ICS 31.200
CCS L 56

团　体　标　准

T/CESA XXX-202X

人工智能芯片 计算机视觉推理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 inference

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 2
5.2 测试对象 2
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 5
6.4 软件生态 6
6.5 应用场景测试 6
7 测试方法 7
7.1 基本技术规格 7
7.2 功能 7
7.3 性能 8
7.4 软件生态 9
7.5 应用场景测试 9
附　录　A （规范性） 推理测试参考框架 11
附　录　B （规范性） 算子及模型列表 12
附　录　C （规范性） 软件生态评测指标 16

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 计算机视觉推理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推理芯片的基本规格、功能、性能、软件生态、应用场景等评测指标及测试方法。

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推理芯片进行测试与评估,也适用于计算机视觉领域深度学习推理芯片产品的采购、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

推理 inference

从给定的前提进行论证并得出结论。

注1:在人工智能领域中,一个前提是一个事实、一个规则、一个模型、一个特征或原始数据。

注2:术语“推理”既指过程也指结果。

[来源: ISO/IEC 2382:2015 2123828 有修改]

3.2

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 2382:2015 3.1.11 有修改]

4 缩略语

下列缩略语适用于本文件。

FPS: 帧率(Frame Per Second)

MAP: 平均精度均值(Mean average precision)

5 测试说明

5.1 测试环境

本文件的测试环境及测试流程均应符合 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》标准的相关要求。

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等深度学习芯片（卡/棒），可通过PCIE、USB等接口与测试主机连接。

5.3 测试内容

计算机视觉推理芯片的测评指标,主要包括基本技术规格、功能、性能、软件生态以及应用场景测试等部分，在依据本文件进行测试的过程中：

a）涉及功能、性能、应用场景等相关指标将通过第三方测试工具进行评测；

b）涉及基本技术规格和软件生态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

6.1.1 内存

内存指代推理芯片的设备端内存（对应GPU的显存），相关指标及参考值见表1。

表1 内存相关指标及参考值

维度	指标	参考值
内存	内存类型	/
	内存容量（GB）	40
	内存带宽（GB/s）	1555

6.1.2 互联

互联指代主机端CPU与设备端推理芯片的通信接口，相关指标及参考值见表2。

表2 互联相关指标及参考值

测试维度	测试指标	参考值
互联	接口类型	PCIe 4.0
	接口带宽（GB/s）	64

6.1.3 算力

算力指代推理芯片所能提供的所有计算能力，包含提供核心矩阵/卷积运算加速的计算单元以及提供相对通用计算的计算单元的算力（对应GPU的Tensor Core与Cuda Core），但对于不同计算单元提供的同样数据类型的算力需要进行区分，相关指标及参考值见表3。

表3　算力相关指标及参考值

测试维度	测试指标	参考值
算力	int8（TOPS）	624
	float16（TFLOPS）	312
	bfloat16（TFLOPS）	312
	float32（TFLOPS）	19.5
	tfloat32（TFLOPS）	156
	float64（TFLOPS）	19.5/9.7

6.1.4　视频图像编解码能力

视频图像编解码能力指代推理芯片所能提供的硬件编解码器的能力，包含视频/图像、编码/解码共四种任务的能力。相关指标及参考值见表4。

表4　视频图像编解码能力相关指标及参考值

测试维度	测试指标	参考值
编解码能力	编码格式、分辨率支持	视频（HEVC、H.264、VP9等）1080p、4K 图像（JPEG、PNG等）4K
	图像最大处理能力（编码格式、路数、分辨率、FPS）	解码：H.265 128路1080P 30FPS 编码：H.265 24路1080P 30FPS
	视频最大处理能力（编码格式、路数、分辨率、FPS）	解码：JPEG 4K 384FPS 编码：JPEG 4K 192FPS

