

团体标准

T/CESA XXX-202X

人工智能芯片 计算机视觉训练用云侧深度学习芯片测试指标与测试方法

AI Chips-Computer Vision-Test metrics and test method of deep learning chips for
cloud side training

征求意见稿

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。
已授权的专利证明材料为专利证书复印件或扉页，已公开但尚未授权的专利申请证明材料为专利公开通知书复印件或扉页，未公开的专利申请的证明材料为专利申请号和申请日期。

202X-XX- XX 发布

202X-XX- XX 实施

中国电子工业标准化技术协会 发布



版权保护文件

版权所有归属于该标准的发布机构，除非有其他规定，否则未经许可，此发行物及其章节不得以其他形式或任何手段进行复制、再版或使用，包括电子版，影印件，或发布在互联网及内部网络等。使用许可可于发布机构获取。

目 次

前 言.....	IV
1 范围.....	1
2 规范性引用文件.....	1
3 术语和定义.....	1
4 缩略语.....	1
5 测试说明.....	1
5.1 测试环境及流程.....	1
5.2 测试对象.....	1
5.3 测试内容.....	2
6 测试指标.....	2
6.1 基本技术规格.....	2
6.2 功能.....	3
6.3 性能.....	3
6.4 软件生态.....	4
7 测试方法.....	5
7.1 基本技术规格.....	5
7.2 功能.....	5
7.3 性能.....	6
7.4 软件生态.....	7
附 录 A（规范性） 算子参数配置.....	9
A.1 算子性能评测配置参数.....	9
附 录 B（规范性） 算子及模型列表.....	12
B.1 算子列表.....	12
B.2 长尾算子列表.....	12
B.3 模型列表.....	13

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由上海商汤科技开发有限公司提出。

本文件由中国电子技术标准化研究院、中国电子工业标准化技术协会归口。

本文件起草单位： 。

本文件主要起草人： 。



人工智能芯片 计算机视觉训练用云侧深度学习芯片测试指标与测试方法

1 范围

本文件规定了计算机视觉领域面向云侧的深度学习训练芯片的基本技术规格、功能、性能、生态与开放性测试指标和测试方法。

本文件适用于芯片生产厂商、应用厂商及第三方机构对计算机视觉领域面向云侧的深度学习训练芯片进行测试与评估,也适用于计算机视觉领域深度学习训练芯片产品的采购、设计。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

T/CESA 1119—2020 人工智能芯片 面向云侧的深度学习芯片测试指标与测试方法

3 术语和定义

T/CESA 1119—2020界定的以及下列术语和定义适用于本文件。

3.1

计算机视觉 **computer vision**

一种具备获取、处理和解释视觉数据能力的功能单元。

[来源:ISO /IEC DIS 22989:2021, 3.1.11, 有修改]

4 缩略语

下列缩略语适用于本文件。

IPS: 每秒处理的图片数(Images Per Second)

API: 应用编程接口(Application Programming Interface)

5 测试说明

5.1 测试环境及流程

本文件的测试环境及测试流程应符合 T/CESA 1119—2020的相关要求。

5.2 测试对象

本文件的测试对象是含有计算机视觉推理用云侧深度学习芯片(卡/棒)的控制主机:指以芯片/卡/

棒形态进行使用的深度学习芯片,如GPU、FPGA以及ASIC等人工智能芯片(卡/棒),可通过PCIE、USB等接口与测试主机连接。

5.3 测试内容

计算机视觉训练芯片的测评指标,主要包括基本技术规格、功能、性能、生态与开放性等部门,在依据本文件进行测试的过程中:

- a) 涉及功能、性能等相关指标将通过第三方测试工具进行评测;
- b) 生态与开放性部分的指标将采信被测对象标称值及其他技术信息,作为先进性的参考。

6 测试指标

6.1 基本技术规格

基本技术规格从算力、内存、通信以及能效比四个方面进行评测,每个子指标的评分计算方式为该子指标测试值与对应基准值的比值,计算公式为:

$$\text{子指标评分} = \frac{\text{子指标测试值}}{\text{子指标基准值}} \dots\dots\dots (1)$$

a) 算力

表1 算力相关指标与参考值

序号	指标名称	指标内容	参考值
1	FP16算力(TFLOPS)	16 bit (1bit sign + 5 bit exponent + 10 bit fraction) 浮点数据的计算能力	
2	FP32算力(TFLOPS)	32 bit (1 bit sign + 8 bit exponent + 23 bit fraction) 浮点数据的计算能力	
3	INT8算力(TOPS)	8 bit整型数据的计算能力	
4	INT16算力(TOPS)	16 bit整型数据的计算能力	
5	BF16算力(TFLOPS)	16 bit (1 bit sign + 8 bit exponent + 7 bit fraction) 浮点数据的计算能力	
6	TF32算力(TFLOPS)	19 bit (1 bit sign + 8 bit exponent + 10 bit fraction) 浮点数据的计算能力	

b) 内存规格

内存是训练芯片的片下存储器(显存),而不是指主机存储器。

表2 内存相关指标与参考值

序号	指标名称	指标内容	参考值
1	容量(GB)	内存容量用字节数进行标称	
2	带宽(GB/s)	芯片的运算单元访问片下存储器的带宽	

c) 通信带宽

表3 通信带宽相关指标与参考值

序号	指标名称	指标内容	参考值
1	主机-设备带宽(GB/s)	训练芯片与主机之间的通信带宽	

2	节点内卡间带宽 (GB/s)	一个计算节点内，两个芯片之间的通信带宽	
---	-------------------	---------------------	--

d) 能效比

表4 能效比相关指标与参考值

序号	指标名称	指标内容	参考值
1	最高浮点算力能效比 (TFLOPS/W)	芯片最高浮点算力与芯片标称功耗TDP的比值	
2	最高整型算力能效比 (TOPS/W)	芯片最高整型算力与芯片标称功耗TDP的比值	

6.2 功能

6.2.1 算子支持程度

训练芯片对算子的支持程度，使用算子支持率进行衡量。算子支持率计算公式如下：

$$\text{算子支持率} = \sum \alpha_i \cdot [\text{算子}_i \text{是否支持}] \dots\dots\dots (2)$$

其中，被测试算子从算子列表（附录 B.1）中获取，每个算子的权重系数 α_i 使用统计方法获得。

6.2.2 模型支持程度

对目前常见深度学习应用领域（例如图像分类、分割、目标检测、NLP、推荐等）中典型模型的支持程度。模型支持率的计算公式如下：

$$\text{模型支持率} = \sum \beta_i \cdot [\text{模型}_i \text{是否支持}] \dots\dots\dots (3)$$

其中，被测试模型以及相应的权重系数 β_i 从模型列表（附录B.2）中获取。

6.2.3 卡间、多机高速通信的功能支持

卡间和多机高速通信分别指“节点内点对点通信”和“跨节点点对点通信”，指标内容如下表。

表5 卡间和多机高速通信功能支持

序号	指标名称	指标内容
1	节点内点对点通信	节点内用于卡间直接通信，CPU-Offload
2	跨节点点对点通信	跨节点卡间直接通信，CPU-Offload

6.2.4 训练性能的数制能力

a) 新型数制

芯片中的运算单元支持TF32、BF16等新型数制。

b) 稀疏计算

芯片在不降低模型训练精度的情况下，支持稀疏矩阵的运算以提高训练性能。

6.3 性能

6.3.1 算子计算性能

算子性能指某一特定输入配置情况下在芯片上的运算时间，不包含数据在主机内存和芯片存储器之间的传输时间。主要考虑 GEMM、Conv2d 和长尾算子在不同输入参数条件下在单芯片上的计算性能，其中长尾算子从被测试算子从算子列表（附录B.1）中获取。单项配置下算子的性能评分如公式（4）所示：

$$\text{算子单项配置性能评分} = \frac{\text{基准耗时}}{\text{计算耗时}} \dots\dots\dots (4)$$

算子性能评分为GEMM、Conv2d和长尾算子测试项的加权平均，其中权重系数依次为{0.3、0.4、0.3}。

6.3.2 通信性能

指算子在单节点多芯片、多节点多芯片条件下的性能表现，包括通信速率和时延。通信速率指消息体字节数与消息体从一个通信节点发出到达另外一个通信节点所需时间的比值（单位：GB/s）。时延指通信节点发送消息体时从开始发送至发送结束所需的时间（单位：ms）。

6.3.3 模型训练性能

主流深度学习模型在不同配置（单卡、多卡）情形下的训练性能。模型训练性能用 IPS 衡量，是指训练过程中每秒钟能处理的图片数，其计算公式如下：

$$\text{IPS} = \frac{\text{batchsize} \cdot \text{卡数}}{\text{每轮迭代时间}} \dots\dots\dots (5)$$

