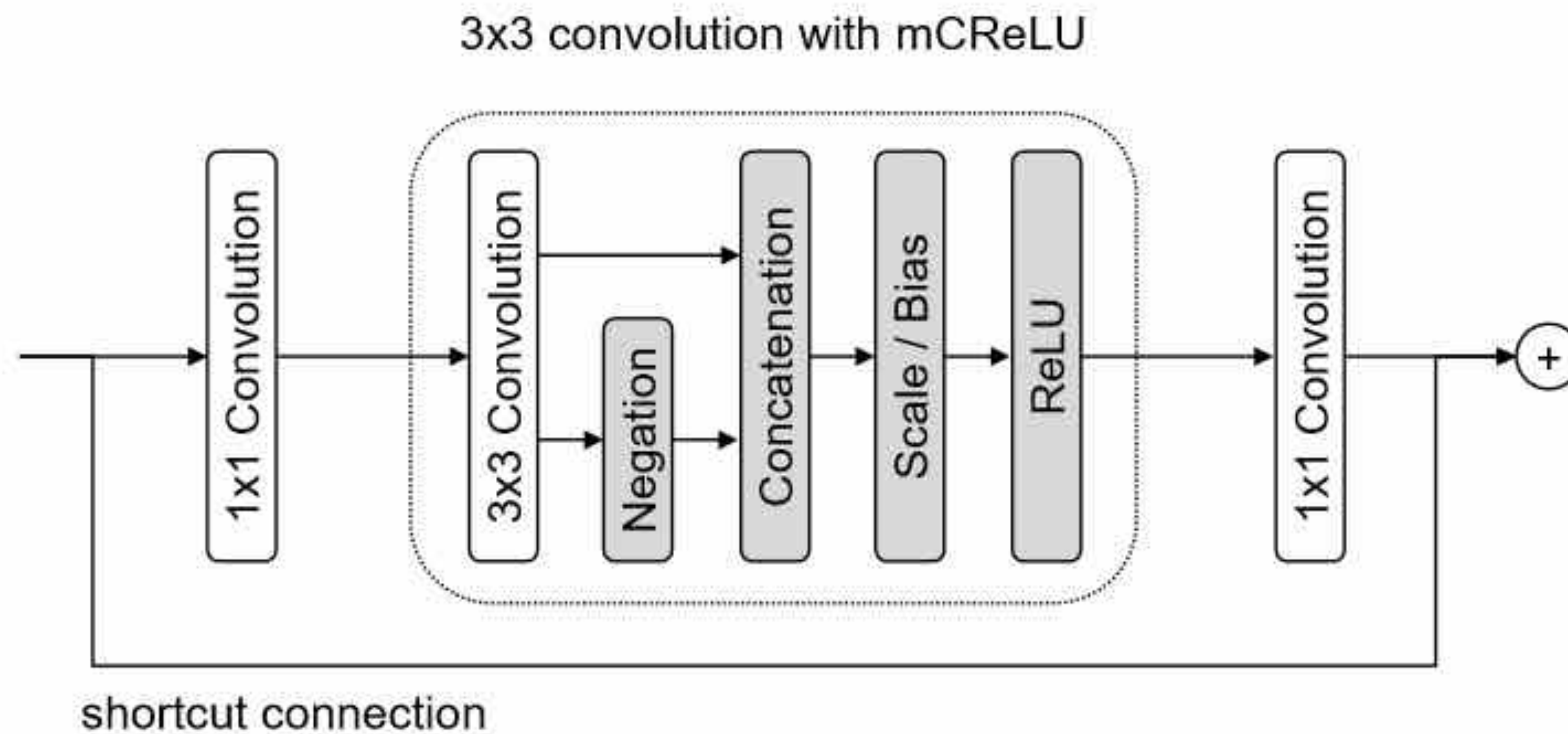
硬件更便宜？

3999美元



硬件更便宜





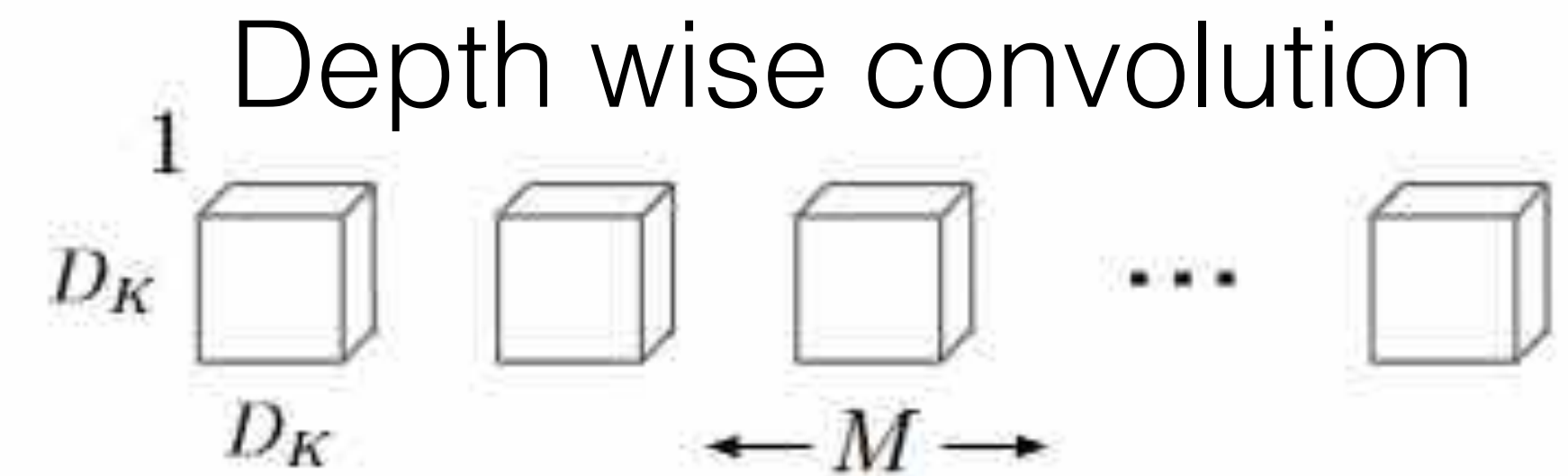
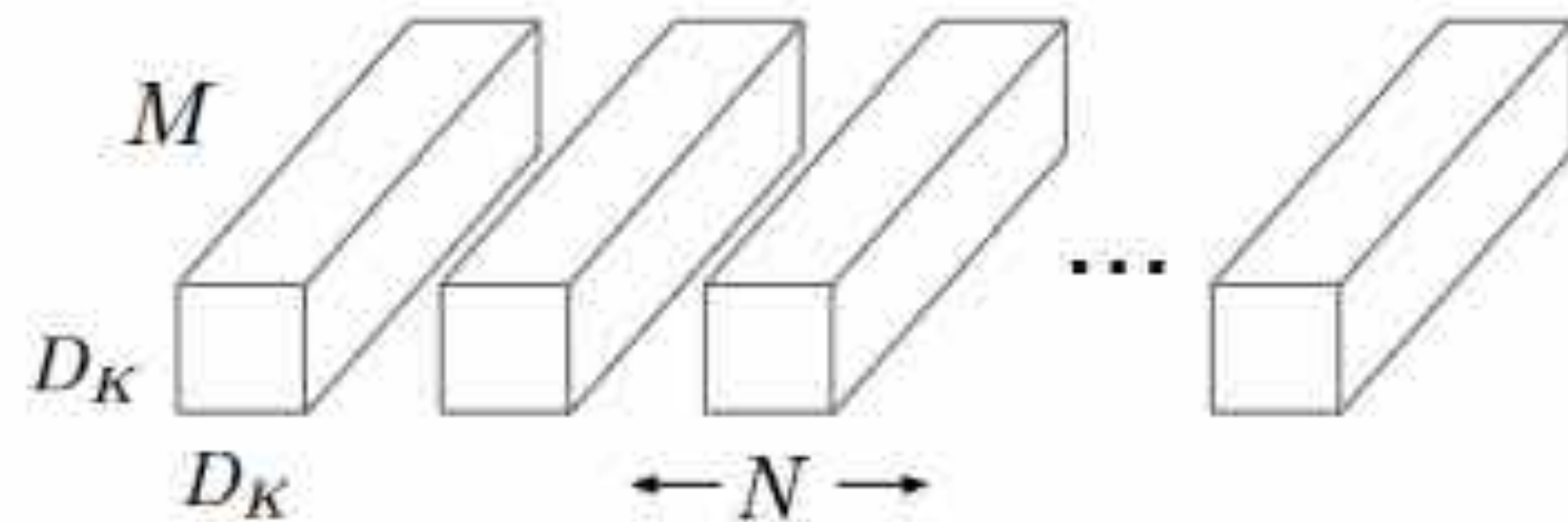
PVANet

