智能驾驶人工智能芯片的计算性能测试研究

赵 瑞[1] 夏显召[1] 吴海文[2*] 李予佳[1] 王清扬[1] 翟瑞卿[1] 李明阳[1] 郝晶晶[3]

〔1.中汽研软件测评（天津）有限公司；2.国家市场监督管理总局认证认可技术研究中心；
3.中国汽车技术研究中心有限公司〕

摘 要：智能驾驶汽车发展迅速，对汽车人工智能（Artificial Intelligence，AI）芯片的计算性能要求持续提升，也引发了行业对于验证芯片计算性能是否满足实际应用需求的关注。然而，目前现有研究多为通用性计算性能测评方法，针对汽车领域应用的AI芯片计算性能有待研究。本文基于这一行业问题，对测试模型进行调研，提出了一套面向智能驾驶应用的AI芯片计算性能测试方法，并选取典型模型对实际产品进行测试，对比分析不同芯片各模型下的测试结果。本文研究内容对于汽车行业进行AI芯片选型也具备重要的参考价值。

关键词：智能驾驶，人工智能芯片，测试评价，计算性能，芯片选型

DOI编码：10.3969/j.issn.1674–5698.2024.10.011

Research on Computing Performance Testing of Artificial Intelligence Chip for Intelligent Driving Vehicles

ZHAO Rui[1] XIA Xian-zhao[1] WU Hai-wen[2*] LI Yu-jia[1] WANG Qing-yang[1]
ZHAI Rui-qing[1] LI Ming-yang[1] HAO Jing-jing[3]

〔1. CATARC Software Testing (Tianjin) Co., Ltd.; 2. China Certification and Accreditation Institute;
3. China Automotive Technology & Research Center Co. Ltd.〕

Abstract: The rapid development of intelligent driving vehicles has continued to raise the computational performance demands for automotive artificial intelligence (AI) chips, and has triggered the industry's concern about verifying whether the computing performance of chips meets the needs of practical applications. However, most existing researches focus on general computing performance measurement methods, and the computing performance of AI chips for automotive applications still needs to be researched. Based on this industrial problem, this paper conducts survey on test models, proposes a set of testing methods of AI chip computing performance for intelligent driving applications, and selects typical models to test different products, and analyzes the test results of different chips under each model. The research has important reference value for AI chip selection in the automotive industry.

Keywords: intelligent driving, artificial intelligence chip, testing and evaluation, computing performance, chip selection

---

基金项目：本文受市场监管总局技术保障专项“汽车芯片安全可靠性认证技术及通用审查方法研究”（项目编号：2023YJ38）资助。

作者简介：赵瑞，硕士研究生，主要从事汽车芯片检测技术研发，牵头及参与制定10余项汽车芯片行业标准，参与国家重点研发计划、省部级重大科研计划等课题。

吴海文，通信作者，博士研究生/博士后，正高级工程师，从事合格评定政策与技术研究。

0　引言

近年来，随着我国智能汽车产业发展加速，智能驾驶、智能座舱系统的装车率逐渐提升，计算芯片已成为支撑汽车智能化的关键部件。针对不同应用场景、不同智能化等级，如何选取与之匹配的计算芯片成为行业关注的热点问题，但现阶段计算芯片的性能测评并未实现行业普遍共识。当前尚未有研究可为行业提供标准化、公开、客观的汽车人工智能（Artificial Intelligence，AI）计算芯片性能测试方法。

为有效探究并解决汽车AI芯片的计算能力测试问题，本文以应用于智能驾驶系统的汽车AI芯片为主要研究对象，对AI芯片测评的研究现状开展分析，并面向智能驾驶实际应用，提出了一套普适性较高的汽车AI芯片的计算能力测试方法。本文使用该方法对芯片产品进行测试验证，并对不同芯片的测试结果进行对比分析。该方法的应用可为企业选择智能驾驶汽车AI芯片提供性能方面的参考，解决当前汽车AI芯片缺失计算能力测试评价方法的问题，具有重要的研究意义和实际应用意义。

