International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

F.748.11

(08/2020)

SERIES F: NON-TELEPHONE TELECOMMUNICATION SERVICES

Multimedia services

Metrics and evaluation methods for a deep neural network processor benchmark

Recommendation ITU-T F.748.11

ITU-T F-SERIES RECOMMENDATIONS

NON-TELEPHONE TELECOMMUNICATION SERVICES

TELEGRAPH SERVICE	
Operating methods for the international public telegram service	F.1–F.19
The gentex network	F.20–F.29
Message switching	F.30–F.39
The international telemessage service	F.40–F.58
The international telex service	F.59–F.89
Statistics and publications on international telegraph services	F.90–F.99
Scheduled and leased communication services	F.100–F.104
Phototelegraph service	F.105–F.109
MOBILE SERVICE	
Mobile services and multidestination satellite services	F.110–F.159
TELEMATIC SERVICES	
Public facsimile service	F.160–F.199
Teletex service	F.200–F.299
Videotex service	F.300–F.349
General provisions for telematic services	F.350–F.399
MESSAGE HANDLING SERVICES	F.400–F.499
DIRECTORY SERVICES	F.500–F.549
DOCUMENT COMMUNICATION	
Document communication	F.550–F.579
Programming communication interfaces	F.580–F.599
DATA TRANSMISSION SERVICES	F.600–F.699
MULTIMEDIA SERVICES	**F.700–F.799**
ISDN SERVICES	F.800–F.849
UNIVERSAL PERSONAL TELECOMMUNICATION	F.850–F.899
ACCESSIBILITY AND HUMAN FACTORS	F.900–F.999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T F.748.11

Metrics and evaluation methods for a deep neural network processor benchmark

Summary

Recommendation ITU-T F.748.11 provides the benchmarking framework, evaluation metrics and methods, and application scenarios for deep neural network processors operating training and inference tasks.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T F.748.11	2020-08-13	16	11.1002/1000/14329

Keywords

Benchmark, cloud, deep neural network processor, metrics, mobile.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

© ITU 2020

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

		Page
1	Scope	1
2	References	1
3	Definitions	1
3.1	Terms defined elsewhere	1
3.2	Terms defined in this Recommendation	1
4	Abbreviations and acronyms	2
5	Conventions	2
6	Overview of a deep neural network processor benchmark	3
6.1	Evaluation object	3
6.2	Evaluation principle	3
7	Architecture framework of the deep neural network processor benchmark	3
7.1	Workload	4
7.2	AI framework	4
7.3	Hardware acceleration SDK	4
7.4	Processor hardware system	4
8	Benchmark metrics and evaluation methods for AI chip under training tasks	5
8.1	Metrics	5
8.2	Benchmark specification	5
9	Benchmark metrics and evaluation methods for AI chip under inference tasks	5
9.1	Metrics	5
9.2	Benchmark specification	6
10	Benchmark application scenarios	7
Appendix I – Reference implementations of benchmark application scenarios		8

Recommendation ITU-T F.748.11

Metrics and evaluation methods for a deep neural network processor benchmark

1 Scope

This Recommendation provides the benchmarking framework, evaluation metrics and methods, and application scenarios for operating training and inference tasks.

It addresses the following subjects:

a) Evaluation method of a deep neural network processor benchmark;

b) Architecture framework of the deep neural network processor benchmark;

c) Metrics of the deep neural network processor benchmark;

d) Benchmark application scenarios of the deep neural network processor benchmark;

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.3172] Recommendation ITU-T Y.3172 (2019), *Architectural framework for machine learning in future networks including IMT-2020*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following term defined elsewhere:

3.1.1 machine learning (ML) [ITU-T Y.3172]: Processes that enable a computational system to understand data and gain knowledge from it without necessarily being explicitly programmed.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 benchmark: Benchmark is an evaluation method with a long-term application in the entire computer field. Example: As computer architecture advanced, it became more difficult to compare the performance of various computer systems simply by looking at their specifications. Therefore, tests were developed that allowed comparison of different architectures (i.e., providing benchmarks).

3.2.2 deep neural network processor: A deep neural network processor is the kind of processor that could accelerate the computing process regarding a deep neural network.

3.2.3 workload: Workload is the benchmark test task for measuring and evaluating artificial intelligence chips, that is, the benchmark test program containing a specific AI task.

3.2.4 test data set: A subset of the data set which is used to test the model after the model that has been initially verified by the validation data set.

3.2.5 model operations: Model operations are the theoretical amount of operations in deep neural networks. For float operations, the unit is the floating point of operations (FLOPs), and for integer operations, the unit is operations (OPs).

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

AI	Artificial Intelligence
API	Application Programming Interface
ASIC	Application-Specific Integrated Circuit
BLEU	Bilingual Evaluation Understudy
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DNN	Deep neural network
DSP	Digital Signal Processing
DUT	Device under test
FLOPs	Floating Point of Operations
FPGA	Field Programmable Gate Array
FPS	Frame Per Second
GPU	Graphic Processor Unit
MAC	Multiply and Accumulate
mAP	Mean Average Precision
NMS	Non-Maximum Suppression
NPU	Neural Processing Unit
OPs	Operations
PSNR	Peak Signal to Noise Ratio
ROC	Receiver Operating Characteristic
SDK	Software Development Kit

5 Conventions

The following conventions are used in this Recommendation:

– The keywords "is required to" indicate a requirement which must be strictly followed and from which no deviation is permitted, if conformance to this Recommendation is to be claimed.

– The keywords "is recommended" indicate a requirement which is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

6 Overview of a deep neural network processor benchmark

6.1 Evaluation object

The evaluation object can be a processor for training and inference tasks at the cloud side, such as an acceleration card, an all-in-one, etc. Or a processor for inference tasks at the mobile side, such as a development board, a mobile phone, and a dedicated chip for applications such as security, autonomous driving, and smart speakers.

6.2 Evaluation principle

6.2.1 Practicality

Benchmarking methods should be able to produce positive results.

6.2.2 Fairness

Benchmarking methods should provide a fair comparison by specifying rules and metrics and be based on objective scientific test data during the evaluation process.

6.2.3 Reproducibility

When the same detected quantity is detected in different detection environments, the consistency of the measurement results is ensured.

6.2.4 Evaluation mechanism

The benchmarking program focuses on the performance of the processor when operating different training/inference tasks using a complete neural network under specified metrics.

6.2.5 Test environment

The benchmark test environment consists of three parts as shown in Figure 6-1, which are respectively the test computer, power consumption monitoring and the device under test (DUT). The test computer is connected through the DUT interface. The power software test tool analyses the DUT's feedback's current and voltage results. Benchmark tools calculate the relevant evaluation metrics.

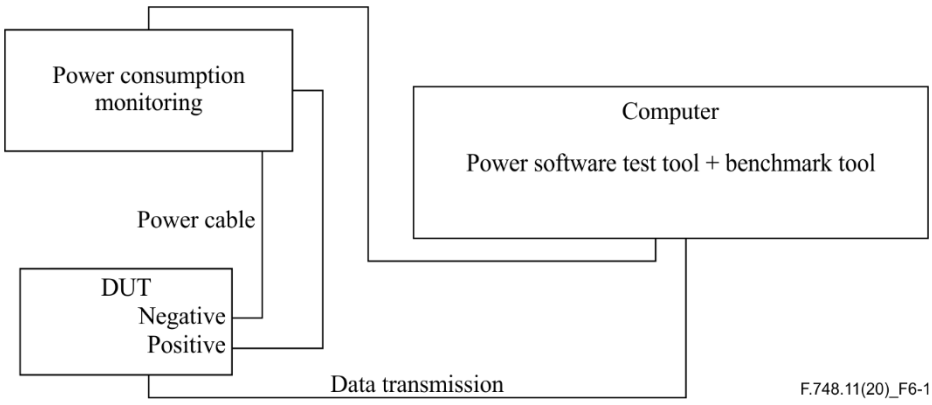

Figure 6-1 – Benchmark test environment

7 Architecture framework of the deep neural network processor benchmark

The benchmarking framework is divided into three layers, as shown in Figure 7-1. These are respectively from the bottom to the top: 1) processor hardware system, 2) hardware acceleration software development kit (SDK), 3) artificial intelligence (AI) framework.

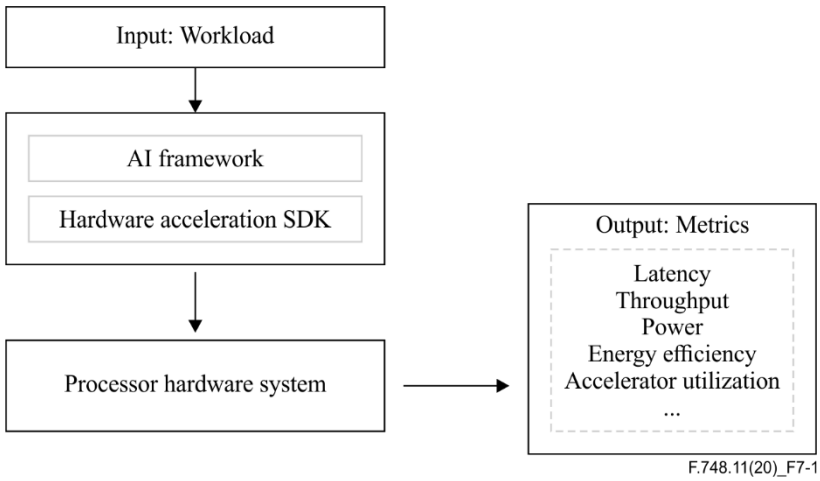

Figure 7-1 – Architecture framework of the deep neural network processor benchmark

As shown in Figure 7-2, according to the deployment location, a benchmark can be divided into cloud side and edge/mobile side and according to the task, it can be divided into training phase and inference phase.

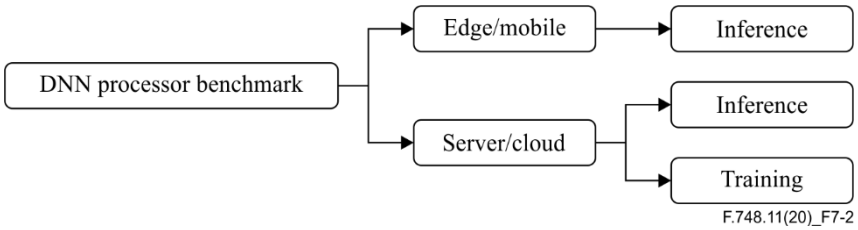

Figure 7-2 – Classification of deep neural network processor benchmark

7.1 Workload

Workload consists of test cases that use neural network models to complete training or inference tasks in specified test scenarios.

7.2 AI framework

The AI framework is a software framework for implementing relevant artificial intelligence models. It includes training frameworks such as TensorFlow, PyTorch, Caffe, PaddlePaddle and Theano, and inference frameworks such as MACE, SNPE, NCNN, TensorFlow Lite, Caffe2go, MDL, Core ML, and others. These frameworks allow us to build deep learning models to complete training or inference tasks.

7.3 Hardware acceleration SDK

The hardware acceleration SDK can support the AI framework upwards, and provide the central processing unit (CPU), graphic processor unit (GPU), digital signal processing (DSP), and neural processing unit (NPU) with required hardware resources for scheduling and use.

7.4 Processor hardware system

The processor hardware system can contain physical devices for artificial intelligence computing, including chips such as CPU, GPU, application-specific integrated circuit (ASIC), field programmable gate array (FPGA), etc., as well as computing servers and mobile terminal devices.

8 Benchmark metrics and evaluation methods for AI chip under training tasks

8.1 Metrics

8.1.1 Training time

The training time is generally defined as the time for training a model on a specific data set to achieve the target accuracy, excluding pre-processing and model loading time.

8.1.2 Accuracy

Accuracy here refers to the accuracy related to the training set. For specific application scenarios and data sets, given the network structure and hyper-parameter configuration, the accuracy metric is used to check whether the training can reach the specified accuracy after a certain number of iterations (training time).

8.2 Benchmark specification

8.2.1 Test data set

The test data set discloses the proportion of the training set and the test data set and the data pre-processing method.

8.2.2 Model

It is required to unify the hyper-parameters, weights' initialization and the format of weights (fp64, fp32, fp16).

8.2.3 Reference implementation

Table 8-1 provides test scenarios, test data sets, target accuracy, and reference models for training benchmarks. All these scenarios are recommended while provided with a training benchmark.

Table 8-1 – Test scenarios for training

Test scenarios	Test data set	Target accuracy	Model
10.1	ImageNet (224x224)	74.9% Top-1 Accuracy	ResNet_50
10.2	COCO 2017	23% mAP	SSD
10.2	COCO 2017	0.377 Box min AP, 0.339 Mask min AP	Mask R-CNN
10.6	WMT English-German	25.0 BLEU	Transformer

9 Benchmark metrics and evaluation methods for AI chip under inference tasks

9.1 Metrics

9.1.1 Inference latency

Inference latency refers to the time that accelerator/processor completes the inference with respect to one sample for a given application scenario with a specific batch size, excluding pre-process and post-process time.

9.1.2 Throughput

Throughput is the number of input data samples that is being processed per second. For an image processing task, the unit of throughput is frames per second (FPS).

9.1.3 Power consumption

Power consumption refers to the average power consumption of the DUT in the process of inference.

9.1.4 Energy efficiency

Energy efficiency refers to the ratio of the throughput with a specific network to the power consumption of the task performed in a given application scenario.

9.1.5 Accelerator utilization

Accelerator utilization refers to the ratio of the actual amount of model operations to the theoretical peak performance under the given application scenario.

Accelerator utilization = (throughput * model operations) / theoretical peak performance.

Taking the convolutional neural network (CNN) as an example, model operations= MAC∗2.

9.1.6 Accuracy

The accuracy of the model inference is related to specific tasks and evaluation metrics. See clause 10 for details.

9.2 Benchmark specification

9.2.1 Test system environment

According to the operating system the test system environment is an environment supported by the device under test (DUT).

9.2.2 Acceleration SDK

The acceleration SDK provides the neural network acceleration engine and its detailed description. It provides interface functions including initialization Init (), pre-processing PreProcess (), loading model LoadModel (), running Run (), unloading model UnloadModel () and post processing PostProcess ().

9.2.3 Model information

Provides relevant information including model categories, original accuracy, data set, input and output node names, the mean of pre-processing and normalized parameters, tensor information (input and output), channel information (RGB / BGR), and data format (NHWC, etc.).

9.2.4 Reference implementation

Table 9-1 provides test scenarios, test data sets, target accuracy, and reference models for an inference benchmark. All these scenarios are recommended while doing a training benchmark.

Table 9-1– Test scenarios for inference

Test scenarios	Test data	Model
10.1	ImageNet	ResNet_50, ResNet_101, Inception_v4, MobileNet_v1, MobileNet_v2
10.2	COCO 2017	Faster-RCNN (with FPN), Yolo V3, SSD
10.3	VOC2012	Deeplabv3+
10.4	2017CVPR	VDSR
10.5	LFW	DeepID3

10 Benchmark application scenarios

Deep learning technologies are most commonly applied within vision, language/audio, and commerce. The application scenario indicates specific tasks developed based on these three application areas. For vision, the application scenarios include classification, recognition, and detection, semantic segmentation, etc. For language and audio, application scenarios could be machine translation and speech recognition. For commerce, the recommendation is the most well-known application scenario. Table 10-1 shows the typical application scenarios with related datasets and reference implementations, which are recommended when conducting benchmark test.

Table 10-1 – Benchmark application scenarios

Area	No.	Application scenarios	Dataset	Reference implementation model	Accuracy
Vision	10.1	Image classification	ImageNet	ResNet/Inception/ MobileNet	Prediction accuracy
	10.2	Object detection	COCO2017/ VOC2012	SSD/Mask R-CNN/YOLO	mAP, NMS
	10.3	Segmentation	VOC2012	Deeplabv3+/FCN	mIoU
	10.4	Super-Resolution	DIV2K	VGG19/VDSR	PSNR,SSIM
	10.5	Face recognition	LFW	VGG-16-Face CNN (224x224)	Rank1,ROC
Language/ Audio	10.6	Translation	WMT Eng-Germ	Transformer	BLEU
	10.7	Speech recognition	LibriSpeech	Deep Speech 2	WER Perplexity
Commerce	10.8	Recommendation	MovieLens-20M	NCF	Prediction accuracy

Appendix I

Reference implementations of benchmark application scenarios

Table I.1 gives a reference implementation of benchmark application scenarios.

Table I.1 – Benchmark test use cases

No.	Application scenarios	Dataset	Source	Model (input size)	Source
10.1	Image classification	ImageNet	http://www.image-net.org/download-images	ResNet_50 (224x224)	https://github.com/KaimingHe/deep-residual-networks
				ResNet-101 (224x224)	https://github.com/KaimingHe/deep-residual-networks
				Inception-v3 (299x299)	https://github.com/tensorflow/models/tree/master/research/slim
				MobileNet-v2 (224x224)	https://github.com/shicai/MobileNet-Caffe
10.2	Object detection	VOC2012	http://host.robots.ox.ac.uk/pascal/VOC/	ssd_mobilenetv1 (300x300)	https://github.com/FreeApe/VGG-or-MobileNet-SSD
				YOLOv3 (320x320)	https://github.com/eric612/MobileNet-YOLO
10.3	Segmentation	VOC2012	http://host.robots.ox.ac.uk/pascal/VOC/	FCN (500x500)	https://github.com/shelhamer/fcn.berkeleyvision.org/tree/master/voc-fcn8s
10.4	Super-Resolution	DIV2K	https://data.vision.ee.ethz.ch/cvl/DIV2K/	VDSR (256x256)	https://github.com/huangzehao/caffe-vdsr
10.5	Face recognition	LFW	http://viswww.cs.umass.edu/lfw/	VGG-16-Face CNN (224x224)	http://www.robots.ox.ac.uk/~vgg/software/vgg_face/

Table I.1 – Benchmark test use cases

No.	Application scenarios	Dataset	Source	Model (input size)	Source
10.6	Translation	WMT Eng-Germ	http://www.statmt.org	Transformer	https://github.com/Kyubyong/transformer
10.7	Speech recognition	Libri Speech	http://www.openslr.org/12/	Deep Speech 2	https://github.com/yao-matrix/deepSpeech2
10.8	Recommendation	Movie Lens-20M	https://grouplens.org/datasets/movielens/	NCF	https://github.com/guoyang9/NCF

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	**Non-telephone telecommunication services**
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2020