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

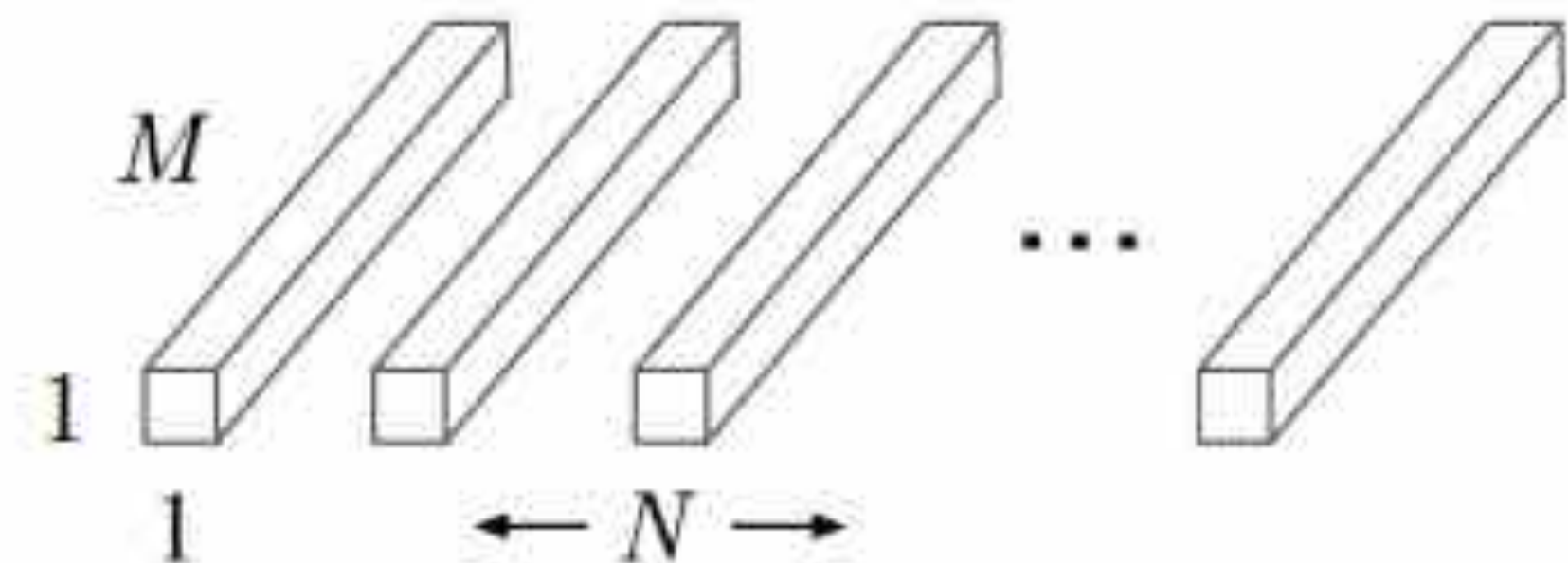


PVANet+fastrRCNN物体检测

Standard convolution



1x1 convolution



MobileNet

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

