# 端侧视频画质增强的软硬件设计方法研究


路 航

(副研究员、中科院青促会成员、计算所新百星)

计算机体系结构国家重点实验室
State Key Laboratory of Computer Architecture, ICT, CAS

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

# 移动端对视频画质增强的需求

**“5G市场，高清拍照和高清视频是用户的核心需求”**。毫无疑问，未来用户在拥有更好的网络下，手机的使用需求才会进一步提高，也将改变用户对社交网络上使用的图片、视频质量要求更高。这与OPPO一直主打的影像概念是一致的。




# 移动端对视频画质增强的算法要求

## 视频任务的需求

### 视频处理任务的需求

**实时性**

直播、会议等任务

**占用计算资源少**

降低延迟，提升用户体验

**通用性**

实际场景比较复杂，图像的退化因素多种多样

**适用移动设备**

移动终端的普及

# 移动端视频画质增强


LA 12 · CELTIC SUBMARÍ · CEU MANU TRIGUEROS · CHEQUIA GROGA · DON ESTEB
PAMESA
TORRECID
Budweiser
Budweiser
VBET
分辨率：720p
画质增强后：2k

# 移动端视频画质增强

# 端-边侧部署AI模型的痛点问题

**训练结果：**

| 指标 | 量化前 | Int8量化后 | Uint8量化后<br>(参数未融合) |
|---|---|---|---|
| 参数量 | 116 KB | 116 KB | 458 KB |
| 计算量 | 6.345 G | 6.345 G | 106 G |
| PSNR | 31.03 db | 30.96 db | **28.93 db** |
| SSIM | 0.9188 | 0.9176 | **0.8522** |