1　智能驾驶AI芯片

1.1　AI芯片发展现状

智能驾驶是AI芯片应用的典型代表。算法、计算能力和大数据是推动智能驾驶汽车崛起的三大要素。这三者必须平衡完美发展，智能驾驶汽车才可能取得良好的发展前景。计算能力是AI的基础，也是智能驾驶复杂数据处理的关键。近年来，由于智能汽车产业的发展，汽车需要处理的数据量呈现爆发式增长，传统的计算架构越来越难以支撑深度学习的海量并行计算需求。因此，AI芯片的技术研发成为研究热门。应用研究方面，国外巨头如：NVIDIA、Google、IBM等国际巨头推出新品，国内地平线、华为、黑芝麻等企业也纷纷布局汽车AI芯片产业，中国AI芯片技术取得了重大的发展[1]。

按照设计架构，AI芯片主要分为GPU、FPGA、ASIC，当前市场上主流应用的AI芯片是GPU。就适用范围而言，GPU为通用型芯片，ASIC为专用型芯片，而FPGA是属于两者之间的半定制化类芯片。综合来看，3种AI芯片各有优劣，GPU运算速率快，通用性较强，开发难度相对较低，预计在目前及未来一段时间都将占据主流地位；ASIC的用量有限，可能难以形成规模化应用；FPGA的量产成本高，与GPU相比开发门槛高。因此目前ASIC与FPGA在AI芯片市场的占比皆不高[2]。

1.2　AI芯片算力

算力是特定场景下对芯片计算能力评价的重要维度。算力大小代表芯片数字化信息处理能力的强弱。自动驾驶场景需要标量、矢量、矩阵3者结合的异构算力，通常可以将算力的综合评价分为两方面，即AI算力和CPU算力。AI算力是AI处理器在特定场景下提供的矢量和矩阵计算能力，也是智能驾驶领域热点的研究方向。AI 算力常用的单位是TOPS（Tera Operations Per Second）或TFLOPS（Tera Floating-point operations per second），1TOPS代表 AI处理器每秒可进行一万亿次（10^{12}）定点操作，1TFLOPS 分别代表 AI 处理器每秒可进行一万亿次（10^{12}）浮点操作。CPU算力是CPU主要提供的标量算力。CPU算力常用的单位是 DMIPS（Dhrystone Million Instructions executed Per Second），其含义为每秒钟执行基准测试程序 Dhrystone 的次数除以1757[2]。

智能驾驶技术的发展极大地提升了其对于芯片算力增长的需求。据统计，当前L2、L3级别自动驾驶计算量已分别达到10TOPS和60TOPS，预计L4级别算力可能会超过100TOPS[3]。大算力的AI芯片可支撑自动驾驶汽车海量的代码运算，为自动驾驶的发展提供保障。然而，在智能驾驶汽车实际应用场景下，AI芯片的最大计算能力并不能达到理论算力值，无法单纯通过产品宣称的理论算力判断不同产品的真实计算性能。因此，建立能够有效地反映汽车AI芯片计算能力的测试指标，并通过实际测试体现计算能力，是具有重大意义的研究工作。

2 AI芯片计算能力测试研究现状

2.1 测试基准

AI芯片的计算性能测评通过基准测试程序实现，当前国内外对于通用的AI芯片性能测试方法已有一定的研究成果和实际应用。AI芯片的性能测试主要依靠使用基准测试集，运行所需神经网络来进行。当前常用的AI芯片基准测试集包括AI benchmark、MLPerf、AIIA DNN benchmark等。此外，众多研究机构开发了面向不同维度的基准测试集，如：Fathom（哈佛大学）、DeepSpeech（百度）、NPUbench（中国科学院）以及AIPerf（清华大学）等，可实现特质化的测试功能。AI benchmark[4]是瑞士苏黎世联邦理工学院开发的专门用于评估AI芯片性能的基准测试集，涵盖了多方面的AI性能，包括计算速度测试；MLPerf[5]是用于测量和提升机器学习软硬件性能的通用基准，包括各个领域的子项，如：图像分类、识别、翻译、语音识别等，测量不同神经网络训练和推理所需的时间和速度；AIIA DNN benchmark[6]是由中国AI产业发展联盟开发的基准测试集，综合5大维度评估AI芯片性能，并根据算力单价比和芯片利用率，反映加速卡性价比与软硬件及存储系统的整体能力。