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

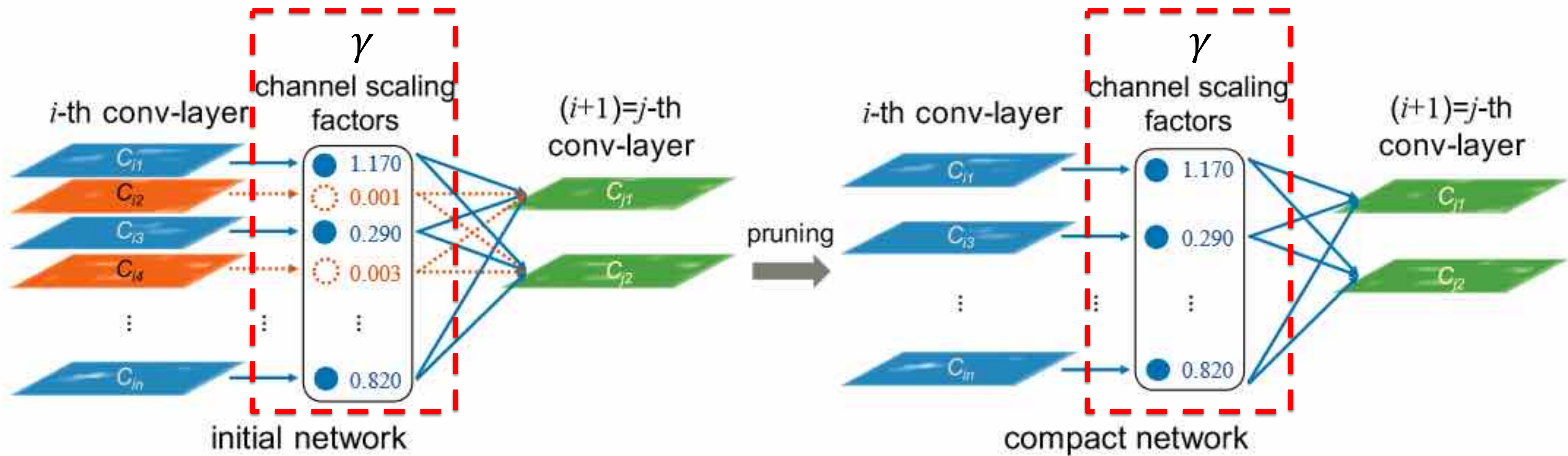
TABLE OF CONTENTES

效率精度平衡的卷积网络

卷积网络的压缩

嵌入式GPU+CPU的加速

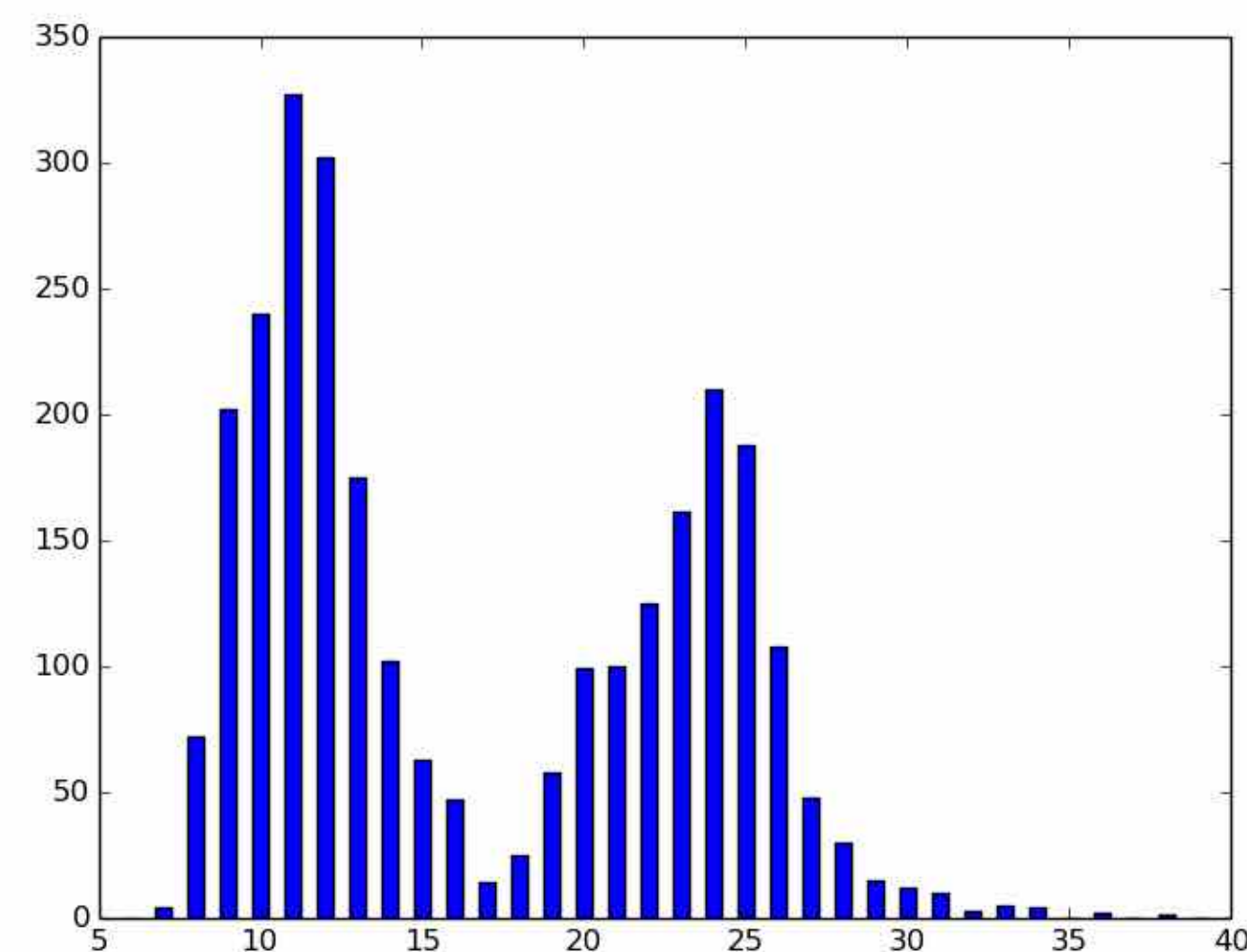
低成本FPGA的加速



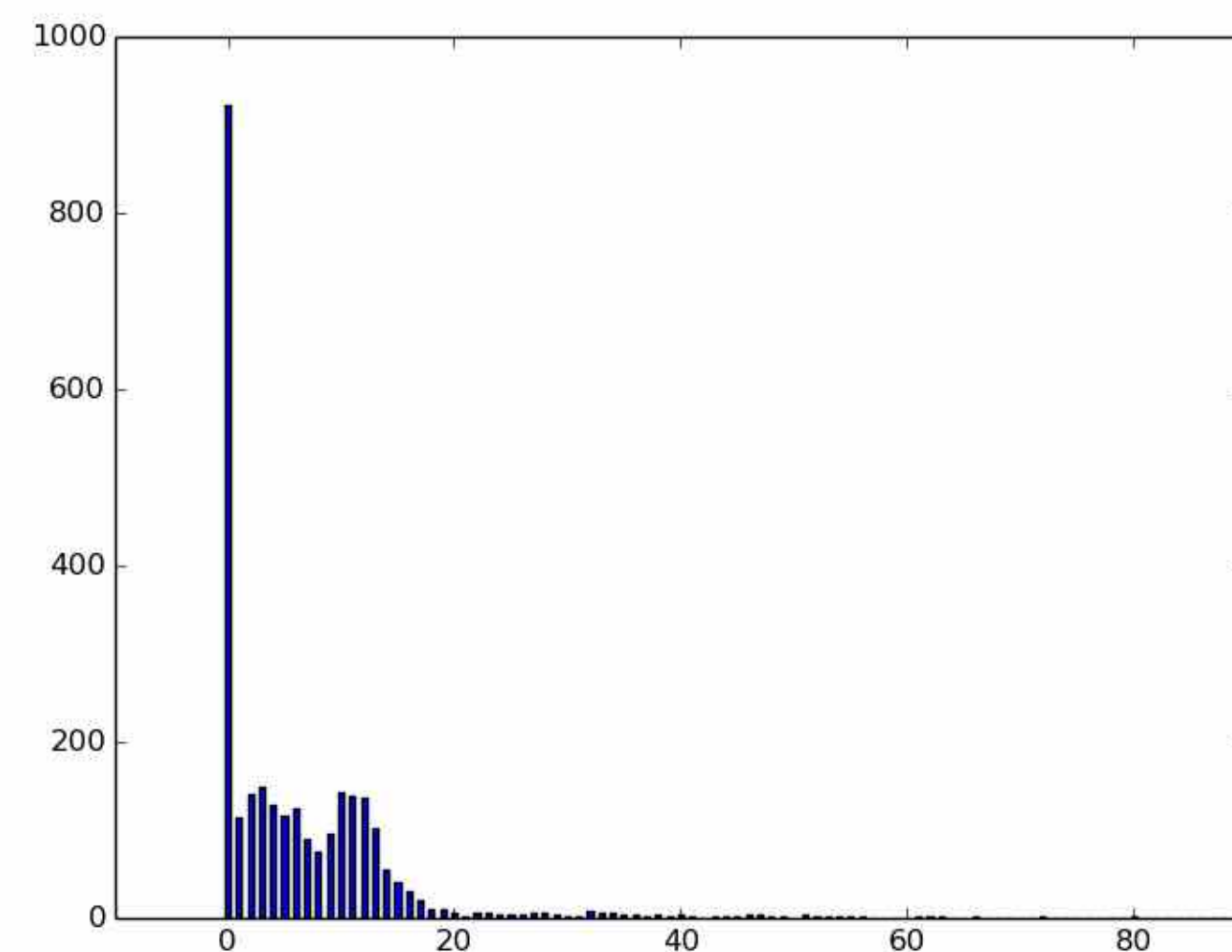