6.1.5　功耗

功耗指代设备端的加速卡的整体功耗，相关指标及参考值见表5。

表5　功耗相关指标及参考值

测试维度	测试指标	参考值
功耗	待机功耗（W）	/
	满载功耗（W）	/
	最大热设计功耗TDP（W）	250

6.1.6　制程

制程指代推理芯片的芯片代工制程。相关指标及参考值见表6。

表6　制程相关指标及参考值

测试维度	测试指标	参考值
制程	芯片代工厂	/
	芯片制程（nm）	/

6.2　功能

6.2.1　算子功能

算子功能主要评测推理芯片对于计算机视觉任务常用的算子的支持度。待测算子采用 ONNX 格式表示，详细的待测算子列表参考附录 B.1，算子类别及权重设置见表 7。

注：算子功能章节的待测算子从附录B.3的待测模型池中挑选而来，并基于算子在待测模型池中所有模型的出现频率和组成模型必要性分为三个级别，并分别赋予权重。

表7　算子类别及权重设置

算子级别	权重	级别描述
第一级算子	0.5	核心算子。组成计算机视觉模型的最基本的算子集合，如Conv等。
第二级算子	0.3	高频算子。除核心算子外，在计算机视觉模型中使用频率较高的算子集合，用于组成某些常用的网络结构，比如Resize常用于FPN结构中。
第三级算子	0.2	领域专用算子。除上述算子外，在特定应用领域需要支持的算子，比如目标检测领域的后处理需要支持NonMaxSuppression。

采用推理芯片在上述三级算子的支持率作为评分依据，支持率如公式（1）所示：

$$\text{级别算子评分} = \frac{\text{当前级别支持的算子数量}}{\text{当前级别算子数量}} \quad \cdots\cdots (1)$$

对于推理芯片在上述三级算子的评分按照赋予的权重进行加权求和得到算子功能评分，如公式（2）下所示：

$$\text{算子功能评分} = \sum \text{级别算子权重} \times \text{级别算子评分} \quad \cdots\cdots (2)$$

6.2.2　模型功能

模型功能主要评测推理芯片在计算机视觉任务中典型模型的支持度。待测模型池基于典型计算机视觉任务的常用公开模型并兼顾实际使用情况，采用 ONNX 格式表示，并作为算子、模型测试的基准模型池。待测模型池分为图像分类、目标检测、语义分割三个领域，详细待测模型列表参考附录 B.3，权重设置见表 8。

表8　模型功能及权重设置

模型领域	权重
图像分类	0.5
目标检测	0.3
语义分割	0.2

采用推理芯片在上述三类模型的支持率作为评分依据，支持率如公式（3）所示：

$$\text{领域模型评分} = \frac{\text{当前领域支持的模型数量}}{\text{当前领域模型数量}} \quad \cdots\cdots (3)$$

对于推理芯片在上述三类模型的评分按照赋予的权重进行加权求和得到模型功能评分，如公式（4）所示：

$$\text{模型功能评分} = \sum \text{领域模型权重} \times \text{领域模型评分} \quad \cdots\cdots (4)$$

6.3 性能

6.3.1 算子性能

算子性能主要评测推理芯片在核心算子和相对耗时算子上的整体性能。待测算子采用 ONNX 格式表示，选择范围见附录 B.1，详细待测算子列表和对应的算子参数列表参考附录 B.2。算子性能及权重设置见表 9。

表9 算子性能及权重设置

算子类别	权重	类别说明
Conv	0.4	Conv是计算机视觉任务中耗时占比最高的算子，其性能决定了推理芯片在大多数模型中的整体性能。
Gemm	0.3	Gemm是实现计算机视觉任务加速的最基本的运算形式，其性能体现了推理芯片核心计算单元的综合性能。
其他算子	0.3	待测模型中包含较多逻辑判断且无法使用矩阵运算有效加速而称为模型性能瓶颈的算子，目前暂选择NonMaxSuppression、TopK、RoiAlign三个算子作为待测算子。

采用推理延迟与参考平台的参考推理延迟的比值作为评分依据，比值越低评分越高。同时考虑到不同推理芯片会支持不同的数据类型且会有不同的推理性能，因此需采用同样的数据类型，分别得到比值后取最大值，如公式（5）所示。

$$\text{特定算子参数配置下算子评分} = \mathbf{Max}\left(\frac{\text{参考平台推理延迟}}{\text{推理芯片推理延迟}}(\text{同数据类型})\right) \quad \cdots\cdots (5)$$