2.2 测评标准

当前，国内已有多个AI芯片测评标准完成指定并发布。中国信息通信研究院起草的行业标准《人工智能芯片基准测试评估方法》[7]于2021年8月发布，标准里规定了AI芯片计算性能基准测试框架、评测指标及评估方法，主要包括基本信息披露和技术测试；中国电子技术标准化研究院起草的团体标准《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》[8]《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》[9]《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》[10]均于2020年10月发布，3项标准分别规定了对云侧、边缘侧、端侧深度学习芯片进行计算性能测试的测试指标、测试方法和要求。

2.3 存在的问题

虽然国内外已经形成多项AI芯片计算性能测试基准或测试标准，但在汽车智能驾驶领域，这些基准或标准并不能完全适用。汽车企业无法通过适用的测试方案验证不同产品的计算性能表现。因此，建立更加适用于智能驾驶领域的测试方案，形成直观、清晰、可对比的测试结果对于AI芯片选型参考具有重要的意义。

3 测试方法论

基于上述研究基础及存在的问题，本文提出了一套测试方法论，包含测试模型选取和测试方案实施要求两部分内容。算法模型和AI芯片都是智能驾驶不同应用场景的运算基础，本文将综合研究AI芯片搭载算法模型的计算性能表现。

3.1 测试模型选取

3.1.1 选取原则

为保证测试的一致性并形成具有可比性的测试结果，测试模型需基于以下3方面进行选取。

（1）测试模型为公版、开源模型，其来源方为行业广泛共识且权威性较高的科学机构，获取渠道为原始发布渠道或者具备行业普遍共识的官方渠道，以保证测试过程中使用的测试模型具备一致性，避免不同获取渠道导致的模型信息差异；

（2）同一测试公版模型的版本一致，若测试模型经过了后处理，则需要明确处理方式和目的，保证同一测试公版模型的一致性和稳定性；

（3）模型需基于汽车计算芯片常用应用场景选取，以提升测试结果的实际应用意义。

3.1.2 选择过程

本文选择的模型经面向行业广泛调研后得出，调研对象包括整车企业、零部件企业、算法企业、芯片企业及测试机构等29家单位。调研内容包含汽车计算芯片常用应用场景、关注的性能指标、当前常用的模型列表及数据集列表，主要调研结果如表1所示。根据调研结果，选取企业选择率超过50%的模型作为测试模型候选列表。

3.2 测试方案

3.2.1 性能指标

表1　模型选取调研结果

模型类型	模型名称	平均关注度	芯片企业关注度	零部件企业关注度	整车企业关注度
图像分类	ResNet18	64%	62.5%	83.3%	64%
	ResNet50	64%	87.5%	33.3%	73%
图像检测	Yolo v3_Darknet53	56%	75%	50%	45%
	Yolo v5m	72%	87.5%	50%	82%
	Yolo v5x	64%	75%	50%	63%
	SSD_MobileNetV1	56%	75%	66.6%	36%
	DETR	56%	62.5%	66.6%	36%
	PointPillars	60%	64.5	83.3%	55%
分割	FCN	64%	75%	66.6%	55%
单目深度估计	MonoDepth	56%	75%	33.3%	55%

根据智能驾驶应用需求，分析计算芯片应用场景，确定对应的性能指标和测试方法。经分析，算法模型作为支撑计算芯片完成计算任务的关键，基于不同应用场景，需适配不同的算法模型。芯片支持算法模型的数量决定计算芯片是否可适配多种任务类型，各算法模型对应的计算速度决定是否可以满足大量数据的任务处理需求，这里以帧率（每秒处理的图像数量）作为AI芯片计算能力体现的关键指标。