$$L(w) = \sum_i l(f(x_i, w), y_i) + \lambda \sum_{\gamma \in \Gamma} \|\gamma\|_1$$

Network slimming

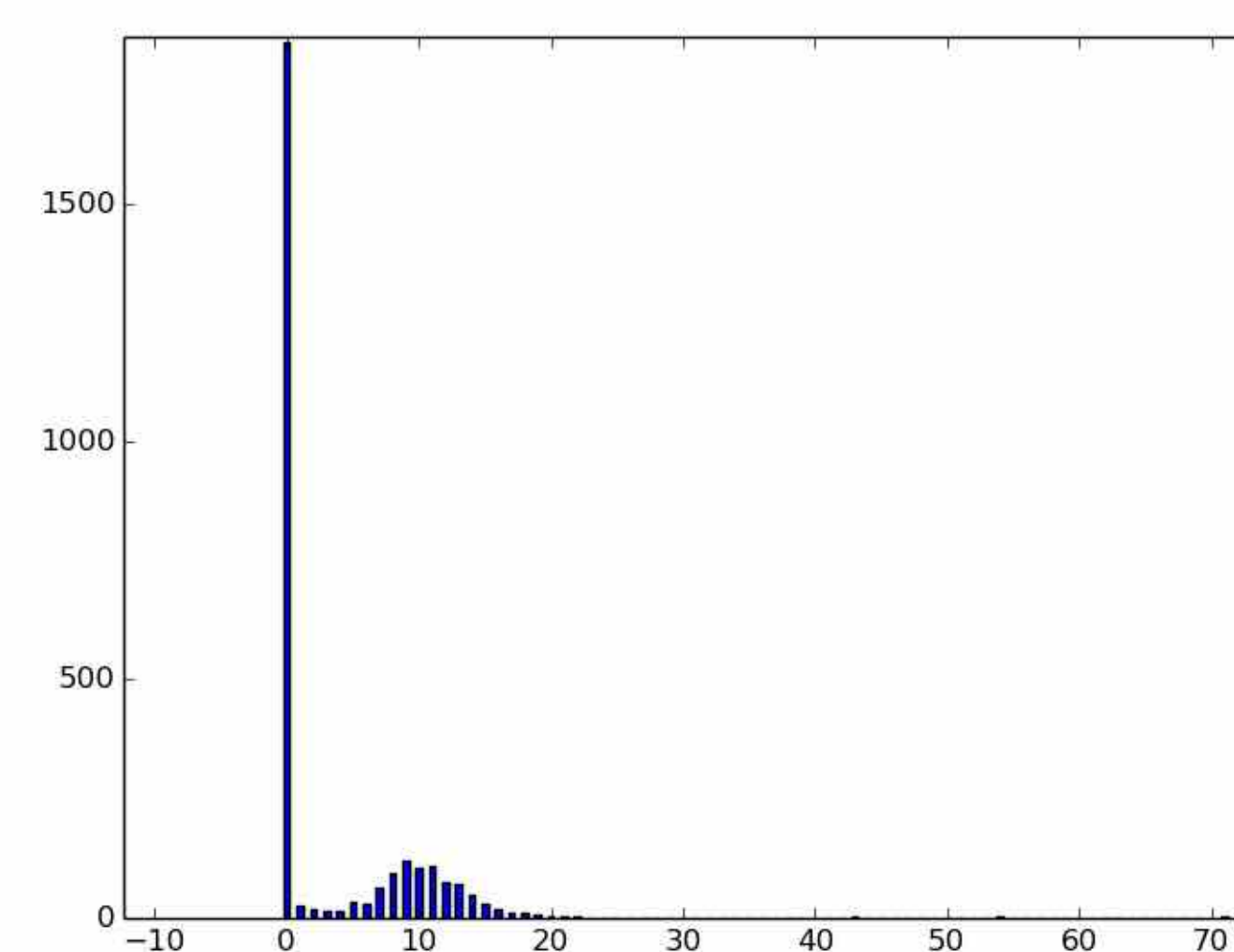
Liu Z, Li J, Shen Z, et al. Learning Efficient Convolutional Networks through Network Slimming. arXiv, 2017.