**原视频**

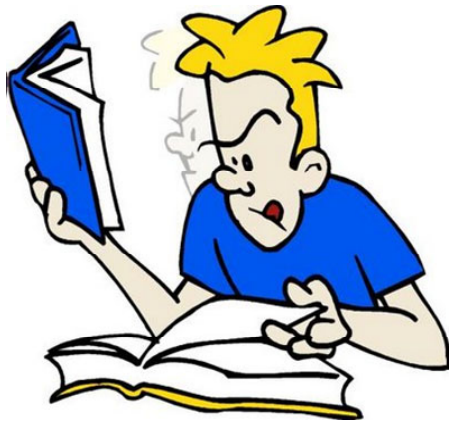

**int8**


**uint8**

**量化训练时间：8块V100，3天！**

# 利用比特稀疏性的通用深度学习加速方法

- 通用性不好
  - 支持模型种类单一，换场景就要换平台 ☹
  - 位宽范围小，大多(u)int8。fp32/16无法满足帧率要求 ☹

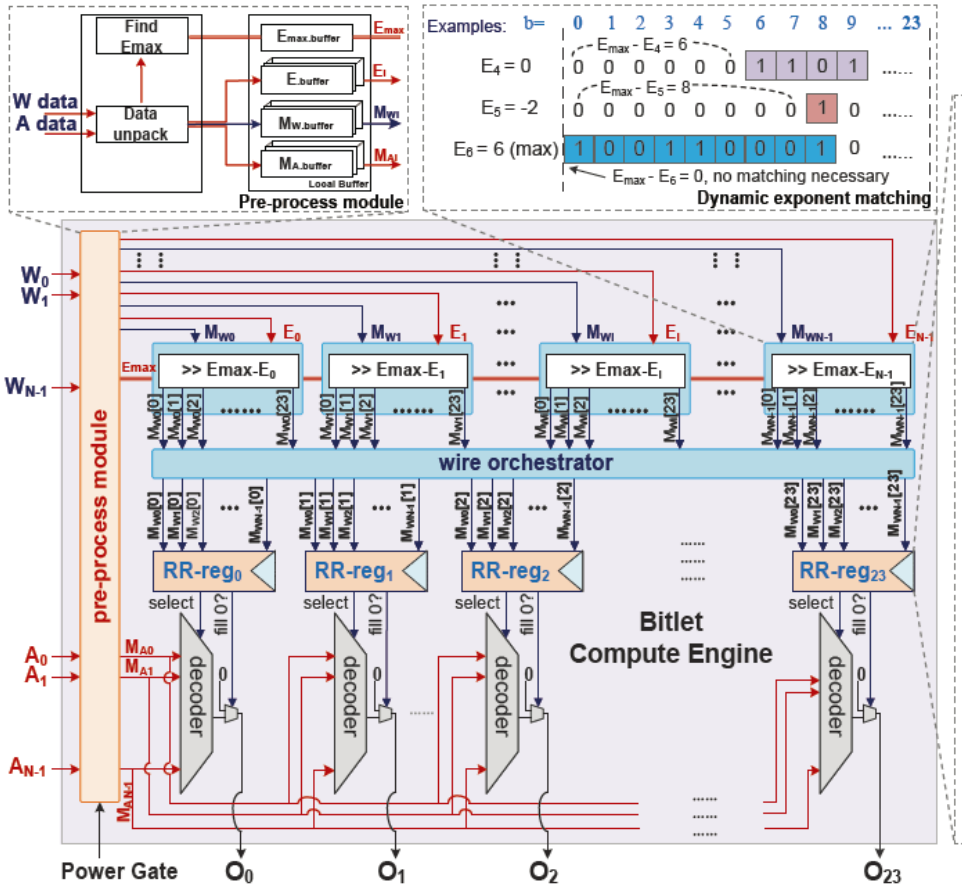


- Our solution
  - 灵活性
    - 支持数据位宽——定点数：1~24bit，浮点数：fp32/16，bfloat16
    - 在同一个计算引擎中实现混合精度计算
    - 自动识别并充分利用比特级稀疏性进行加速

    → High TOPs/W ☺
  - 通用性
    - 定点、浮点加速集成于同一计算引擎，无需设计额外硬件支持所有AI场景
  - 搭载硬件剪枝

    → High TFLOPS/W ☺

# 利用比特稀疏性的通用深度学习加速方法




**MICRO'21 | Oct 18-22nd, 体系结构领域顶会！清北中科院，天津大学与腾讯上榜！**

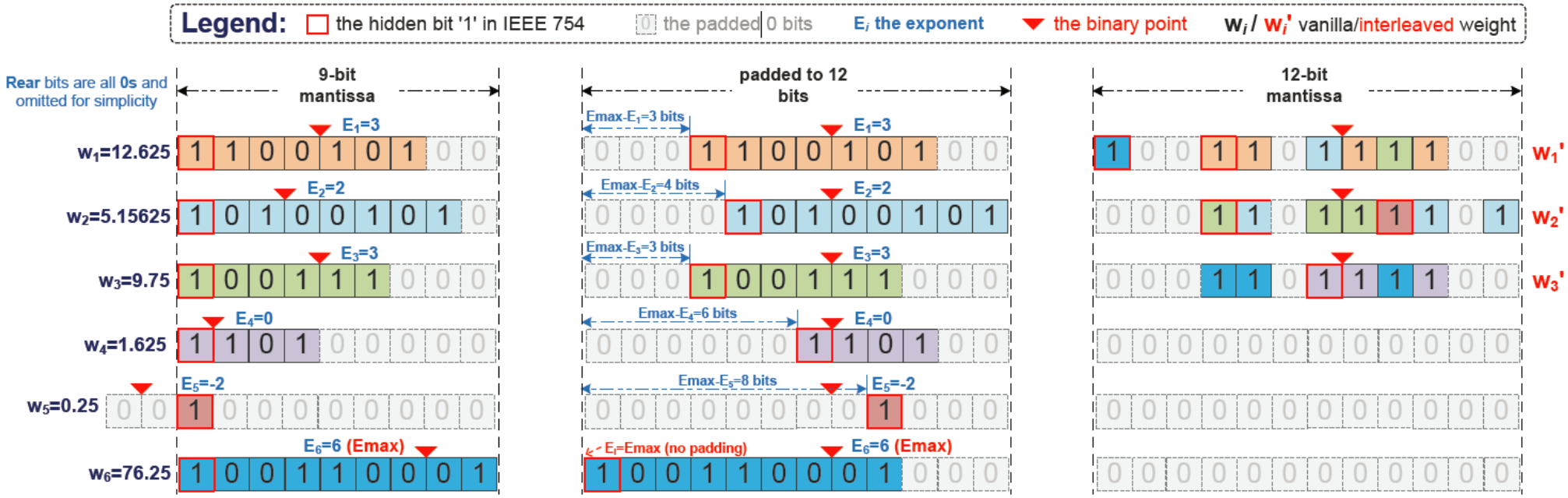
原创 CS Conferences CS Conferences 2021-10-13 10:30