3.2.2 测试方案

AI芯片的使用既需要芯片硬件性能作为基础，也需要软件为芯片基于应用场景的开发提供支撑，因此本文制定了工具链、芯片及算法模型这种软硬结合的测试方案。计算芯片适配算法模型需要综合考虑工具链的模型转换能力和计算芯片硬件的计算能力，因此，本测试方案将计算芯片与其配套工具链作为整体进行评估，充分反应计算芯片的软硬件结合综合性能。本测试方案依托算法模型，打造更贴近于实际应用的性能指标和测试方法，如图1所示。

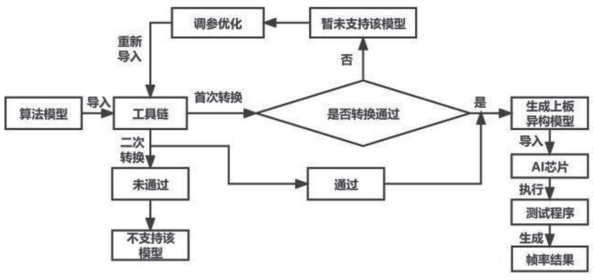

图1　测试方案示意图

3.2.3 测试项目说明

本文测试项目包括模型兼容性测试和帧率及时延测试。

模型兼容性测试：通过芯片和工具链是否支持batch 1的公版模型转换，判断工具链的公版模型覆盖度，即芯片产品对于算法模型的兼容性，记录转换通过和不通过的模型数量，根据通过模型数量和测试模型总数量的比值大小判断产品的兼容性高低程度。

帧率及时延测试：在进行帧率测试时根据测试要求对应设置不同batch参数，根据不同模型大小设置size参数值，生成板上可运行的异构模型，按照测试要求设置测试参数如：线程数，优化模式等，调用测试程序执行测试。Batch值是一次训练所选取的样本数，同时可以反映芯片和工具链的能力，batch数值的不同会影响帧率，一般会将batch数值设置为batch1、batch2、batch4、batch8等多种参数，记录不同batch测试结果，利用batch1的帧率值取倒数计算时延，并将最佳测试结果对应的测试设置记录，作为该产品最优性能的详细体现。

4 测试结果

依据测试方案，通过摸底实验已经取得国内外多款芯片测试结果，包含国内外的4款智能驾驶汽车AI芯片产品，以A、B、C、D代表，如图2所示。

通过测试，本文也得出一些结论，首先是根据部分模型帧率测试结果可以看出，不同产品宣

称算力与实际计算测试结果并无正相关，如图2中C产品算力与A相比差很多，但图3中C产品的模型EfficientNet-lite0和MobileNetV2测试结果却比A更好，因此可以得出，基于模型的帧率测试可以更真实地反映产品的计算能力；其次，根据测试结果也可以看出，不同产品的模型覆盖数量即产品的兼容性存在差异，单一产品不同输入样本数量（batch）测试结果也存在较大差异，分别如图4和5所示。

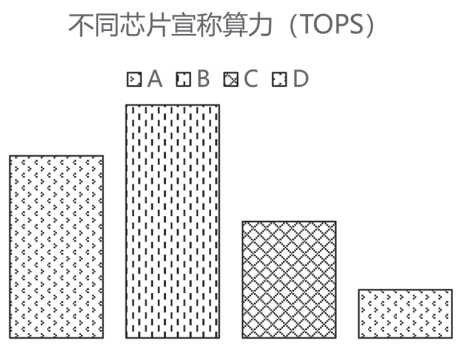

图2 不同产品宣称算力结果图

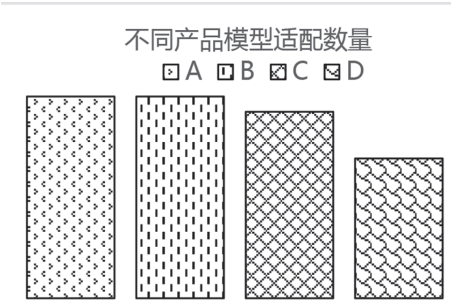

图4 不同产品模型适配数量结果对比图

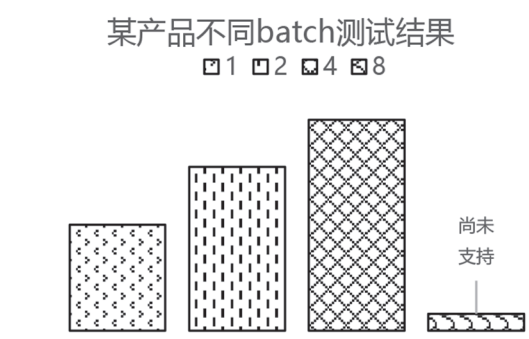

图5 某产品不同batch测试结果对比图

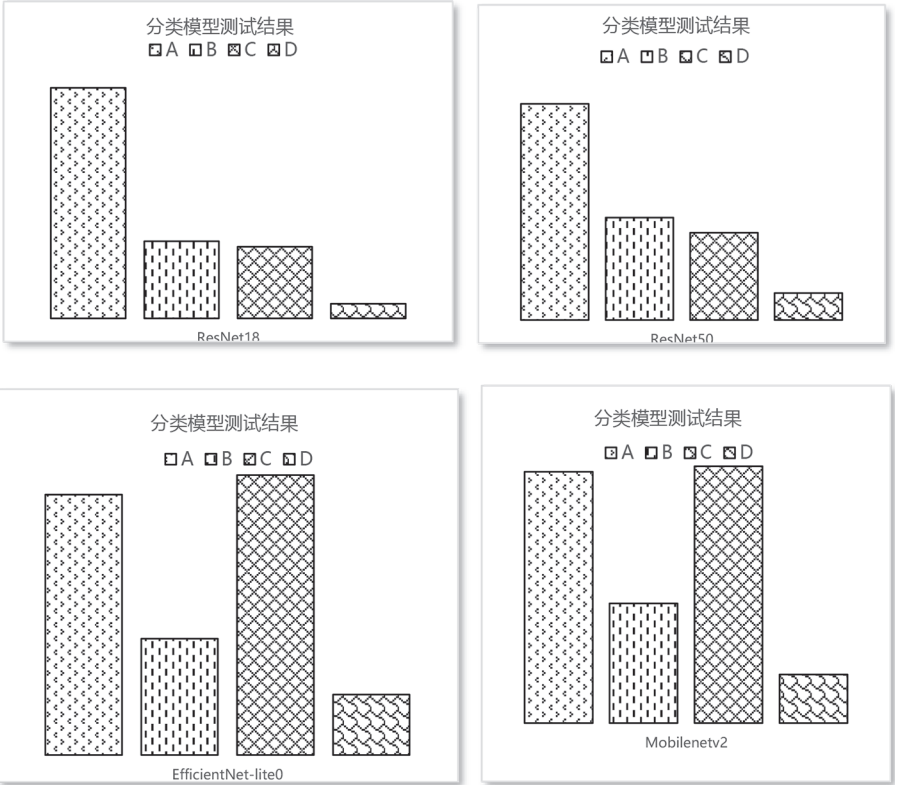

图3 不同产品测试结果对比图

在测试过程中，除了对模型转换batch参数、模型size等参数进行规定设置，还通过设置不同线程、不同计算核、不同优化模式等参数实现对不同产品的优化测试，获得最优性能，如图6所示。将最优性能与规范测试性能对比，既可展示企业产品的综合性能，又可展现产品计算能力的优化空间水平。通过测试也得出，不同公版模型的渠道、版本等对测试结果影响较大。因此，测试的开展应保证输入模型属性的一致性，保证测试结果的对比性。