$$\gamma = 0$$

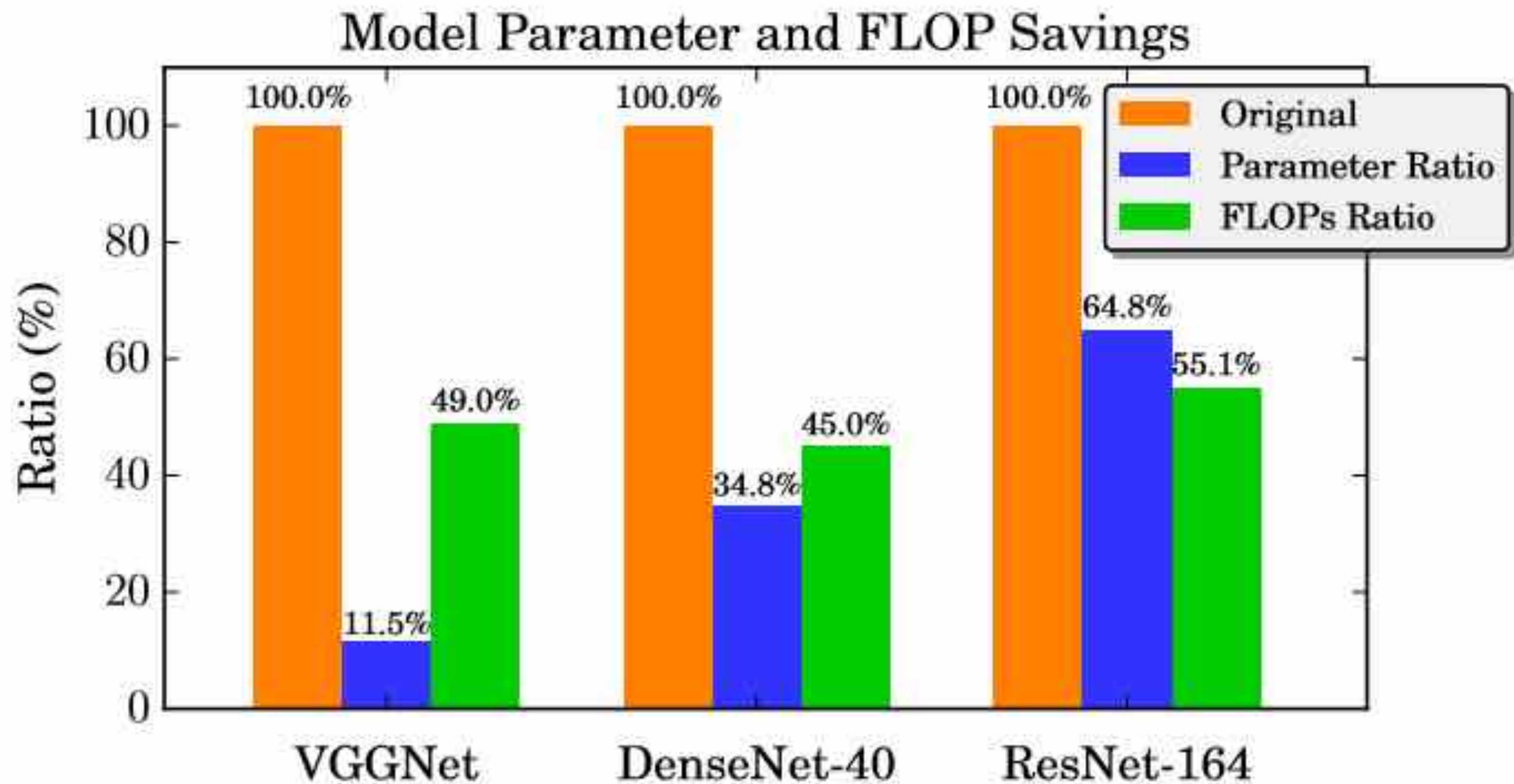


$$\gamma = 0.0001$$



$$\gamma = 0.001$$

网络参数稀疏化效果



网络压缩结果

TABLE OF CONTENTES

效率精度平衡的卷积网络

卷积网络的压缩

嵌入式GPU+CPU的加速

低成本FPGA的加速

摄像头

预处理

神经网络

后处理

buffer

Pipeline

当前无法显示该图像。

当前无法显示该图像。

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

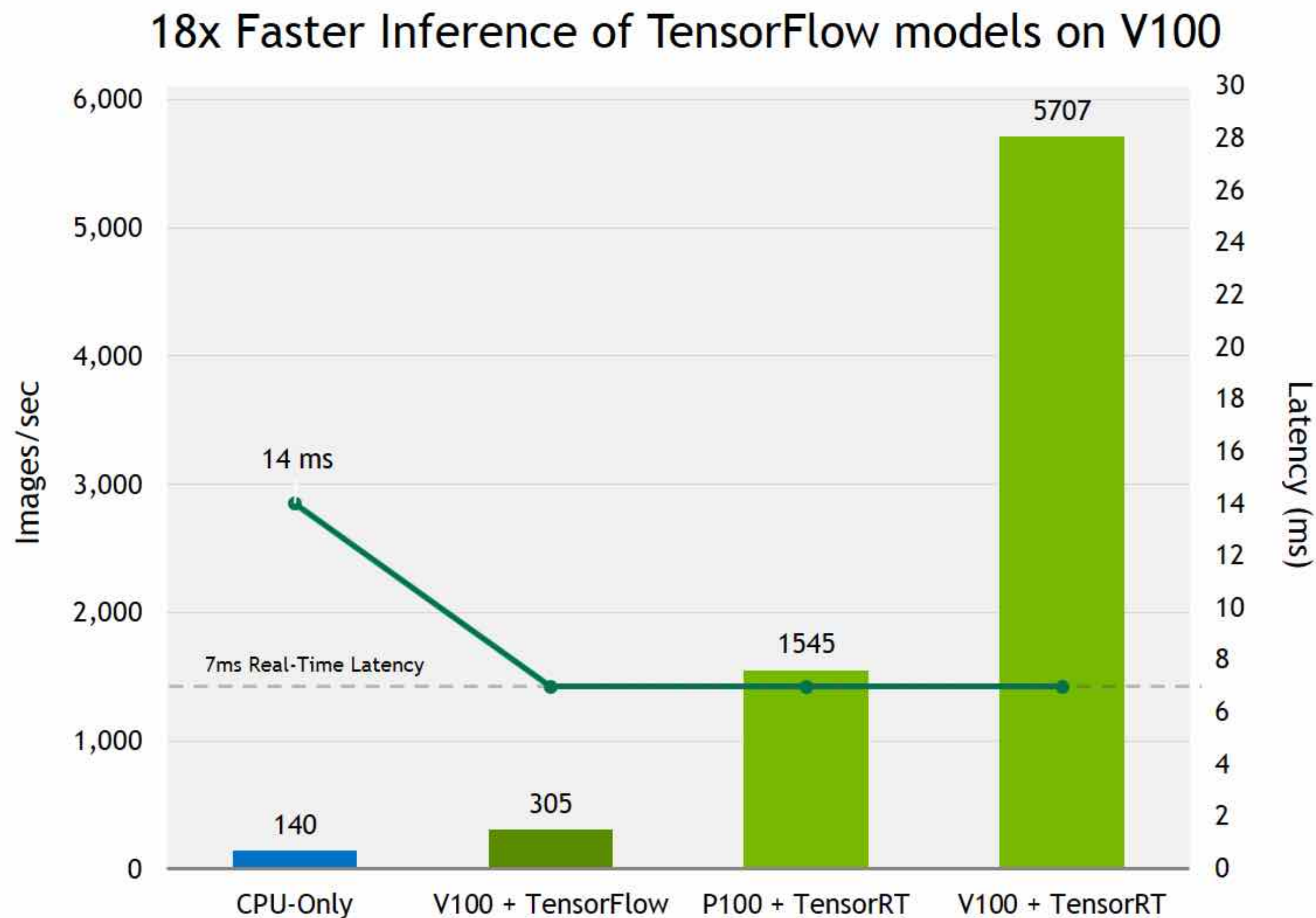


TABLE OF CONTENTES

效率精度平衡的卷积网络

卷积网络的压缩

嵌入式GPU+CPU的加速

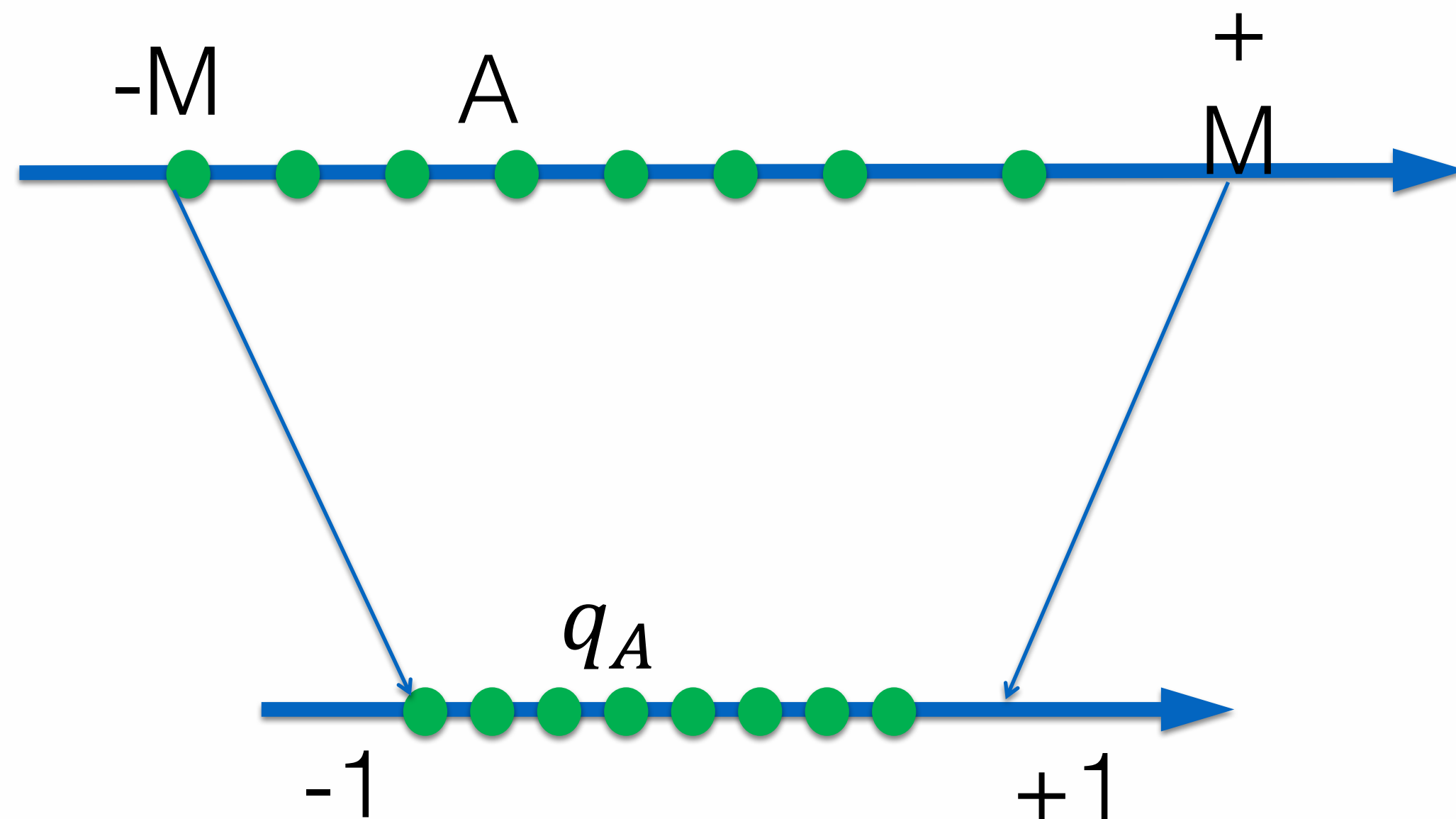
低成本FPGA的加速

FPGA定点化

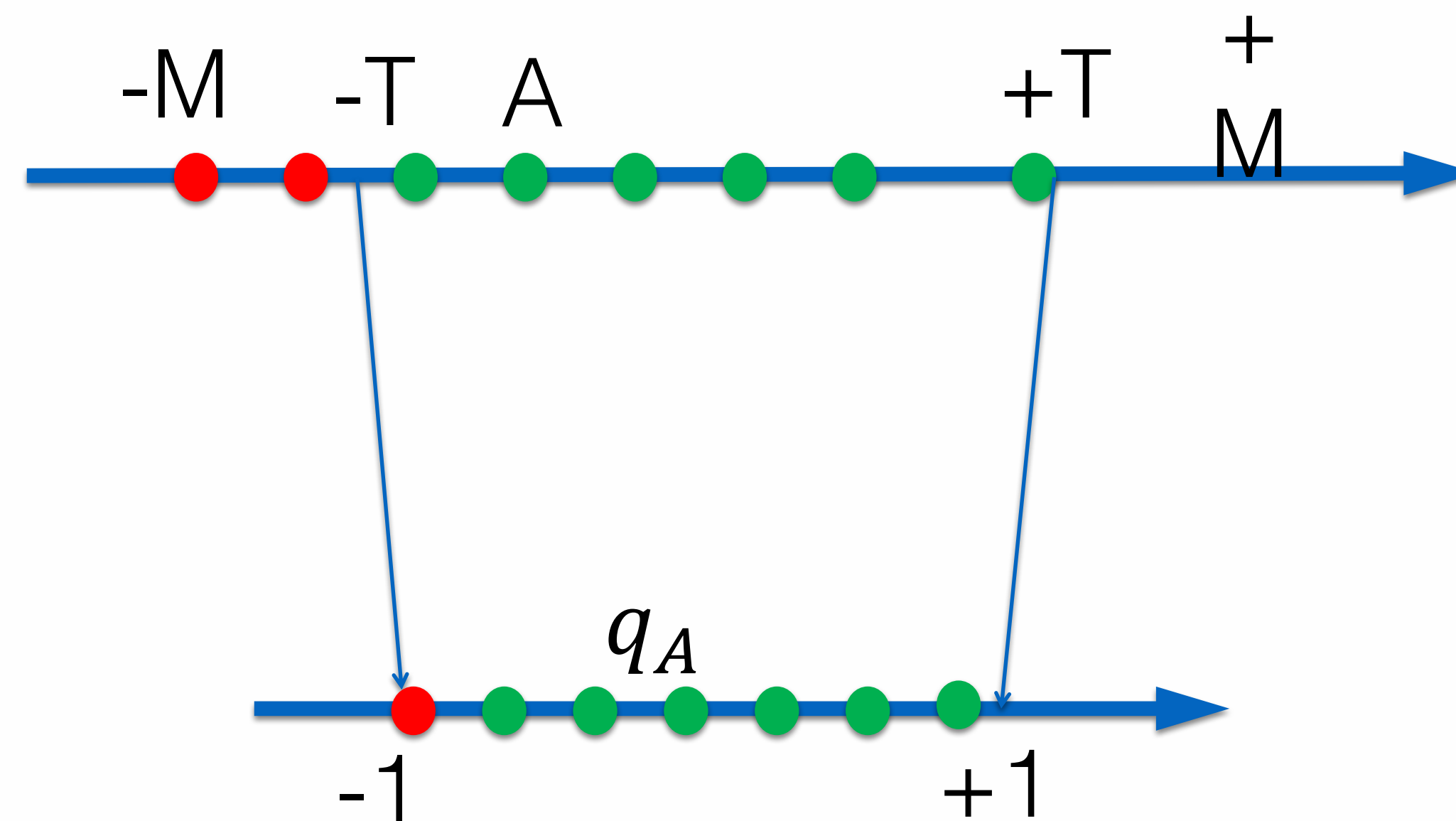
- 表示范围FP32 : $-3.4 \times 10^{38} \sim 3.4 \times 10^{38}$, INT8 : $-128 \sim 127$
- 定点小数 $A = (A_0)A_1 \cdots A_k \cdots A_n$ (A_0 为符号位, A_i 为0/1)代表的小数为
$$(-1)^{A_0} [A_1 * 2^{k-1} + A_2 * 2^{k-2} + \cdots + A_k * 2^0 + \cdots + A_n * 2^{k-n}]$$
- 定点小数表示范围在 $\pm 2^k (1 - 0.5^n)$ 之间, 精度 (最小单位) 为 2^{k-n}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