收录于话题
#近期顶会
66个 >

动动大拇指 快快关注哦~

The International Symposium on Microarchitecture (MICRO)，国际微架构研讨会(MICRO)，是介绍和讨论微体系结构、编译器、硬件/软件接口以及高级计算和通信系统设计的主要论坛。MICRO是CCF A类会议，H5指数45，Impact Score高达6.76，在体系结构领域具有极高的评价。MICRO的目标是将微体系结构、编译器和系统领域的研究人员聚集在一起进行技术交流。

# 利用比特稀疏性的通用深度学习加速方法




(a) **Step 1:** preprocessing the floating-point weights

(b) **Step 2:** dynamic exponent matching

(c) **Step 3:** bit distillation

## 技术特点：

- MAC计算整体输出结果，无中间结果；利用bit中暴露的大量稀疏性加速

# 利用比特稀疏性的通用深度学习加速方法

- **峰值算力比较：**
  - **fp32：0.36TFLOPs/W；INT8：1.3TOPs/W @TSMC 28nm**

| | Accelerator ASICs | | | | | | GPUs | | |
|---|---|---|---|---|---|---|---|---|---|
| **Chip** | Eyeriss [14] | SCNN [32] | Stripes [22] | Laconic [36] | **Bitlet (Ours)** | | Titan V | Titan Xp | Tegra X2 |
| **PEs/Cores** | 168 | 64 | 4096 | 192 | **32** | | 5120 | 3840 | 256 |
| **Precision** | 16b | 16b | 1~16b | 1~16b | **fp32/16, 1~24b** | | fp32/16, 8b | fp32, 8b | fp32/16 |
| **Technology** | 65nm TSMC | 16nm TSMC | 65nm TSMC | 65nm TSMC | **28nm TSMC** | **65nm TSMC** | 12nm TSMC | 16nm TSMC | 16nm TSMC |
| **Freq. (MHz)** | 250 | 1000 | 980 | 1000 | **1000** | | 1455 | 1582 | 854 |
| **PEAK Performance (GOPs)** | 23.1 | 2000 | – | – | **204.8 (fp32)<br>372.35 (16b)<br>744.7 (8b)** | | 14900 (fp32)<br>29800 (fp16) | 12150 (fp32) | 750.1(fp32)<br>1330 (fp16) |
| **Power** | 278mW | – | – | – | **570mW(fp32)<br>432mW(16b)<br>366mW(8b)** | **1829mW(fp32)<br>1390mW(16b)<br>1199mW(8b)** | 250W | 250W | 15W |
| **PEAK Power Efficiency (GOPs/W)** | 83.09 | – | – | 441 (16b)<br>805 (8b) | **359.15 (fp32)<br>667.97(16b)<br>1335.93 (8b)** | **111.97 (fp32)<br>267.87 (16b)<br>621.10 (8b)** | 59.6(fp32)<br>119.2(fp16) | 48.6 (fp32) | 50.0 (fp32)<br>88.7(fp16) |
| **Area ($mm^2$)** | 12.25 | 7.9 | 122.1 | 1.59 | **1.54** | **5.80** | – | – | – |