根据测试结果，我们形成了同一模型、统一测试条件下不同产品的帧率结果“天梯图”，如图7所示。

该图直观呈现不同模型下各产品性能的排名情况，为汽车行业相关企业提供清晰的选型参考。

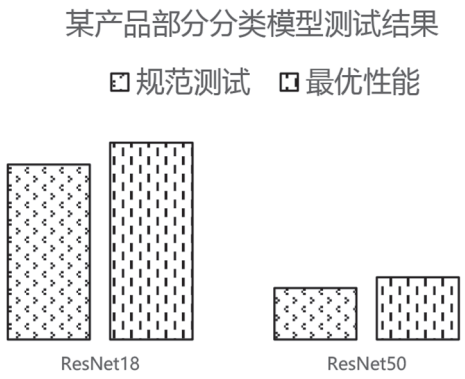

图6　规范测试和优化测试产品帧率结果图

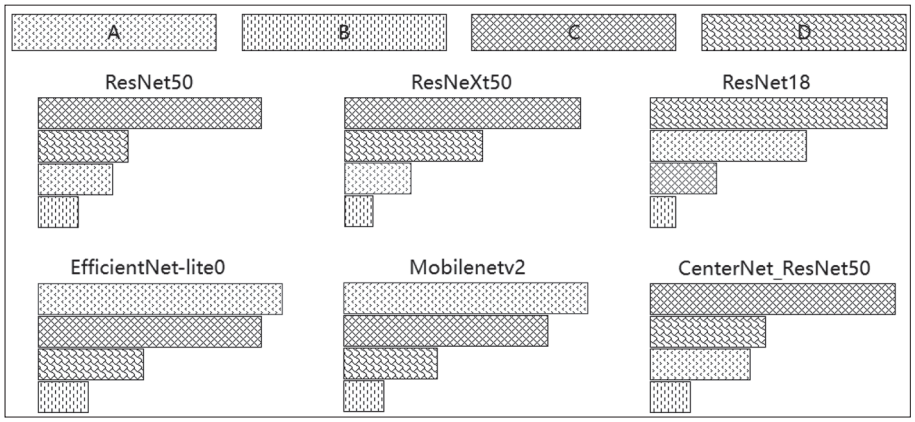

图7　产品计算性能天梯图

5　结 论

本文基于对现有通用AI芯片性能测试的研究成果分析，提出了适用于智能网联汽车领域，针对智能驾驶应用的AI计算芯片性能测试方法。本文基于行业调研结果选取多个性能测试模型，对多款汽车AI芯片进行测试，验证测试方法对不同芯片产品的适用性。对测试结果进行深入分析，测试结果可提供两方面的选型参考，（1）功能性验证，评估工具链支持公版模型的数量和转换能力，评估产品的通用性能力；（2）性能验证，对不同产品的帧率测试结果进行对比验证，评估产品的性能高低。本文提出的测试方法可用于汽车AI芯片计算性能的测试评价，并可以根据应用需求选取不同测试模型进行测试，测试结果具有可比性和一致性，可为企业进行芯片选型提供重要参考。在后续研究中可探索面向智能座舱应用场景的常用测试模型和测试方法。

参考文献

[1] 商惠敏. 人工智能芯片产业技术发展研究[J]. 全球科技经济瞭望, 2021, 36(12): 24–30.

[2] 葛悦涛,任彦. 2020 年人工智能芯片技术发展综述[J]. 无人系统技术, 2021, 4(12):14–19.

[3] 张天雷,任秉韬,郑思仪,等. 中国智能驾驶2107发展报告[M]. 北京：电子工业出版社，2018.

[4] IGNATOV A, TIMOFTE T, KULIK A, et al. AI benchmark: All about deep learning on smartphones in 2019 [EB/OL]. [2022–07–08]. https://arxiv.org/pdf/1910. 06663.pdf.

[5] GitHub Copilot X. MLPerf[EB/OL]. [2022–07–08]. https:// github.com/mlperf.

[6] 国际电信联盟第十六研究组. Metrics and evaluation methods for a deep neural network processor benchmark: ITU–T F.748.11[S]. Switzerland: International Telecommunication Union, 2020.

[7] 中国信息通信研究院. YD/T 3944—2021,人工智能芯片基准测试评估方法[S]. 北京: 中国电子工业标准化技术协会, 2021.

[8] 中国电子技术标准化研究院. 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[S]. 北京:中国电子工业标准化技术协会, 2020.

[9] 中国电子技术标准化研究院. 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[S]. 北京: 中国电子工业标准化技术协会, 2020.

[10] 中国电子技术标准化研究院. 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[S]. 北京: 中国电子工业标准化技术协会, 2020.