ICS 33.050.20

M 30

电信终端产业协会标准

T/TAF 063-2020

移动智能终端人工智能性能基准测试方法

Benchmark Test Methods for AI Performance of Intelligent Terminal

2020-08-04 发布　　2020-08-04 实施

电信终端产业协会　发布

目　次

目　次 I
前　言 II
引言 III
标动智能终端人工智能性能基准测试方法 1
1 范围 1
2 规范性引用文件 1
3 文件清单的排列顺序： 1
4 术语和定义 1
4.1 神经网络模型 1
4.2 推理集 1
4.3 端侧人工智能推理框架 1
4.4 模型转换工具 2
4.5 深度学习编译器 2
4.6 基准测试例 2
4.7 终端硬件 2
5 测试概述 2
5.1 测试构架 2
5.2 通用测试方法 3
5.3 性能指标监测 3
6 图像处理测试方法 3
6.1 图像分类测试方法 3
6.2 人脸识别测试方法 4
6.3 目标语义分割测试方法 5
6.4 图片超分辨率测试方法 6
6.5 目标检测测试方法 7
7 视频处理测试方法 8
7.1 视频目标检测测试 8
附　录　A （规范性附录） 标准修订历史 9
附　录　B 图像语义分割测试类别 9
附　录　C 图像超分辨率测试推断集 10
附　录　D 目标检测类别 10
参考文献 13

前　言

本标准按照 GB/T-2009 给出的规则起草。

本标准中的某些内容可能涉及专利。本标准的发布机构不承担识别这些专利的责任。

本标准由电信终端产业协会提出并归口。

本标准起草单位：中国信息通信研究院、维沃移动通信有限公司、OPPO广东移动通信有限公司

本标准主要起草人：解谦，卢炳全，高立发，贾利敏

引　　言

随着人工智能的飞速发展，为满足低响应时间，高安全可靠性以及在任意使用环境下（如无网络）使用AI场景，一部分AI应用将以部署在终端设备的方式运行，如移动智能手机，平板电脑等。一款智能移动终端AI处理性能的好坏，一般可以通过基准测试的方式衡量。本标准基于移动终端推理框架技术，提出一个合理、公平，能反映出终端实际的AI处理能力的基准测试方法，包括对AI基准测试在不同应用场景的数据集的技术要求，测试方法和评测指标，旨在让终端AI处理性能测试得到可靠的，可比较的，能体现终端AI处理能力差异的评测结果，推动智能移动终端向AI终端发展。

移动智能终端人工智能性能基准测试方法

1 范围

本标准规定了通过使用端侧人工智能推理框架在移动智能终端侧进行推理计算的基准测试的方法，可以对终端基于神经网络模型的计算性能进行评估。评测场景包括图像处理、视频处理等不同场景，针对不同场景测试集，测试方法和评测指标提出要求。

本标准适用于具备智能操作系统的移动智能终端，包括数字移动电话机，平板电脑以及其他数字移动通信终端设备。

2 规范性引用文件

下列文件对于本标准的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本标准。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本标准。

3 文件清单的排列顺序：

a）国家标准；
b）行业标准；
d）国内有关文件；
e）国际标准（含ITU标准、ISO/IEC标准等）；
f）ISO或IEC有关文件；
g）其他国际标准以及其他国际有关文件。

4 术语和定义

4.1 神经网络模型

封装了神经网络算法和参数的特定格式的文件，用于人工智能计算。本标准的神经网络模型应为经过训练且达到一定准确率的模型。

4.2 推理集

作为人工智能推理计算的输入数据集，可以为图片，视频等格式的数据或文件。

4.3 端侧人工智能推理框架

端侧人工智能推理框架部署在移动智能终端上，通常由模型转换工具和深度学习编译器组成。端侧人工智能推理框架可以分为通用框架和专用框架，通用框架能跨平台运行，能在多种芯片平台上运行的人工计算，如TensorFlow Lite，Paddle Lite等。专用框架指仅能在指定的部分芯片平台上运行的人工智能计算，如SNPE，HiAI等。在测试过程中需要指明使用的端侧人工智能推理框架。

4.4 模型转换工具

模型转换工具能将输入的神经网络模型，根据移动终端特点进行剪裁压缩和优化，具有减小模型体积、优化算法操作和参数精度等功能。

4.5 深度学习编译器

用于解决深度神经网络模型在使用不同底层硬件计算芯片计算的适配等问题，为上层应用的执行提供硬件加速能力。

4.6 基准测试例

基准测试例为指定测试场景下，使用神经网络模型推理算法对推理测试集进行推理测试的测试例。

4.7 终端硬件

参与人工智能处理的硬件，包括CPU、GPU、AI硬件加速单元,内存、电池等。

5 测试概述

5.1 测试构架

基准测试指通过运行一段（一组）程序或者操作，来评测终端相关性能的活动。移动智能终端人工智能性能基准测试指通过端侧人工智能推理框架,运行不同的神经神经网络模型和测试负载进行推理运算，以此来综合评价测试对象的AI计算性能。移动智能终端人工智能性能基准测试包括图像处理、视频处理测试。具体测试框架如下，见图1：

图1 人工智能基准测试构架

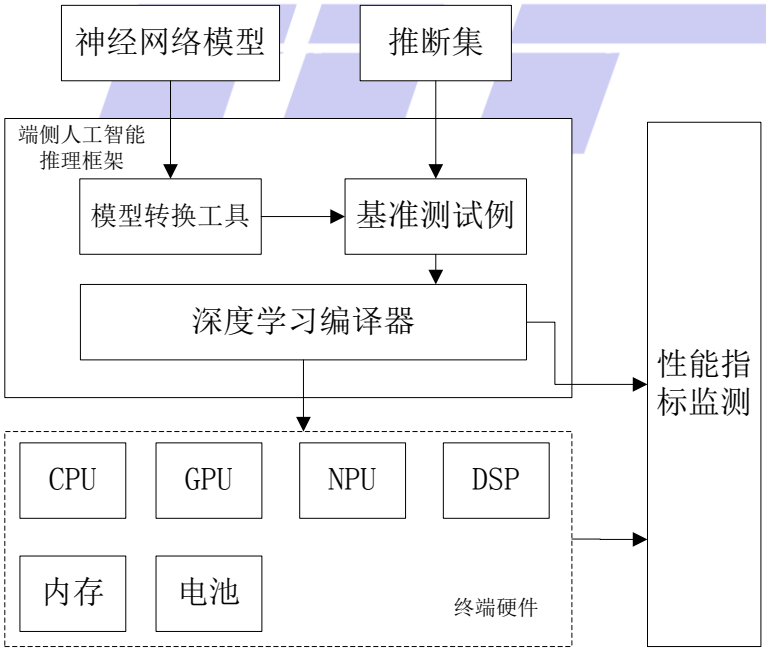

图 1 移动智能终端人工智能性能基准测试构架

5.2 通用测试方法

a) 移动智能终端初始化，包括屏蔽测试无关的其他应用、后台功能、调整屏幕亮度、记录初始电量等,使得每次测试前终端的运行状态保持一致；
b) 使用模型优化工具将预训练的模型文件离线转换为移动智能终端上可以直接运行的模型文件，并进行优化；
c) 将测试例推理集的图像或视频资源进行缩放、通道转换等预处理工作；
d) 将预处理的测试图像或视频资源输入优化后的模型进行推理测试；
e) 测试过程中通过软件方式或其他方式记录模型指标和硬件性能指标。

5.3 性能指标监测

性能指标包括检测人工智能推断计算性能的模型性能指标和硬件性能指标。模型性能指标参见第5章内容。硬件性能指标为通用测试指标包括功耗，内存，CPU等，具体为：

1) 功耗为测试过程中损失的电量百分比；

2) 内存平均占用为测试过程中测试工具占用的平均内存量；

3) 测试过程中CPU平均使用率（可选）；

4) 测试过程中的CPU平均工作温度（可选）；

5) 测试过程中电池平均工作温度（可选）。

6 图像处理测试方法

6.1 图像分类测试方法

测试编号	1
测试名称	图像分类测试
测试描述	根据各自在图像信息中所反映的不同特征，把不同类别的目标区分开来的图像处理方法。
推理集要求	推理集应由公开渠道可自由获取的非商业用途图片数据构成，可选的公开数据集包括如下图片集： 1. CIFAR-100； 2. Caltech_256； 3. ImageNet。 进行基准测试时，应从公开数据集的测试集中随机抽取10000张图片，分类类型不少于100类。
模型要求	评测模型可以选择下表所列深度学习模型： 1. Inception v3； 2. MobileNet V1。
测试步骤	1) 加载数据集中的图片到终端内存，并完成图像缩放、通道转换等预处理工作；

	2） 评测软件记录本次图片推理前的时间戳； 3） 将内存中预处理后的数据输入推理模型； 4） 记录模型输出结果和该时刻的时间戳； 5） 重复步骤a)-d，直到数据集所有图片完成测试，输出记录，计算指标； 6） 测试需要使用 float 精度或 int 精度的模型分别进行测试。
测试指标	TOP1准确率（VTop1，单位：%）：在一次推理结果分类排序中，只有当概率最高的结果为正确分类，本次推理结果才能判定为正确，统计所有图片的推理结果，用正确推理图片数量除以图片总数，得到TOP1准确率。 $$VTop1=\frac{TP1}{TP1+FN1}\times 100\%$$ TP1：推理结果中，Top1 分类正确的图片数量； FN1：推理结果中，Top1 分类不正确的图片数量。
	TOP5准确率（VTop5，单位：%），在一次推理结果分类排序中，概率排名前五的结果中包含正确的分类，本次推理结果判定为正确，统计所有图片的推理结果，用正确推理图片数量除以图片总数，得到TOP5准确率。 $$VTop5=\frac{TP5}{TP5+FN5}\times 100\%$$ TP5：推理结果中，Top5 分类正确的图片数量； FN5：推理结果中，Top5 分类不正确的图片数量。
	单张图片推理时间（Inference Time，单位：毫秒）：记录一组图片推理总耗时，计算出单张图片平均推理时间： $$Inference\ Time=\frac{TN}{N}$$ TN：一组图片推理总耗时； N：该组图片数量。

6.2 人脸识别测试方法

测试编号	2
测试名称	人脸识别测试
测试描述	针对人脸照片进行特征提取和比对，并根据终端的平均处理时长，量化移动终端的性能。
推理集要求	推理集应由公开渠道可自由获取的非商业用途图片数据构成，可选的公开数据集包括如下图片集： 1. Labeled Faces in the Wild Home (LFW) 2. MegaFace 3. PubFig: Public Figures Face Database 4. Colorferet 进行基准测试时，应从公开数据集的测试集中随机抽取10000组，选取对象

<table>
<tr><td></td><td>按照不同年龄段和不同性别两个维度选取，至少包括男性儿童，女性儿童，男性成人，女性成人，男性老人，女性老人。</td></tr>
<tr><td>模型要求</td><td>评测模型可以选择下表所列深度学习模型：
1. facenet</td></tr>
<tr><td>测试步骤</td><td>1) 选取符合要求的推理集作为测试样例，建立对应的文件列表；
2) 将文件列表送入对比识别算法程序，开始执行程序；
3) 从推理算法程序读取文件列表时开始计时，记录200组图片对比完成所需要的时间和对比结果；
4) 与数据库中的图像关系对比，计算测试样例的正确通过率，错误接受率。统计错误率，错误接受率为百万分之一，千分之一，万分之一处的正确通过率；
5) 测试需要使用 float 精度或 int 精度的模型分别进行测试。</td></tr>
<tr><td>测试指标</td><td>正确通过率(Pass Rate，PR，单位：%)在真实的验证过程中（正确指纹）同一个人的样本被判断为同一个人的比对次数占总比对次数的比例：
$$PR=\frac{TP}{TP+FN}\times 100\%$$
TP：同一个人的样本对被判断为同一个人的比对次数；
FN：同一个人的样本对被判为不同人的比对次数。</td></tr>
<tr><td></td><td>错误接受率(False Acceptance Rate，FAR，单位：%)在冒充攻击尝试（错误指纹）中被错误接受的比例：
$$FAR=\frac{FP}{TN+FP}\times 100\%$$
FP：不同人的样本对被判为同一个人的比对次数；
TN：不同人的样本对被判为不同人的比对次数。</td></tr>
<tr><td></td><td>单张图片推理时间（Inference Time，单位：毫秒）：记录200组图片推理总耗时，计算出单张图片平均推理时间：
$$\text{Inference Time}=\frac{TN}{N}$$
TN：一组图片推理总耗时；
N：该组图片数量。</td></tr>
</table>

6.3 目标语义分割测试方法

测试编号	3
测试名称	图像语义分割测试
测试描述	图像语义分割（Image Semantic Segmentation）融合了传统的图像分割和目标识别两个任务，将图像分割成一组具有一定语义含义的块，并识别出每个分割块的类

	别，最终得到一幅具有逐像素语义标注的图像。
推理集要求	推理集应由公开渠道可自由获取的非商业用途图片数据构成，可选的公开数据集包括如下图片集： 1. PASCAL VOC2012 进行基准测试时，应从公开数据集的测试集中随机抽取1000张，至少包括附录B的分类。
模型要求	评测模型可以选择表所列深度学习模型： 1. unet； 2. deeplabv3。
测试步骤	使用训练好的神经网络算法对推理集图片进行语义分割： a) 测试过程记录每个数据的推导时间（入口和出口时间差）； b) IoU计算方法： 1) 分别加载标注图和结果图； 2) 根据标注的对象颜色和结果图中对象颜色，统计颜色吻合的像素点； 3) 根据标注对象颜色和结果图对象颜色，统计色块像素； 4) 根据统计结果计算IoU； 5) 其他分类范围也用相同的方式分别计算IoU； c) 测试需要使用float精度或int精度的模型分别进行测试。
测试指标	分割类别 支持分割的对象类别，记录识别出超出推理集要求的种类个数和少于推理集要求的种类个数之和。
	测试集的平均mIoU： $$mIoU = \frac{1}{N}\sum_{t=1}^{N}(IoU)$$ IoU: Intersection over Union，用于评价单一目标上检测的准确度。 IoU为推理结果区域与实际目标区域的交集比并集。
	单张图片推理时间（Inference Time，单位：毫秒）：记录一组图片推理总耗时，计算出单张图片平均推理时间： $$\text{Inference Time}=\frac{TN}{N}$$ TN：一组图片推理总耗时 N：该组图片数量

6.4 图片超分辨率测试方法

测试编号	4
测试名称	图片超分辨率测试
测试描述	指由一幅低分辨率图像或图像序列恢复出高分辨率图像。

<table>
<tr><td>推理集要求</td><td>推理集应由公开渠道可自由获取的非商业用途图片数据构成，可选的公开数据集见附录C。进行基准测试时，应从公开数据集的测试集中随机抽取10000张图片。</td></tr>
<tr><td>模型要求</td><td>评测模型可以选择下表所列深度学习模型：
1. SRCNN
2. vdsr</td></tr>
<tr><td>测试步骤</td><td>a) 依据具体的使用场景先将推断集图片压缩，然后使用训练好的神经网络算法对压缩图片进行超分放大。
b) 测试过程记录每个数据的推导时间（入口和出口时间差）；
c) 使用超分放大图片和原始图片质量计算评测指标；
d) 测试需要使用float精度或int精度的模型分别进行测试。</td></tr>
<tr><td rowspan="3">测试指标</td><td>PSNR（峰值信噪比）值
$$MSE=\frac{1}{N}\sum_{i=1}^{N}(x(i)-y(i))^2$$
$$PSNR=10*\log_{10}(\frac{L^2}{MSE})$$
$x(i),y(i)$：图像 $x,$y 像素值；
L：像素值的动态范围，一般取255；
N： 图像 $x,$y 的像素数（ $x,$y 分辨率相同）。</td></tr>
<tr><td>SSIM（结构相似度）值
$$SSIM(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$$
μ_x,μ_y： 图像 $x,$y 的均值；
σ_x^2,σ_y^2： 图像 $x,$y 的方差；
σ_{xy}： 图像 $x,$y 的协方差；
$c_1=(k_1L)^2$， $c_2=(k_2L)^2$：用来维持稳定的常数， L 是像素值的动态范围，一般取255， $k_1=0.01,k_2=0.03$。</td></tr>
<tr><td>单张图片推理时间（Inference Time，单位：毫秒）：记录一组图片推理总耗时，计算出单张图片平均推理时间：
$$\text{Inference Time}=\frac{TN}{N}$$
TN：一组图片推理总耗时
N：该组图片数量</td></tr>
</table>

6.5 目标检测测试方法

测试编号	5
测试名称	目标检测测试
测试描述	目标检测，也称为目标提取，是一种基于目标几何和统计特征的图像分割技术。其综合了图像分割和识别，能够提取图片中的对象类别以及具体位置信息。
推理集要求	推理集应由公开渠道可自由获取的非商业用途图片数据构成，可选的公开数据集包括如下图片集： 1. COCO 进行基准测试时，应从公开数据集的测试集中随机抽取10000张图片，分类类型见附录D，每类随机选取20张样本图片。
模型要求	评测模型可以选择下表所列深度学习模型： 1. MobileNetV2-SSD, 2. ResNet-SSD.
测试步骤	1) 加载数据集中1张图片到终端内存，并完成图像缩放、通道转换等预处理工作； 2) 评测软件记录本批次图片推理前的时间戳； 3) 将内存中预处理后的数据单张输入推理模型； 4) 记录模型输出结果和该时刻的时间戳； 5) 重复步骤a)-d，直到数据集所有图片完成测试，输出记录，计算指标； 6) 测试需要使用float精度或int精度的模型分别进行测试。
测试指标	准确度 $mAP^{0.5}$ ：在IoU阈值为0.5的前提下，在所有类别上的mAP值。 mAP: Mean Average Precision，用于评价在全部测试样本上的准确度。与 IoU 设置紧密相关。 IoU: Intersection over Union，用于评价单一目标上检测的准确度。 IoU 为推理结果区域与实际目标区域的交集比并集。
	单张图片推理时间（ Inference Time，单位：毫秒）：记录一组图片推理总耗时，计算出单张图片平均推理时间： $$\text{Inference Time}=\frac{TN}{N}$$ TN：一组图片推理总耗时； N：该组图片数量。

7 视频处理测试方法

7.1 视频目标检测测试

测试编号	7
测试名称	对视频中的内容进行目标检测处理
测试描述	根据各自在图像信息中所反映的不同特征，把不同类别的目标区分开来的图像处理方法。

推理集要求	拍摄一段有代表性的3分钟街景视频，确保内容明确、目标丰富。将视频帧进行人工切割和目标标注，以此形成最终视频输入样本。至少包括建筑，汽车，行人，交通指示牌。
模型要求	见 6.5
测试步骤	1) 按每帧将视频中的图像取出进行处理； 2) 其余测试方法参考5.6。
测试指标	速度FPS：Frame Per Second，每秒钟最大能处理的图片张数。
	准确度 $mAP^{0.5}$ ：在IoU阈值为0.5的前提下，在所有类别上的mAP值。 mAP: Mean Average Precision，用于评价在全部测试样本上的准确度。与 IoU 设置紧密相关。 IoU: Intersection over Union，用于评价单一目标上检测的准确度。 IoU 为推理结果区域与实际目标区域的交集比并集。

附 录 A
（规范性附录）
标准修订历史

修订时间	修订后版本号	修订内容

附 录 B
图像语义分割测试类别

序号	父类	子类
1	人（Person）	人（person）
2	动物(Animal)	鸟（bird）
3	动物(Animal)	猫（cat）
4	动物(Animal)	牛（cow）
5	动物(Animal)	狗（dog）
6	动物(Animal)	马（horse）
7	动物(Animal)	羊（Sheep）
8	交通工具(Vehicle)	飞机（aeroplane）
9	交通工具(Vehicle)	自行车（bicycle）

10	交通工具(Vehicle)	船（boat）
11	交通工具(Vehicle)	巴士（bus）
12	交通工具(Vehicle)	车（car）
13	交通工具(Vehicle)	摩托车（motorbike）
14	交通工具(Vehicle)	火车（train）
15	Indoor(室内家具)	瓶子（bottle）
16	Indoor(室内家具)	椅子（chair）
17	Indoor(室内家具)	餐桌（dining table）
18	Indoor(室内家具)	盆栽（potted plant）
19	Indoor(室内家具)	沙发（sofa）
20	Indoor(室内家具)	电视/监视器（tv/monitor）