对于相同算子下多个算子的参数配置，需要体现推理芯片的综合性能，因此采用均等权重，如公式（6）所示。

$$\text{算子评分} = \frac{\sum \text{特定算子参数配置下算子评分}}{\text{算子的算子参数配置数量}} \quad \cdots\cdots (6)$$

推理芯片在上述三类算子的评分按照赋予的权重应进行加权求和得到算子性能评分，如公式（7）所示。考虑到不同的应用场景对于推理芯片在推理延迟和推理带宽上有不同的需求，因此推理芯片在单 batch 和多 batch 下的性能需要分别评测。

$$\text{算子性能评分} = \sum \text{算子类别权重} \times \text{算子评分} \quad \cdots\cdots (7)$$

6.3.2 模型性能

模型性能主要评测推理芯片在各计算机视觉领域中典型模型的整体推理性能。待测模型采用 ONNX 格式表示，选择范围见附录 B.2，待测模型列表参考附录 B.4，算子性能及权重设置见表 10。

表10 模型性能及权重设置

模型领域	权重
图像分类	0.5
目标检测	0.3
语义分割	0.2

采用推理延迟与参考平台的参考推理延迟的比值作为评分依据，比值越低评分越高。同时考虑到不同推理芯片会支持多种数据类型的模型推理，且模型推理已经比较接近实际应用场景需求，因此对于每个待测模型，只设置一个参考推理延迟，而对于推理芯片所支持的多种数据类型，分别算出比值并取最大值，如公式（8）所示。考虑到不同的应用场景对于推理芯片在推理延迟和推理带宽上有不同的需求，因此推理芯片在单batch和多batch下的性能需要分别评测。

$$\text{模型评分} = \mathbf{Max}\left(\frac{\text{参考推理延迟}}{\text{推理芯片推理延迟（指定数据类型）}}\right) \quad \cdots\cdots\cdots\cdots (8)$$

对于同一领域内的模型采用均等权重，如公式（9）所示。

$$\text{模型领域评分} = \frac{\sum \text{对应领域的模型评分}}{\text{对应领域下模型数量}} \quad \cdots\cdots\cdots\cdots (9)$$

对于推理芯片在多个领域的评分按照相应的权重进行加权求和得到模型性能评分，如公式（10）所示。

$$\text{模型性能评分} = \sum \text{模型领域权重} \times \text{模型领域评分} \quad \cdots\cdots\cdots\cdots (10)$$

6.4 软件生态

软件生态应从多个维度评测推理芯片的软件栈支持的特性，具体评测指标参考附录B.6。

6.5 应用场景测试

应用场景测试章节主要评测包含推理芯片在内的系统在实际计算机视觉任务的完整pipeline下的综合性能，包含推理芯片在内的设备端的总线接口、视频/图像解码、图像前处理、模型推理、后处理等能力。

具体待测任务参考附录B.5的应用场景测试-待测任务列表。

通过测试，采用实测整体吞吐量与参考平台的参考吞吐量的比值作为评分依据，比值越高评分越高。数据类型在满足精度损失要求下，选择吞吐量最高时对应的数据类型，如公式（11）所示。

$$\text{任务评分} = \mathbf{Max}\left(\frac{\text{推理芯片吞吐量(指定数据类型)}}{\text{参考吞吐量}}\right) \quad \cdots\cdots (11)$$

对推理芯片在附录B.5中的多个应用场景测试任务的评分赋予均等的权重，如公式（12）所示。

$$\text{应用场景测试评分} = \frac{\sum \text{任务评分}}{\text{待测任务数目}} \quad \cdots\cdots (12)$$