FP32 Value $A =$ FP32 scale factor $s_A * \text{INT8 Value } q_A$

$$\bullet \sum_j A_j B_j = s_A s_B \sum_j q_{Aj} q_{Bj}$$

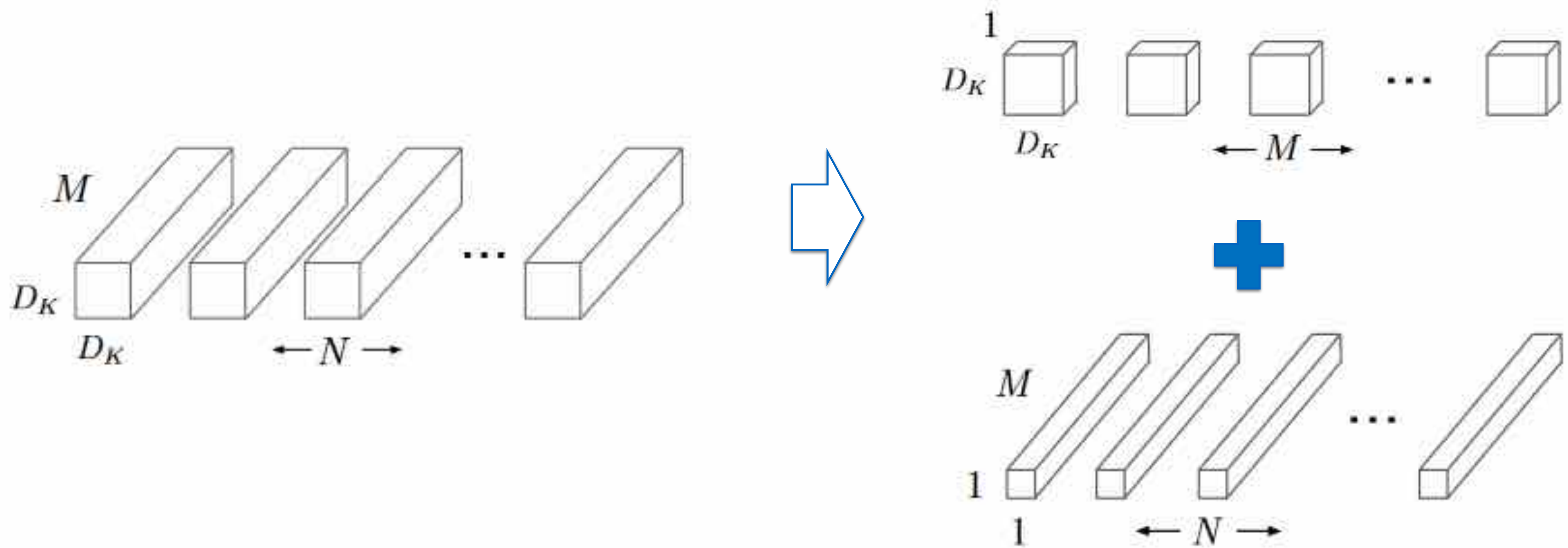


$M = \max |A| \leq 2^k$
 表示范围大，精度差



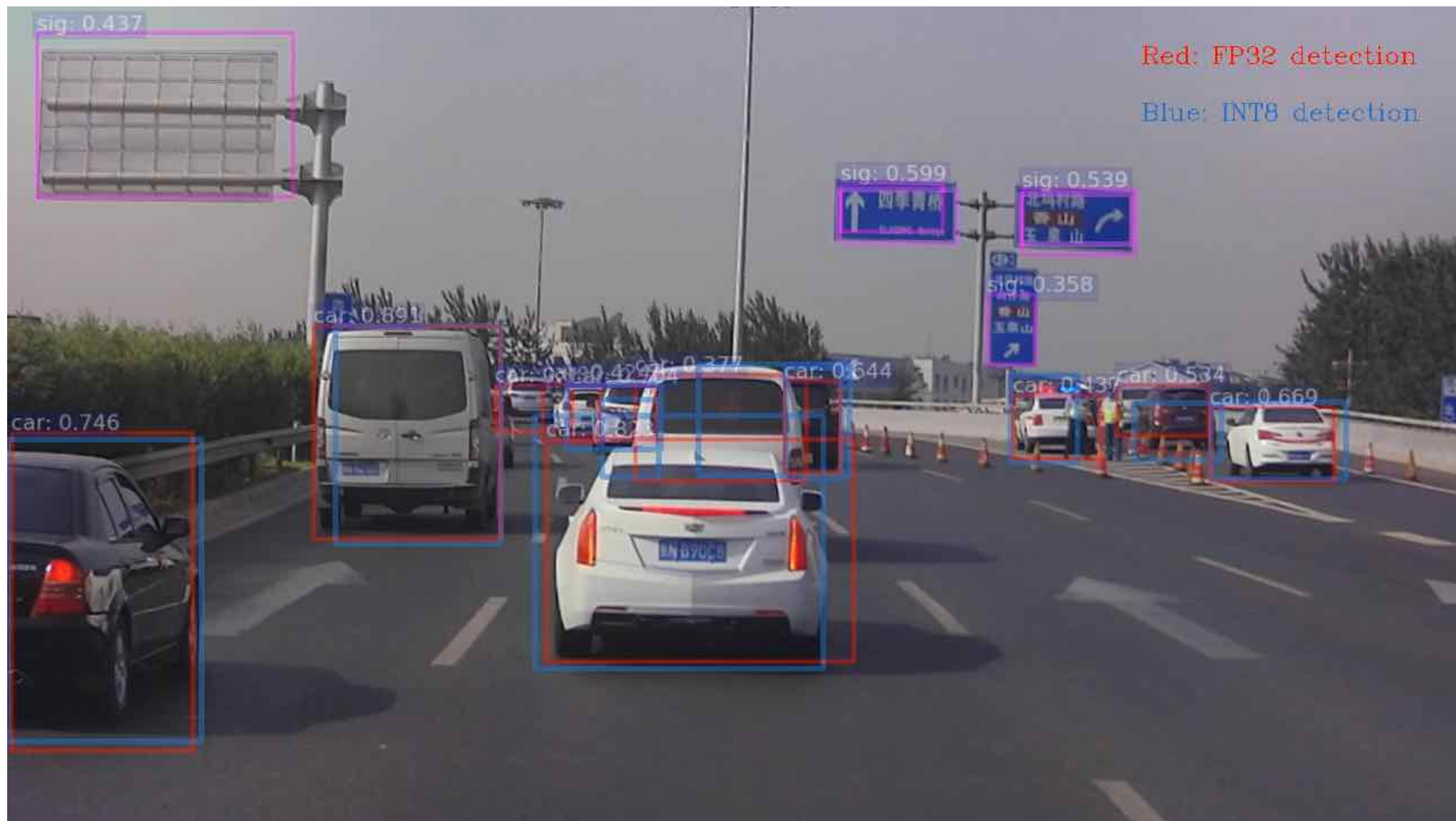
$M > T = 2^k$
 表示范围小，精度好

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加速比较差，但适合CPU和定制计算设备



FPGA+MobileNet物体检测

UISEE



UISEE Visual Perception demo video

Thanks!