# 利用比特稀疏性的通用深度学习加速方法

- **FPGA上，本IP和赛灵思浮点数IP计算时间比较（64个浮点MAC）：**

| 资源消耗 | 本IP | 赛灵思浮点乘法IP v12.0 | | | |
|---|---|---|---|---|---|
| LUT | **2558** | 35,773 | 34,083 | 33,270 | 33,419 |
| FF | **2040** | 8505 | 21,217 | 35,462 | 41,674 |
| 最大频率 | 300MHz | 100MHz | 200MHz | 300MHz | 400MHz |
| 延迟 | **40** | 6 | 24 | 48 | 73 |

# 利用比特稀疏性的通用深度学习加速方法

首次搭载硬件剪枝技术

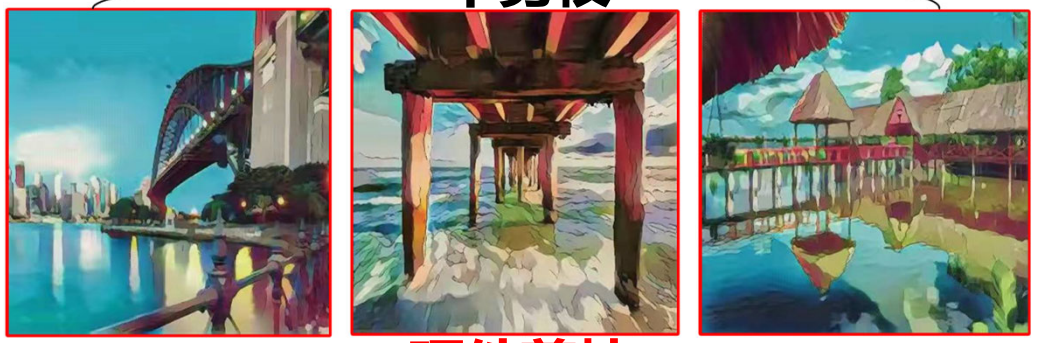
| 模型 | 指标 | 不剪枝 | 硬件剪枝 |
|---|---|---|---|
| **ResNet-50** | Top-1 (%) | 76.13 | 76.01 |
| **MobileNetV2** | Top-1 (%) | 71.88 | 71.29 |
| **YoloV3** | mAP 0.5:0.95 | 0.336 | **0.338** |
| **Multi-Pose** | mAP 0.5:0.95 | 0.59 | 0.57 |
| **lapSRN** | PSNR | 31.65 | 31.44 |
| | SSIM | 0.90 | 0.89 |
| **DCPDNet** | SSIM | 0.78 | **0.78** |
| **DenseNet-161** | Top-1 (%) | 77.14 | **77.14** |
| **FCOS** | mAP 0.5:0.95 | 0.38 | **0.38** |
| **CartoonGAN** | **下一页直接比较视觉效果** | | |
| **Transformer** | BLEU4 | 40.83 | **40.92** |
| **C3D** | Top-1 (%) | 97.31 | 97.27 |
| **D3DNet** | PSNR | 36.05 | 36.04 |
| | **SSIM** | **0.94** | **0.94** |