模型性能评分为所有测试模型评分的加权平均。

$$\text{模型性能评分} = \sum_{i=1}^N \gamma_i \cdot \text{模型}_i \text{的性能评分} \dots\dots\dots (6)$$

其中，模型i的性能评分为：

$$\text{模型}_i \text{的性能评分} = \frac{1}{M} \cdot \sum_{j=1}^M \frac{\text{测试性能}_i^j}{\text{基准性能}_i^j}, M = 3 \dots\dots\dots (7)$$

式中：

M——单机1卡、单机4卡和单机8卡3种测试配置。

模型测试参数配置以及训练数据集详见附录B.2。

6.4 软件生态

6.4.1 生态

生态指芯片的基本软件栈，并考虑芯片在公开市场的部署规模。评测内容主要包含如下几点：

- a) 支持用户对芯片进行软件开发的运行时库、编译工具链和调试调优工具。

表6 基本软件栈支持度指标

序号	指标名称	指标内容	必要/可选指标
1	驱动支持	是否包含驱动以及提供驱动API用于软件开发	必要

2	运行时库	是否包含运行时库	必要
3	编译工具链	是否提供编译工具链对用户程序进行编译	必要
4	调试工具	是否提供调试工具对芯片的代码进行调试排错	必要
5	调优工具	是否提供调优工具对芯片的代码实现进行性能分析、调优	必要

- b) 芯片的高性能计算库。主要包括计算库的数量、计算库提供的算子/函数的数量、提供计算库的性能三个方面。

表7 高性能计算库指标

序号	指标名称	指标内容	必要/可选指标
1	第1级计算库	是否包含深度神经网络（DNN）库、线性代数库等，并使用典型算子去测试其计算性能对芯片算力的利用率	必要
2	第2级计算库	是否包含其他计算库，例如随机数生成库等	必要

- c) 高性能通信库支持程度。覆盖主机-芯片之间、节点内芯片间以及跨节点芯片间三种场景的高性能通信库，以及是否支持常见的通信原语，如 All-Reduce、Reduce-Scatter、Broadcast 等。

6.4.2 开放性

开放性评测中的指标包含开放的芯片指令集或虚拟指令集、开放的设备代码编译器等，详见表8。

表8 开放性指标

序号	指标名称	指标内容	必要/可选指标
1	编程模型	编程模型、线程模型、存储层级设计是否和业界主流异构计算模型保持兼容	必要
2	编程接口	编程接口（如设备管理、流的使用与管理、同步机制等）是否与主流异构计算的编程接口保持兼容	必要

7 测试方法

7.1 基本技术规格

算力、内存、通信等子指标的测试均采用厂商提供的标称值。

7.2 功能

7.2.1 测试目标

测试训练芯片以及其软件栈是否支持附录B.1（算子列表）和附录B.2（模型列表）所列的算子与模型。

7.2.2 测试准备

功能测试需要被测方提供以下内容：

- 应提供处于最佳工作环境、厂商标配的主机配置、厂商标配的训练芯片产品形态；
- 应提供训练芯片软件栈的相关技术文档。

7.2.3 测试要求

待测算子和待测试模型应满足以下要求：

- d) 支持至少一种数值精度（FP32、FB16、TF32、BF16、INT8 和 INT16）的实现；
- e) 训练芯片执行该算子的输出结果应与 ONNX Runtime CPU（v1.10.0, Intel i7-8700@3.2GHz）的输出结果进行比较，两者误差在可接受范围内；
- f) 若输出参数是张量，对张量中每一个元素与标准输出结果中对应元素进行比较；
- g) 测试模型中至少 95%的算子在训练芯片上执行，且关键算子（包括卷积、矩阵乘、归一化、激活函数、池化）在训练芯片上执行；
- h) 模型测试使用超参与附录 B.2 保持一致，在满足模型测试精度要求的前提下，不限定训练使用的数值精度。

7.2.4 算子功能测试流程

表11 算子功能测试流程

序号	步骤	步骤描述
1	参数配置	给定输入参数，使用该算子在ONNX Runtime CPU实现进行计算，获得在该输入配置下的标准输出结果。
2	算子执行	使用上述输入参数，在训练芯片上执行该算子，获得相应的测试输出结果。
3	精度对比	将测试输出结果与标准输出结果进行对比，计算相对误差和绝对误差。

7.2.5 模型功能测试流程

表12 模型功能测试流程

序号	步骤	步骤描述
1	给定参数	给定模型测试数据集、超参配置、要求训练轮数以及测试精度要求。
2	模型运行	在以训练芯片为基础的计算系统上，使用指定数据集和超参进行训练。
3	精度对比	当训练轮数达到训练要求的轮数时，测试模型在指定测试数据集上的精度。

7.3 性能

7.3.1 测试目标

测试训练芯片以及其软件栈在附录B.1 算子列表和B.2 模型列表下的训练性能。

7.3.2 测试准备

性能测试需要被测方提供以下内容：

- a) 应提供处于最佳工作环境、厂商标配的主机配置、厂商标配的训练芯片产品形态；
- b) 应提供训练芯片软件栈的相关技术文档。

7.3.3 测试要求

测试训练芯片性能有以下测试要求：

- i) 应在在不同通信负载和通信节点条件下，测试 All-Reduce 算子的算法带宽（GB/s）和通信延迟（ms）；

j) 应在不同的配置下（单机1卡、单机4卡、单机8卡等），测试模型训练性能。

7.3.4 算子性能测试流程

表13 算子性能测试流程

序号	步骤	步骤描述
1	参数配置	准备输入数据，并将算子执行所需的所有输入数据传输至训练芯片存储器。
2	热身轮	在芯片上执行算子M (M<10) 次，作为性能测试的热身轮。
3	耗时测试	将算子在芯片上连续运行特定次数 N (N 介于 1000 和 100000 之间，测试人员在测试过程中根据实际情况指定)，取运算时间的均值；
4	精度测试	算子在某一特定输入配置下的计算时间与相应的基准时间的比值即为该输入参数配置下的性能评分。数值精度可取泛单精度 (FP32、TF32 等) 和泛半精度 (FP16、BF16 等)，基准性能也有两种精度的基准值，被测芯片的某个算子的评分系数选取两种数制精度下的最高值。
5	结果确认	该测试条件下的算子必须确保精度满足要求，评测要求参考第 7.2.1 章节。

7.3.5 模型性能测试流程

表14 模型性能测试流程

序号	步骤	步骤描述
1	参数配置	准备模型训练所需的参数、数据集，训练过程不能对设定参数进行修改。
2	热身轮	启动模型训练，执行M (M<3) 轮 (epoch) 训练作为热身轮。
3	测试执行	至少执行一个完整的训练轮 (epoch)，根据第6.3.3章节中IPS定义计算模型的训练性能。

7.4 软件生态

7.4.1 软件生态

测试芯片应支持必要的基本软件栈、高性能计算库、高性能通信库以及产品部署规模。

a) 基本软件栈

表14 基本软件栈测试方法

序号	指标名称	CUDA 对应	是否支持
1	驱动支持	cuda driver	
2	运行时库	cuda runtime	
3	编译工具链	nvcc	
4	调试工具	cuda-gdb	
5	调优工具	nvprof	

b) 高性能计算库

表15 高性能计算库测试方法

序号	指标名称	CUDA 对应	是否支持
1	第 1 级计算库	cuda, cublas	

2	第2级计算库	cusparse、curand	
---	--------	-----------------	--

c) 高性能通信库

通信库应支持常见的通信原语如All-Reduce、Reduce-Scatter、Broadcast等，CUDA对应的高性能通信库为NCCL。

7.4.2 开放性

训练芯片应考虑开放性相关指标：

表16 开放性测试方法

序号	子指标评测内容	是否支持
1	芯片指令集或虚拟指令集的开放程度	
2	是否开放设备代码编译器（或部分组件）用于极致性能调优	
3	编程接口和编程模型是否与主流异构计算生态兼容或可类比	



附 录 A
(规范性)
算子参数配置

A.1 算子性能评测配置参数

算子性能评测中所有测试算子以及相应的输入配置参数列如以下：

a) GEMM

GEMM算子的定义请参见ONNX-Operator-Gemm, 测试中参数 M, N, K 的取值如下表所示。参数 ($transA, transB$) 分别取 (N, N)、(N, T)、(T, N) 和 (T, T), 参数 C 为大小为 (M, M) 且值随机生成的矩阵, 参数 α, β 取默认值。综合上述参数配置项, 最终测试配置项数为 $224 = 56 \times 4$ 。

表A.1 GEMM测试输入参数配置

序号	M	N	K	序号	M	N	K
1	8	16	32	29	64	16	4096
2	8	128	32	30	64	128	4096
3	8	1024	32	31	64	1024	4096
4	8	7680	32	32	64	7680	4096
5	8	16	256	33	2048	16	32
6	8	128	256	34	2048	128	32
7	8	1024	256	35	2048	1024	32
8	8	7680	256	36	2048	7680	32
9	8	16	1536	37	2048	16	256
10	8	128	1536	38	2048	128	256
11	8	1024	1536	39	2048	1024	256
12	8	7680	1536	40	2048	7680	256
13	8	16	4096	41	2048	16	1536
14	8	128	4096	42	2048	128	1536
15	8	1024	4096	43	2048	1024	1536
16	8	7680	4096	44	2048	7680	1536
17	64	16	32	45	2048	16	4096
18	64	128	32	46	2048	128	4096
19	64	1024	32	47	2048	1024	4096
20	64	7680	32	48	2048	7680	4096
21	64	16	256	49	1760	6574	1760
22	64	128	256	50	35	8467	2048
23	64	1024	256	51	7680	16	2560
24	64	7680	256	52	6144	32	2816
25	64	16	1536	53	512	16	1024
26	64	128	1536	54	3072	128	512
27	64	1024	1536	55	256	1024	4096
28	64	7680	1536	56	512	32	512

b) Conv2d

表A.2 Conv2d测试输入参数配置

序号	W	H	C	N	K	S	R	pad_w	pad_h	s_h	s_v
1	224	224	3	8	64	3	3	1	1	1	1
2	112	112	64	8	128	3	3	1	1	1	1
3	56	56	128	8	256	3	3	1	1	1	1
4	28	28	256	8	512	3	3	1	1	1	1
5	14	14	512	8	512	3	3	1	1	1	1
6	7	7	512	8	512	3	3	1	1	1	1
7	224	224	3	32	64	3	3	1	1	1	1
8	112	112	64	32	128	3	3	1	1	1	1
9	56	56	128	32	256	3	3	1	1	1	1
10	28	28	256	32	512	3	3	1	1	1	1
11	14	14	512	32	512	3	3	1	1	1	1
12	7	7	512	32	512	3	3	1	1	1	1
13	224	224	3	256	64	3	3	1	1	1	1
14	112	112	64	256	128	3	3	1	1	1	1
15	56	56	128	256	256	3	3	1	1	1	1
16	28	28	256	256	512	3	3	1	1	1	1
17	14	14	512	256	512	3	3	1	1	1	1
18	7	7	512	256	512	3	3	1	1	1	1
19	224	224	3	32	64	7	7	3	3	2	2
20	28	28	192	32	32	5	5	2	2	1	1
21	28	28	192	32	64	1	1	0	0	1	1
22	14	14	512	32	48	5	5	2	2	1	1
23	14	14	512	32	192	1	1	0	0	1	1
24	7	7	832	32	256	1	1	0	0	1	1
25	7	7	832	32	128	5	5	2	2	1	1
26	224	224	3	512	64	7	7	3	3	2	2
27	28	28	192	512	32	5	5	2	2	1	1
28	28	28	192	512	64	1	1	0	0	1	1
29	14	14	512	512	48	5	5	2	2	1	1
30	14	14	512	512	192	1	1	0	0	1	1
31	7	7	832	512	256	1	1	0	0	1	1
32	7	7	832	512	128	5	5	2	2	1	1
33	480	48	1	16	16	3	3	1	1	1	1
34	240	24	16	16	32	3	3	1	1	1	1
35	120	12	32	16	64	3	3	1	1	1	1
36	60	6	64	16	128	3	3	1	1	1	1
37	108	108	3	8	64	3	3	1	1	2	2
38	54	54	64	8	64	3	3	1	1	1	1
39	27	27	128	8	128	3	3	1	1	1	1

40	14	14	128	8	256	3	3	1	1	1	1
41	7	7	256	8	512	3	3	1	1	1	1
42	56	56	64	16	64	3	3	1	1	1	1
43	56	56	64	16	256	1	1	0	0	2	2
44	28	28	128	16	128	3	3	1	1	1	1
45	28	28	128	16	512	1	1	0	0	2	2
46	14	14	256	16	256	3	3	1	1	1	1
47	14	14	256	16	1024	1	1	0	0	2	2
48	7	7	512	16	512	1	1	0	0	1	1
49	7	7	2048	16	512	1	1	3	3	2	2
50	56	56	64	512	64	3	3	1	1	1	1
51	56	56	64	512	256	1	1	0	0	2	2
52	28	28	128	512	128	3	3	1	1	1	1
53	28	28	128	512	512	1	1	0	0	2	2
54	14	14	256	512	256	3	3	1	1	1	1
55	14	14	256	512	1024	1	1	0	0	2	2
56	7	7	512	512	512	1	1	0	0	1	1
57	7	7	2048	512	512	1	1	3	3	2	2
58	112	112	64	8	64	1	1	0	0	1	1
59	56	56	64	8	256	1	1	0	0	1	1
60	112	112	64	128	64	1	1	0	0	1	1
61	56	56	64	128	256	1	1	0	0	1	1
62	112	112	64	512	64	1	1	0	0	1	1
63	56	56	64	512	256	1	1	0	0	1	1

附 录 B
(规范性)
算子及模型列表

B.1 算子列表

表B.1 算子列表

序号	算子列表
1	conv1d, conv2d, conv3d, batch_norm, relu, max_pool1d, max_pool2d, max_pool3d, conv_transpose1d, conv_transpose2d, conv_transpose3d, softmax, softmin, cross_entropy, binary_cross_entropy, dropout, select, randperm, mm, bmm, matmul, max, min, mean, add, sub, sum, div, mul, eq, gt, topk, stack, cat, split, sort, fill, arange, reshape, scatter, nonzero, layer_norm, interpolate, sigmoid, avg_pool1d, avg_pool2d, avg_pool3d, flatten, unsqueeze, squeeze, SGD, sin, cos, sinh, cosh, log, log2, exp, exp2, sqrt, fmod, sign, pow, neg, abs, floor, index_select, masked_select, permute, where, clamp, repeat, transpose, leaky_relu, prelu, log_softmax, instance_norm, Adam, Nms, RoiAlign, SyncBatchNorm, GlobalMaxPool, GlobalAveragePool, adaptive_avg_pool1d, adaptive_avg_pool2d, adaptive_avg_pool3d, adaptive_max_pool1d, adaptive_max_pool2d, adaptive_max_pool3d, ...

B.2 长尾算子列表

表B.2 长尾算子列表

序号	算子	序号	算子
1	bbox2delta	21	Fcos_matcher
2	bbox_overlaps	22	Index2d
3	Delta2bbox	23	Intersect
4	Compute_locations	24	Jaccard
5	Batched_nms	25	Legacy_bbox2delta
6	Bbox2roi	26	Margin_loss
7	Bbox2offset	27	Mask_predictor
8	L2_loss	28	Masks_to_boxes
9	Aeloss	29	Offset2bbox
10	Bmn_loss	30	Partialconv2d
11	Box_area	31	Shift
12	Box_iou	32	Random_sampler
13	Boxes_for_nms	33	Sanitize_coordinates
14	Bucket2bbox	34	Tblr2bbox
15	Center_size	35	Valid_flags
16	Centernet_keypoint	36	Position_embedding_sine
17	Crop	37	Position_embedding_learned

18	Edge_smoothloss		38	Msms_clsfc
19	Focal_loss		39	Maxiou_matcher_match
20	Gaussian_focal_loss		40	Map_roi_levels

B.3 模型列表

表B.3 模型列表

类别	模型	数据集	权重
分类	Resnet50_v1.5	ImageNet ILSVRC-2012	
	Inception_v3	ImageNet ILSVRC-2012	
	VGG16	ImageNet ILSVRC-2012	
	SE-Resnet50	ImageNet ILSVRC-2012	
	MobileNet_v2	ImageNet ILSVRC-2012	
	ShuffleNet_v2	ImageNet ILSVRC-2012	
	DenseNet121	ImageNet ILSVRC-2012	
	Swin Transformer	IN1K	
检测	Faster-RCNN-R50	COC02017	
	Mask-RCNN-R50	COC02017	
	Cascade-RCNN-R50	COC02017	
	Retinanet	COC02017	
	Yolo_v3	COC02017	
	FCOS-R50	COC02017	
	SSD300	COC02017	
	CenterNet-R18	COC02017	
	SOLO	COC02017	
	Swin Transformer-T-Mask-RCNN	COC02017	
分割	DeepLabV3-R50	VOC2012	
	UNet		
	FCN-R50	cityscapes	
	PSPNet-R50	cityscapes	
	APCNet-R50		

参 考 文 献

- [1] T/CESA 1120—2020 人工智能芯片 面向边缘侧的深度学习芯片测试指标与测试方法
[2] T/CESA 1121—2020 人工智能芯片 面向端侧的深度学习芯片测试指标与测试方法