附 录 C
图像超分辨率测试推断集

序号	数据集名称	数量	分辨率	格式	种类
1	BSDS300	300	（435,367）	JPG	动物，建筑，食物，风景，人物，植物等
2	BSD500	500	(432,370)	JPG	动物，建筑，食物，风景，人物，植物等
3	DIV2K	1000	(1972,1437)	PNG	环境，植物，动物，手工制品，人物，风景等
4	General-100	100	(435,381)	BMP	动物，日用品，食物，人物，植物，地质等
5	L20	20	(3843,2870)	PNG	动物，建筑，风景，人物，植物等
6	Manga109	109	(826,1169)	PNG	漫画
7	OutdoorScene	10624	(553,440)	PNG	动物，建筑，草，山，植物，天空，水
8	PIRM	200	(617,482)	PNG	环境，植物，自然风景，人物等
9	Set5	5	(313,336)	PNG	小孩，鸟，蝴蝶，头，女人
10	Set14	14	(492,446)	PNG	人类，动物，昆虫，花，蔬菜，漫画等
11	T91	91	(264,204)	PNG	车，花，水果，人脸等
12	Urban100	100	(984,797)	PNG	建筑，城市，结构等

附 录 D
目标检测类别

序号	COCO类别编号	目标类别	父类
1	1	人 person	人 Person
2	2	自行车 bicycle	交通工具 Vehicle
3	3	汽车 car	交通工具 Vehicle
4	5	飞机 airplane	交通工具 vehicle

5	7	火车 train	交通工具 vehicle
6	9	船 boat	交通工具 vehicle
7	10	交通信号灯 traffic light	室外 outdoor
8	11	消防栓 fire hydrant	室外 outdoor
9	12	路标 street sign	室外 outdoor
10	13	停止标识 stop sign	室外 outdoor
11	16	鸟 bird	动物 animal
12	17	猫 cat	动物 animal
13	18	狗 dog	动物 animal
14	19	马 horse	动物 animal
15	20	羊 sheep	动物 animal
16	26	帽子 hat	动物 accessory
17	27	登山包 backpack	配件 accessory
18	28	雨伞 umbrella	配件 accessory
19	29	鞋子 shoe	配件 accessory
20	30	眼镜 eye glasses	配件 accessory
21	31	手包 handbag	配件 accessory
22	35	滑雪 skis	运动 sports
23	37	运动球 sports ball	运动 sports
24	38	风筝 kite	运动 sports
25	44	瓶子 bottle	厨房 kitchen
26	45	盘子 plate	厨房 kitchen
27	47	杯子 cup	厨房 kitchen
28	50	勺子 spoon	厨房 kitchen
29	51	碗 bowl	厨房 kitchen
30	52	香蕉 banana	食物 food
31	53	苹果 apple	食物 food
32	59	披萨 pizza	食物 food
33	61	蛋糕 cake	食物 food
34	62	椅子 chair	家具 furniture
35	63	长椅 couch	家具 furniture
36	64	盆栽 potted plant	家具 furniture
37	65	床 bed	家具 furniture
38	66	镜子 mirror	家具 furniture
39	68	窗户 window	家具 furniture
40	69	桌子 Desk	家具 furniture
41	71	门 Door	家具 furniture
42	72	电视 Tv	电子产品 electronic
43	73	笔记本电 laptop	电子产品 electronic
44	74	鼠标 mouse	电子产品 electronic

45	76	键盘 keyboard	电子产品 electronic
46	77	移动电话 cell phone	电子产品 electronic
47	82	冰箱 refrigerator	加电 appliance
48	84	书 book	室内 indoor
49	85	闹钟 clock	室内 indoor
50	89	吹风机 hair drier	室内 indoor

参 考 文 献

电信终端产业协会团体标准

标准中文名称

T/TAF 063—2020

电信终端产业协会印发
地址：北京市西城区新街口外大街 28 号
电话：010-82052809
电子版发行网址：www.taf.org.cn