团体标准《深度学习推理引擎应用编程接口（API）规范》（征求意见稿）编制说明

**一、工作简况**

《深度学习推理引擎应用编程接口（API）规范》规定了深度学习推理引擎应用编程接口规范，定义了模型管理接口、设备管理接口、输入输出接口、推理接口四类接口规范。本标准适用于部署在设备上的深度学习推理引擎，包括云侧（IDC）推理引擎及端侧（移动端、边缘端）推理引擎，可用于指导推理引擎API的设计与开发，个别条款不适用于特殊行业，专业应用。

2020年7月14日，新一代人工智能产业技术创新战略联盟、中关村视听产业技术创新联盟发布公告，《深度学习推理引擎应用编程接口（API）规范》正式立项团体标准，团体标准计划号为2020071401。

“指令集与开发接口专题组”自2019年12月第9次AITISA会议开始，讨论启动《深度学习推理引擎应用编程接口（API）规范》的标准工作，制定的标准应在满足开发者及用户对终端应用多样化需求的同时，解决深度学习应用开发存在的重复开发及各引擎效果不一致的问题。经过专题组对应用需求、技术创新的讨论后，于2020年4月第10次AITISA会议形成《深度学习推理引擎应用编程接口（API）规范》提案征集书（N1184）与标准的第一份草案（N1185）。后专题组通过4次会议讨论（2020年6月线上会议、2020年8月线上会议、2020年12月线上会议、2021年6月海花岛会议），于2021年6月形成了标准第一次组内征求意见稿（N1351）。2021年12月，经过6个月专题组意见征求，标准收到修改建议1项，编制组认真讨论研究了意见后，最终形成了N1424 CD2.0版本。

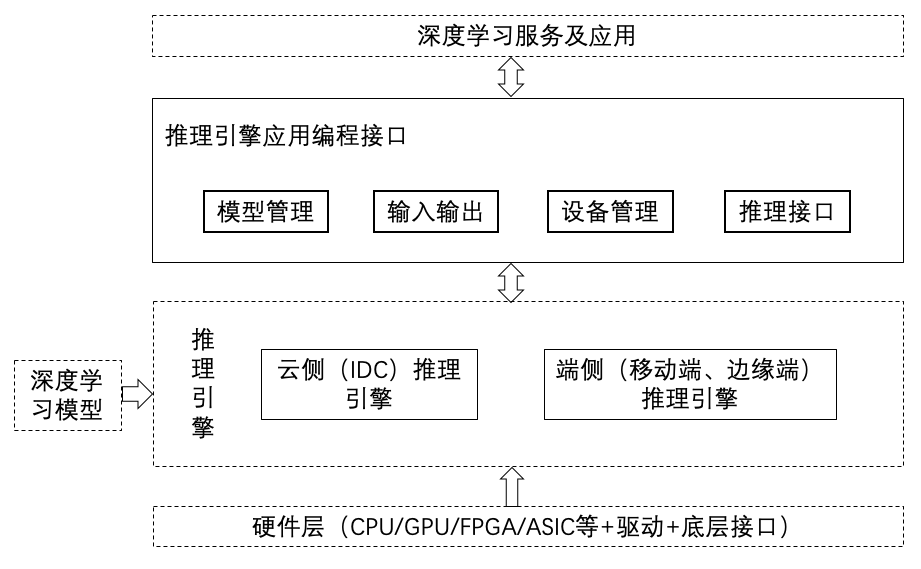
本标准起草单位：北京百度网讯科技有限公司、北京大学、清华大学、北京旷视科技有限公司、上海商汤智能科技有限公司。

本标准起草人： 胡晓光、杨超、翟季冬、梅敬青、蒋慧、严春伟、范睿博、李笑如、李雨芮、马艳军、于佃海、徐浩、蒋晓琳、张艺伯、吴庚

**二、标准编制原则和确定主要内容的论据及解决的主要问题**

深度学习推理引擎是一种专门用于深度学习预测或推理的软件框架，提供一系列高层接口用于加载深度学习模型，搭载并适配硬件执行模型中涉及的所有运算，获得推理结果，可应用于图像识别、自然语言处理等场景。国内外科技厂商推出了一系列的端侧深度学习推理引擎，如飞桨、NCNN、TensorFlow Lite 、MNN、MegEngine 等。伴随着深度学习推理引擎的发布，多种推理引擎API（应用程序编程接口）也开始涌现出来。API类型数量众多，且每种深度学习引擎的API拥有不同的编写特点，不利于应用程序开发者广泛的采用深度学习技术，可能会导致深度学习生态的碎片化。因此，如能屏蔽不同的推理引擎对应用服务开发者的接口差异，制定标准化的开发接口，将会进一步有力地促进深度学习推理技术的良性发展。

为面向深度学习推理引擎提供统一应用编程接口，促使各类推理引擎与应用之间互联互通，提升推理引擎和应用开发者效率，北京百度网讯科技有限公司、北京大学、清华大学、北京旷视科技有限公司、上海商汤智能科技有限公司组成了编制组，进行标准制定。深度学习推理引擎应用编程接口（API）位于推理引擎与应用层之间，上层应用可通过API进行深度学习模型的调用。



深度学习推理引擎应用编程接口主要包括以下部分：

1. 模型管理接口：用于网络接口与参数读取，模型编译优化及自定义算子注册；

2. 设备管理接口：用于设备能耗及线程管理；

3. 输入输出接口：用于输入数据传入及输出数据获取；

4. 推理接口：用于推理实例创建、推理执行以及推理计算。

**三、主要试验[或验证]情况分析**

本文件所涉及的深度学习推理引擎应用编程接口，包括模型管理接口、设备管理接口、输入输出接口、推理接口，广泛被技术类企业实现及应用。

**四、知识产权情况说明**

本文件对潜在的知识产权情况进行了说明。

**五、采用国际标准和国外先进标准情况**

本部分未采用对应国际标准，而是采用自主制定的技术路线。同时，在本标准的制定过程中，积极采纳“指令集与开发接口专题组”会员单位修改建议。

**六、与现行相关法律、法规、规章及相关标准的协调性**

符合我国有关的现行法律、法规。

**七、重大分歧意见的处理经过和依据**

无。

**八、标准性质的建议**

建议发布为推荐性标准。

**九、贯彻标准的要求和措施建议**

建议通过项目资助或技术手段，推广国内科技企业采用标准定义的推理引擎API接口，为加强深度学习推理引擎领域厂商之间的协作，降低应用开发者使用深度学习技术的成本，促进推理引擎生态的良性发展和资源的整合利用。

**十、替代或废止现行相关标准的建议**

无。

**十一、其它应予说明的事项**

无。

团体标准《深度学习推理引擎应用编程接口（API）规范》编制工作组

2021-12-21