7 测试方法

7.1 基本技术规格

基本技术规格章节的各项指标的测试结果目前采用推理芯片的相关技术文档中的标称值以及厂商提供的其他相关技术信息。

7.2 功能

7.2.1 测试目标

测试推理芯片以及其软件栈是否支持附录B.1（算子功能-待测算子列表）和附录B.3（模型功能-待测模型列表）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：

a) 应提供处于最佳工作环境、厂商标配的主机配置、厂商标配的推理芯片产品形态；
b) 应提供推理芯片软件栈的相关技术文档；
c) 应提供单算子的编译库与运行库；
d) 应提供模型的编译库与运行库；
e) 宜提供模型量化工具。

7.2.3 测试要求

推理芯片以及其软件栈应满足以下要求：

a) 支持给定单算子或者模型的编译。其中，不要求单算子与推理芯片底层算子是一对一的关系，可以是多个底层算子组成的子图，即计算行为一致即可；
b) 支持至少一种数据类型。对于给定单算子或者模型，其输入输出要求是float32，若使用整型计算，量化与反量化应由推理芯片软件栈支持解决；
c) 支持给定单算子的运行，且在相同输入下，推理芯片的输出应与ONNX Runtime CPU(v1.10.0, Intel i7-8700@3.2GHz)的输出对齐，或在可接受精度损失范围内；
d) 支持给定模型的运行，且在模型指定的数据集输入下，推理芯片的输出精度与模型提供的官方精度的误差不超过1%。

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	数据准备	针对每一个算子，给定标准输入输出数据，其中标准输入数据为算子对应的float32类型的测试输入数据，标准输出数据为在标准输入数据下的ONNX Runtime CPU的输出，通常也为float32类型，某些情况下会包含int32等其他数据类型输出。
2	算子编译与运行	调用推理芯片的单算子编译库、量化工具（如必要）、运行库完成单算子的编译与运行，在给定的标准输入数据下获取推理芯片输出数据。
3	精度对比	使用绝对误差和相对误差综合比较标准输出数据和推理芯片输出数据，判断其是否在可接受精度损失范围内，对于不同的数据类型有不同的精度要求。

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	数据准备	针对每一个模型，给定测试数据集和官方测试精度。
2	模型编译与运行	调用推理芯片的模型编译库、量化工具（如必要）、运行库完成模型的编译与运行，在给定的测试数据集下测试得到推理芯片测试精度（top1/top5、mAP、mIoU等）。
3	精度对比	直接比较推理芯片测试精度与模型官方测试精度的误差（通常要求误差不超过1%）。

7.3 性能

7.3.1 测试目标

测试推理芯片以及其软件栈在附录B.2 算子性能-待测算子与算子参数列表的多算子和多算子参数以及附录B.4 模型性能-待测模型列表中的多模型下的推理耗时，其中单batch与多batch分别测试。

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、厂商标配的推理芯片产品形态；
b) 应提供推理芯片软件栈的相关技术文档；
c) 应提供单算子的编译库与运行库；
d) 应提供模型的编译库与运行库；
e) 宜提供模型量化工具；
f) 宜提供profiling工具。

7.3.3 测试要求

测试推理芯片在单算子和模型的推理耗时有以下测试要求：

a) 应先完成单算子或者模型的精度测试，且能满足精度要求；
b) 运行端宜采用C/C++接口，以保证测试结果的准确度；

   注：使用C的time模块或者C++的chrono模块测试推理前后的时间差（设备端）作为推理耗时，不包含主机端与设备端间数据传输的耗时。如推理芯片有提供profiling工具，可采用profiling工具的测试结果作为参考，但不作为推理耗时测试结果。

c) 单batch下应测试最短的推理耗时，多batch下应测试能达到最大吞吐量的推理耗时；

   注：其中，多batch会选择8、16、32、64、128、256作为固定测试档位以减少测试量，多batch在推理芯片上的高效调度需由厂商提供的软件栈或提供必要的测试样例。

d) 对于单算子或者模型的量化场景而言，在实际测试中runtime阶段因量化或格式转换引入的时间应作为推理耗时。

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	数据准备	针对每一个算子，给定标准输入输出数据，其中标准输入数据为算子对应的float32类型的测试输入数据，标准输出数据为在标准输入数据下的ONNX Runtime CPU的输出，通常也为float32类型，某些情况下会包含int32等其他数据类型输出。
2	单算子编译	调用推理芯片的单算子编译库、量化工具（如必要）完成单算子的编译，需包含单batch以及多batch的编译。
3	设备环境准备	完成必要的设备runtime管理，并完成必要的warmup，并将输入数据放在设备端内存上。
4	单算子运行	完成单算子在单batch下和多batch下的推理，记录推理前后时间差，并辅以profiling工具测试记录。
5	设备环境重置	重置设备runtime环境。

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	数据准备	针对每一个模型，给定测试数据集和官方测试精度。
2	模型编译	调用推理芯片的模型编译库、量化工具（如必要）完成模型的编译，需包含单batch以及多batch的编译。
3	设备环境准备	完成必要的设备runtime管理，并完成必要的warmup，并将输入数据放在设备端内存上。
4	模型运行	完成模型在单batch和多batch下的推理，记录推理前后时间差，并辅以profiling工具测试记录。
5	设备环境重置	重置设备runtime环境。

7.4 软件生态

软件生态章节的各项指标的评分应根据推理芯片的相关技术文档、配套的软件栈以及厂商提供的其他相关技术信息确定。

7.5 应用场景测试

7.5.1 测试目标

测试推理芯片以及其软件栈在附录B.5 应用场景-待测任务列表的整体吞吐量（fps）。

7.5.2 测试准备

应用场景测试需要被测方提供以下内容：

a) 应提供功能与性能测试所要求的材料；

b) 应提供应用场景测试的测试样例；

注：每一个应用场景对应一个样例，同时需提供详细的文档说明。

c) 测试样例应包含明确的运行时的各项信息以及测试结果。

7.5.3 测试要求

测试推理芯片在待测任务下的整体吞吐量应满足以下要求：

a) 应用场景测试的测试用例由推理芯片厂商提供，以满足对应模型功能测试的精度要求下追求系统整体的最大吞吐量；

b) 对主机端CPU不做要求，以厂商推荐的主机端配置为准；

注：但不能使用主机端CPU完成诸如数据前后处理、视频图像编解码，只做计算调度使用，如因推理芯片无法共相应的能力需要主机端CPU协助计算，应单独标明。

c) 整体流程符合基本要求，采用给定的输入测试数据，保证测试吞吐量的真实性；

d) 测试精度满足模型功能测试的精度要求。

7.5.4 测试流程

推理芯片厂商提供的实际应用场景的测试样例应满足以下测试流程：

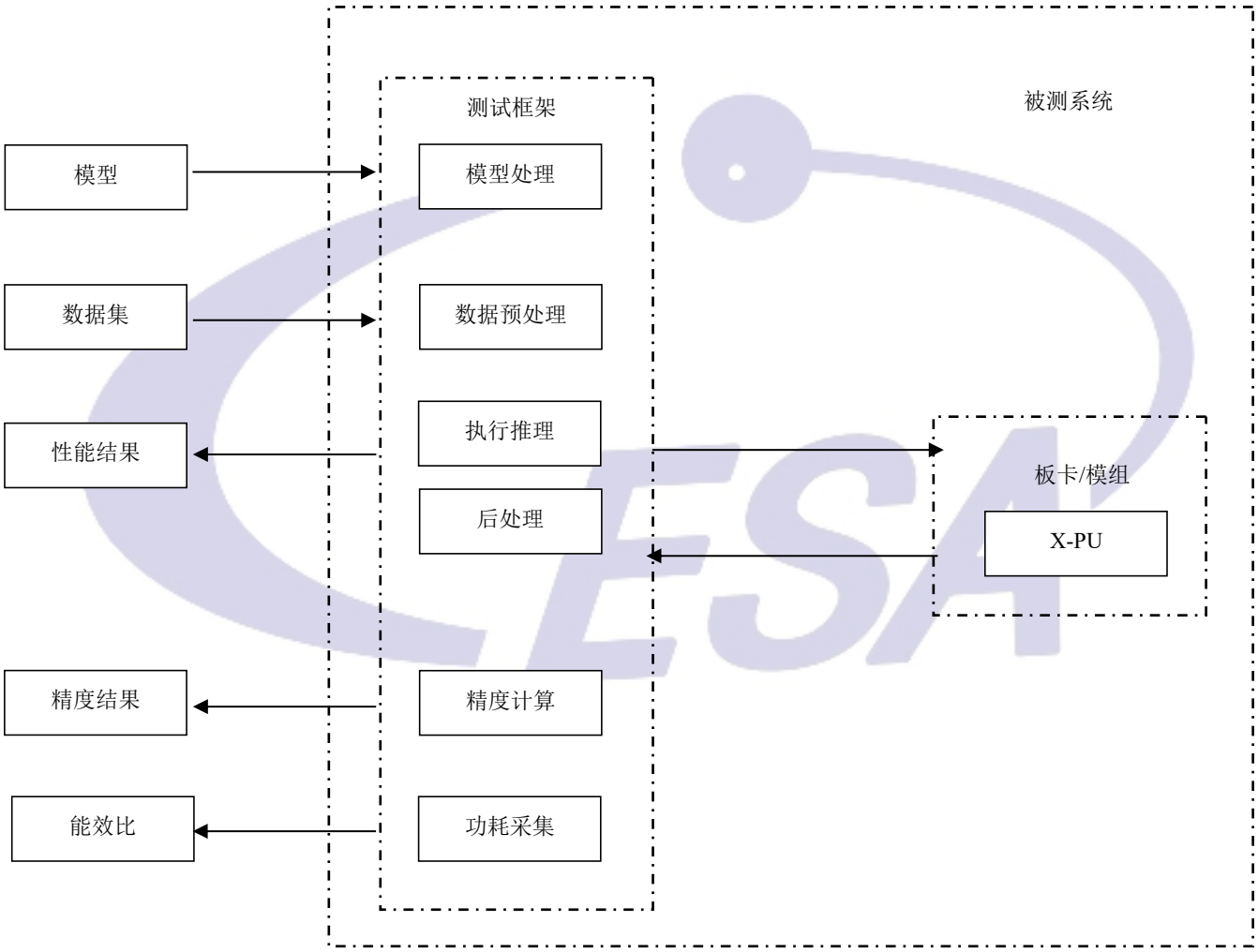
表15 应用场景的测试流程

序号	步骤	步骤说明
1	准备数据	准备输入图像数据或者视频数据。
2	数据传输	将输入数据由主机端传输到设备端。
3	解码	图像或者视频数据解码。
4	前处理	完成必要的图像前处理。
5	模型推理	完成对应模型的推理。
6	后处理	完成必要的后处理操作。
7	数据传输	将输出数据由设备端传输到主机端。
8	精度对比	将输出数据与参考输出数据进行对比得到整体精度。
9	多路配置	上述步骤可以并行执行多路处理流程以达到系统整体最大吞吐量。

附　录　A
（资料性）
推理测试参考框架

A.1　推理测试框架

推理测试框架说明见图A.1。



注：X-PU为被测计算机视觉推理用深度学习芯片，包括GPU、NPU等。

1.输入，本标准指定的模型和数据集；

2.模型处理，可选，如量化、使用模型转换工具对模型进行转换；

3.送测方提供使能其系统的接口供测试框架调用，比如初始化，模型加载、执行、卸载。

图A.1 推理测试框架

附　录　B
（规范性）
算子及模型列表

B.1　算子功能测试列表

表B.1　算子功能测试列表

级别	算子	备注
第 1 级算子	Add, AveragePool, Concat, Conv, Gemm, GlobalAveragePool, GlobalMaxPool, MaxPool, Relu, Reshape	
第 2 级算子	Argmax, BatchNormalization, ConvTranspose, LeakyRelu, MatMul, PRelu, Resize, RoiAlign, Slice, Squeeze, Unsqueeze, Transpose, Sigmoid, Softmax	
第 3 级算子	Cast, Clip, ConstantOfShape, Div, Exp, Expand, Equal, Greater, Less, LRN, Not, Mul, NonMaxSuppression, Range, ReduceMax, ScatterND, Shape, Sub, Tile, TopK, Where	

B.2　算子性能配置列表

表B.2　Conv算子计算配置

序号	C_i	W	H	C_o	kernel	pad	stride
1	3	224	224	64	3	1	1
2	64	112	112	128	3	1	1
3	128	56	56	256	3	1	1
4	256	28	28	512	3	1	1
5	512	14	14	512	3	1	1
6	512	7	7	512	3	1	1
7	3	224	224	64	7	3	2
8	192	28	28	32	5	2	1
9	192	28	28	64	1	0	1
10	512	14	14	48	5	2	1
11	512	14	14	192	1	0	1
12	832	7	7	256	1	0	1
13	832	7	7	128	5	2	1
14	1	480	48	16	3	1	1
15	16	240	24	32	3	1	1
16	32	120	12	64	3	1	1
17	64	60	6	128	3	1	1
18	3	108	108	64	3	1	2
19	64	54	54	64	3	1	1

20	128	27	27	128	3	1	1
21	128	14	14	256	3	1	1
22	256	7	7	512	3	1	1
23	64	56	56	64	3	1	1
24	64	56	56	256	1	0	2
25	128	28	28	128	3	1	1
26	128	28	28	512	1	0	2
27	256	14	14	256	3	1	1
28	256	14	14	1024	1	0	2
29	512	7	7	512	1	0	1
30	2048	7	7	512	1	3	2
31	64	112	112	64	1	0	1
32	64	56	56	256	1	0	1

表B.3　Gemm算子计算配置

序号	M	N	K	A Transpose	B Transpose
1	1	1000	2048	0	1
2	1	4096	9216	0	1
3	1	4096	4096	0	1
4	1	1000	4096	0	1
5	1	1000	1280	0	1
6	1	1000	1024	0	1
7	1	1000	512	0	1
8	1	4096	25088	0	1
9	1	1024	12544	0	1
10	1	1024	1024	0	1
11	1	1024	81	0	1
12	1	1024	320	0	1

表B.4　NonMaxSuppression算子计算配置

序号	max_output_boxes_per_class	iou_threshold	score_threshold	boxes(dim)	scores(dim)
1	1000	0.7	0	(1, 1000, 4)	(1, 1, 1000)
2	100	0.5	0.05	(1, 80000, 4)	(1, 1, 80000)
3	200	0.5	0.05	(1, 3900, 4)	(1, 80, 3900)

注：上述各项配置的 centor_point_box 设置为 0。

表B.5　TopK算子计算配置

序号	axis	largest	sorted	X(dim)	K

1	-1	1	1	(1, 72)	5/20
2	-1	1	1	(1, 1000)	8/100/400/600
3	-1	1	1	(1, 80000)	8/100/400/600
4	-1	1	1	(1, 312000)	8/100/400/600

表B.6 RoiAlign算子计算配置

序号	output_height	output_width	spatial_scale	input (dim)	rois (dim)
1	7	7	0.25	(1, 256, 140, 84)	(1000, 5)
2	7	7	0.125	(1, 256, 70, 42)	(1000, 5)
3	7	7	0.0625	(1, 256, 35, 21)	(1000, 5)
4	7	7	0.03125	(1, 256, 18, 11)	(1000, 5)
5	14	14	0.25	(1, 256, 140, 84)	(100, 5)
6	14	14	0.125	(1, 256, 70, 42)	(100, 5)
7	14	14	0.0625	(1, 256, 35, 21)	(100, 5)
8	14	14	0.03125	(1, 256, 18, 11)	(100, 5)

注：上述各项配置的coordinate_transformation_mode为half_pixel，sampling_ratio为0，mode为avg。

B.3 模型功能测试列表

表B.7 模型功能测试列表

类别	网络模型	数据集(精度验证)	备注
图像分类	alexnet	ImageNet2012 val （Top-1, Top-5）	来自 openmmlab 的模型，使用项目提供的模型转换脚本将模型转换成 onnx 模型，脚本路径 https://github.com/open-mmlab/mmclassification/blob/master/tools/deployment/pytorch2onnx.py。
	googlenet		
	zfnet		
	vgg16		
	efficientnet		
	squeezenet		
	resnet50-v1		
	resnext50		
	se-resnet50		
	shufflenet-v2		
	mobilenet-v2		
	inception-v3		
	densenet121		
目标检测	faster-rcnn	COCO val2017 (mAP)	来自 openmmlab 的模型，使用项目提供的模型转换脚本将模型转换成 onnx 模型，脚本路径 https://github.com/open-mmlab/mmdetection/blob/master/tools/deployment/pytorch2o
	mask-rcnn		
	retinanet		
	fsaf		

	ssd		nnx.py
	yolo-v3		
	fcos		
	solo		
语义分割	deeplab-v3	Cityscapes（mIoU）	来自 openmmlab 的模型，使用项目提供的模型转换脚本将模型转换成 onnx 模型，脚本路径 https://github.com/open-mmlab/mmsegmentation/blob/master/tools/pytorch2onnx.py
	pspnet		
	unet		
	fcn		

B.4 模型性能测试列表

表B.8 模型性能测试列表

类别	网络模型
图像分类	ResNet50，ShuffleNetV2，MobileNetV2，VGG16
目标检测	Faster R-CNN，Mask R-CNN，Retinanet，YOLOv3
语义分割	DeeplabV3，PSPNet

B.5 应用场景测试列表

表B.9 应用场景测试列表

待测任务	测试数据集	测试模型	输入尺寸（CHW）	精度验证
图像分类	ImageNet2012-val	ResNet50	3*224*224	Top-1,Top-5
目标检测	COCO_val2017	YOLOv3	3*416*416	mAP

附　录　C
（规范性）
软件生态评测指标

C.1　软件生态评测指标表

表C.1　软件生态评测指标表

评测维度	评测指标	
模型编译	支持接入多种前端模型格式	支持caffe模型接入
		支持onnx模型接入
		支持pytorch模型接入
		支持tensorflow模型接入
		支持其它格式的模型接入
	支持通过IR api构造网络，适配自研模型格式	
	支持离线模型编译，不依赖硬件设备	
	支持标记额外的输出tensor	
运行时管理	支持动态batch size设置	
	支持动态输入（height width）设置	
	支持动态图网络模型	
	支持异步推理	
	支持多种tensor layout设置	NCHW
		NHWC
		NHWCD
		NCDHW
	支持设置输入输出tensor的user buffer，用于数据零拷贝优化	
	支持获取输入输出tensor的信息	获取tensor name信息
		获取tensor shape信息
		获取tensor量化参数信息
模型量化	支持多种量化方案	A/W都支持非对称量化
		A/W之一支持非对称量化
		A/W仅支持对称量化
	支持多种量化粒度	支持分层量化
		支持分通道量化
	支持浮点数据类型的scale量化参数	
	支持多种算子的量化计算	支持conv算子量化
		支持gemm算子量化
		支持pool算子量化
		支持add算子量化
		支持concat算子量化
	支持多种量化精度	支持int16量化

		支持int8量化
		支持int4量化
	支持混合精度量化	
	支持用户设置量化参数	支持用户设置权重的量化参数
		支持用户设置激活的量化参数
性能加速	支持稀疏计算，且能够进行加速计算	
图像处理能力	支持图像处理的resize计算加速	
	支持图像处理的crop计算加速	
	支持图像处理的warp affine计算加速	
	支持图像处理的color convert计算加速	
扩展功能	支持编程语言级别的底层开发	
	支持用户自定义算子接入	
	编译推理框架开源	
	提供推理性能profiling工具	
	提供代码调试工具	
	提供精度对比（逐层）工具	
	支持float数据类型的输入tensor	
	模型编译过程与运行时过程可分离	
	软件栈支持多种系统平台部署	模型编译器支持linux系统部署
		模型编译器支持windows系统部署
		模型编译器支持macos系统部署
		运行时库支持linux系统部署

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