# 利用比特稀疏性的通用深度学习加速方法

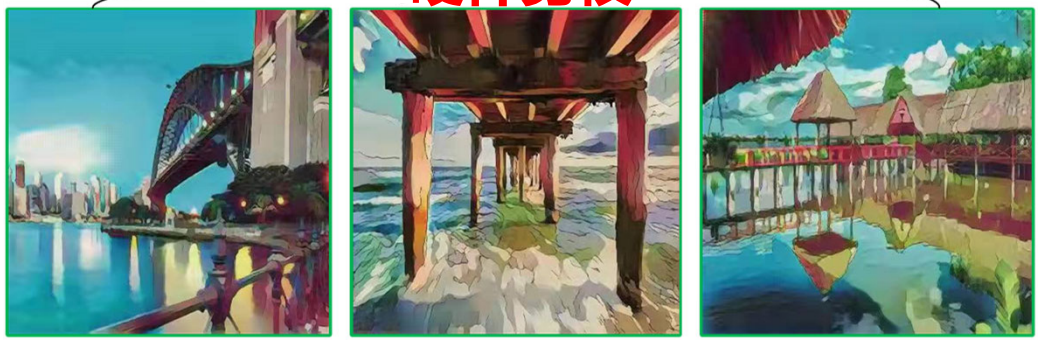
## 直接效果比较（CartoonGAN）：

Original Images w/o Cartoon Style Transfer

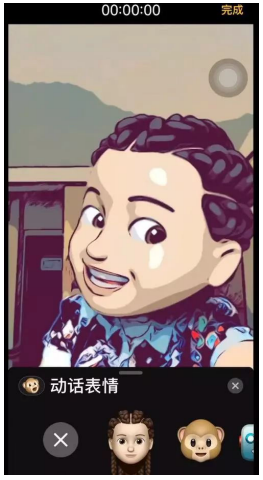
不剪枝

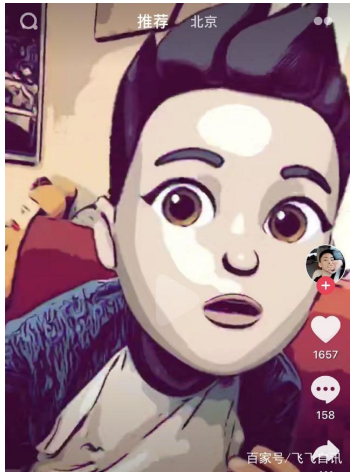


硬件剪枝



抖音爆火漫画脸教程，看看你的手机支持吗？



# 利用比特稀疏性的通用深度学习加速方法

- **直接效果比较（LapSRN）：**

Original Image
(195x195)

不剪枝

硬件剪枝

# 利用比特稀疏性的通用深度学习加速方法

- **硬件剪枝的速度提升（64个浮点MAC的cycle数）：**

| 模型 | 不剪枝 | 硬件剪枝 | 速度提升 |
|---|---|---|---|
| **ResNet-50** | 62.5 | 58.88 | **6%** |
| **MobileNetV2** | 72.5 | 63.19 | **13%** |
| **YoloV3** | 81.76 | 67.71 | **17%** |
| **Multi-Pose** | 74.29 | 66.64 | **10%** |
| **lapSRN** | 85.09 | 76.59 | **10%** |
| **DCPDNet** | 71.92 | 62.00 | **14%** |
| **DenseNet-161** | 56.5 | 55.06 | **3%** |
| **FCOS** | 77.58 | 68.62 | **12%** |
| **CartoonGAN** | 63.38 | 58.81 | **7%** |
| **Transformer** | 61.34 | 41.02 | **33%** |
| **C3D** | 64.41 | 56.31 | **13%** |
| **D3DNet** | 69.58 | 63.0 | **9%** |

# 利用比特稀疏性的通用深度学习加速方法

## 云-边-端通用的加速方法（即将推出IP）

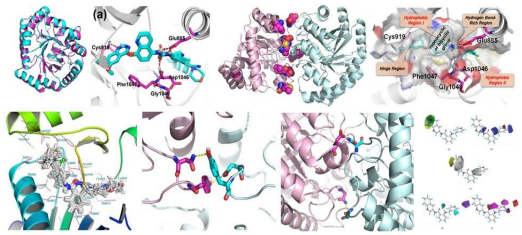
### Features：一个硬件加速IP，云-边-端多场景使用。

- 通用性（General Purpose）——平衡能效和通用性，多种精度支持集成于同一个加速引擎，高度灵活可配置，无额外硬件设计开销。

- 多精度支持（multi-precision support）——支持浮点数（fp32/16，bfloat16），定点数（1~24bit位宽），实现混合精度计算；使用本IP在浮点数模式下的推理时间与普通加速IP的int8推理时间相当；

- 硬件剪枝进一步提升推理性能——业界还没有带硬件剪枝技术的NPU！



# 感谢各位聆听！


路 航

(副研究员、中科院青促会成员、计算所新百星)



计算机体系结构国家重点实验室
State Key Laboratory of Computer Architecture, ICT, CAS

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES