白皮书

为您的AI模型选择最佳的存储器

AI如何使用不同形态的存储器

目录

概述

人工智能 (AI) 被公认为是一个非常占用存储的应用。幸运的是，Ambiq的Apollo4 Plus SoC提供了丰富的存储器类型和配置供您选择。决定使用哪种存储器并如何使用它可能需要进行一些实验。因此，我们进行了一些实验并提供了结果给您参考。正如您下面将看到的，有许多好的选项可帮助满足您对AI应用的设计要求。

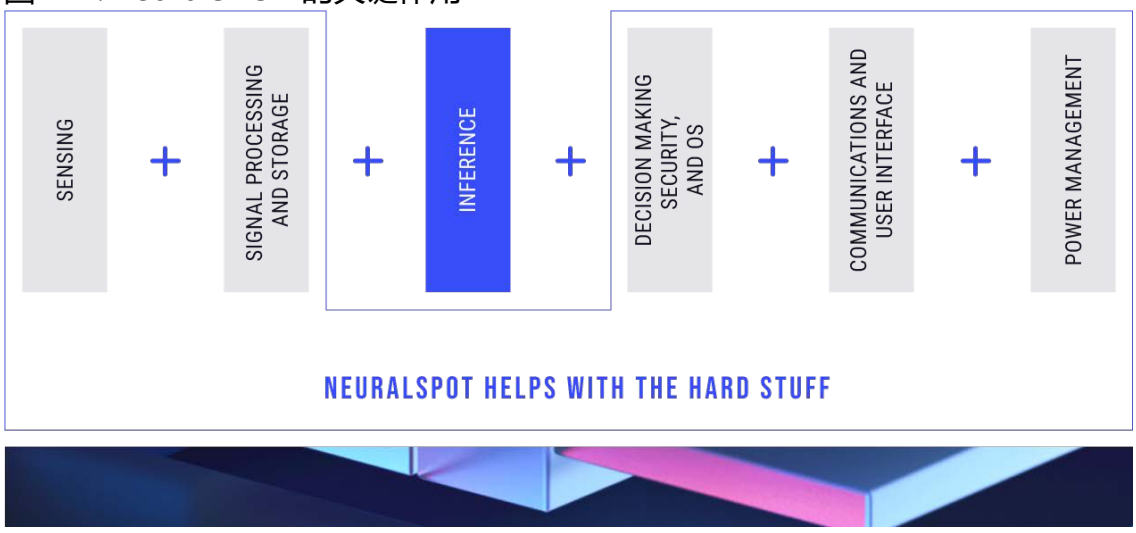
Ambiq AI 简介

Ambiq专注于超低功耗SoC，旨在使由电池供电的智能物联网终端解决方案成为现实。如今，几乎每一个终端设备都融入了AI特性，包括异常检测、语音驱动的用户界面、音频事件检测和分类以及健康监测。Ambiq的超低功耗、高性能平台非常适合实现这类AI特性，我们致力于通过提供以开发者为中心的开源工具包、软件库和参考模型来使产品方案实施尽可能简单，从而加速AI功能的开发。

neuralSPOT® - 简化您的AI部署

neuralSPOT®是一个真正意义上的面向AI开发者的SDK：它包含将AI模型移植到Ambiq平台上所需的一切。在这里您能找到用于与传感器通信、管理SoC外设、控制电源和内存分配的库，用于从您的笔记本电脑或PC轻松调试模型的工具以及将所有这些全部结合起来的示例。

图 1-1: neuralSPOT 的关键作用



要深入了解neuralSPOT®如何成为您的AI开发团队提供强大助力，请参阅我们的白皮书并访问我们的GitHub。

Ambiq ModelZoo

Ambiq的ModelZoo是一个基于neuralSPOT®构建的AI参考模型集合，可以帮助您的团队在Apollo4 Plus上轻松实现AI模型的开发和部署。它包括用于语音接口、语音增强和ECG分析的开源模型，还有复现我们模型的训练结果进而提供训练您自己的模型所需的一切。

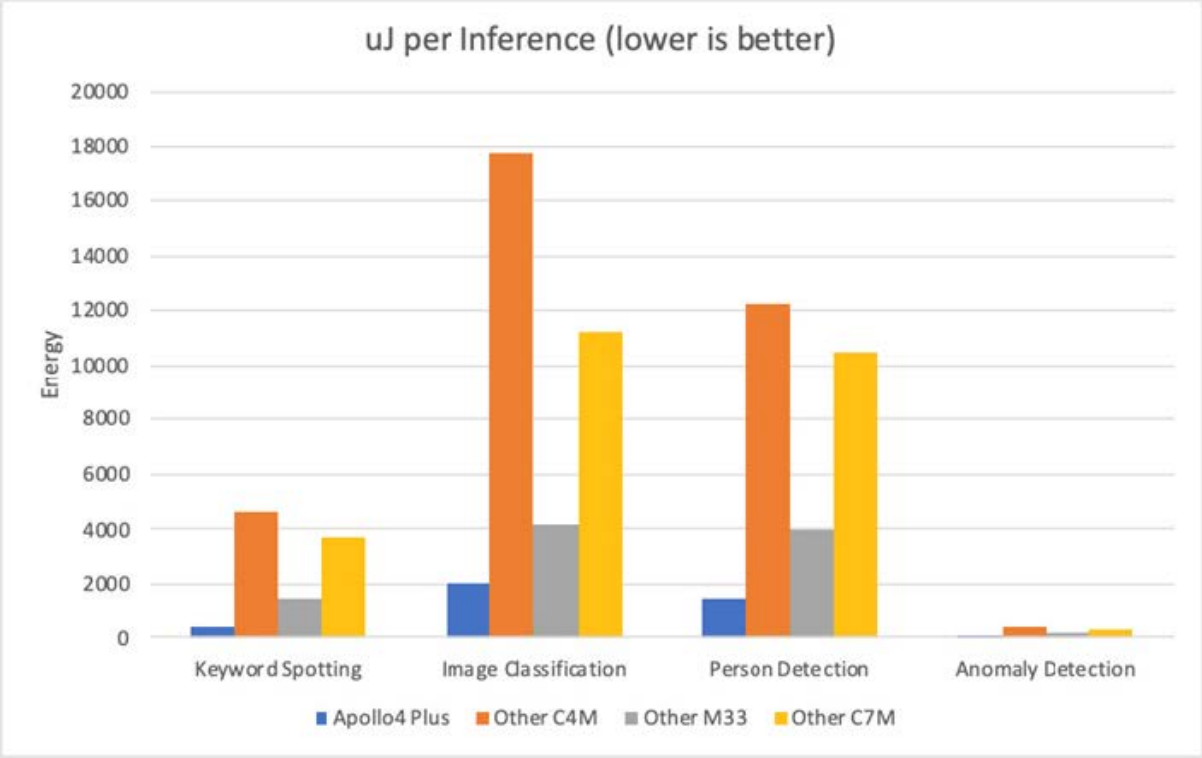
模型 描述

- **NN 语音** - 集合了3种专注于语音的模型：语音活动检测、关键词识别和语意识别。
- **心律失常分类**- 基于单导联心电图传感器检测多种类型的心脏疾病。
- **语音增强** - 基于TinyLSTM的音频模型，可以从语音中去除噪声。

Ambiq AI 性能评估

我们已对Apollo4 Plus平台进行了性能评估，并得出了不错的结果。我们基于MLPerf的评估测试用例可以在我们的benchmark repository中找到，包括有关如何复现我们的结果的说明。

图 1-2: Ambiq Apollo4 Plus 使用MLPerf 进行的性能评估测试



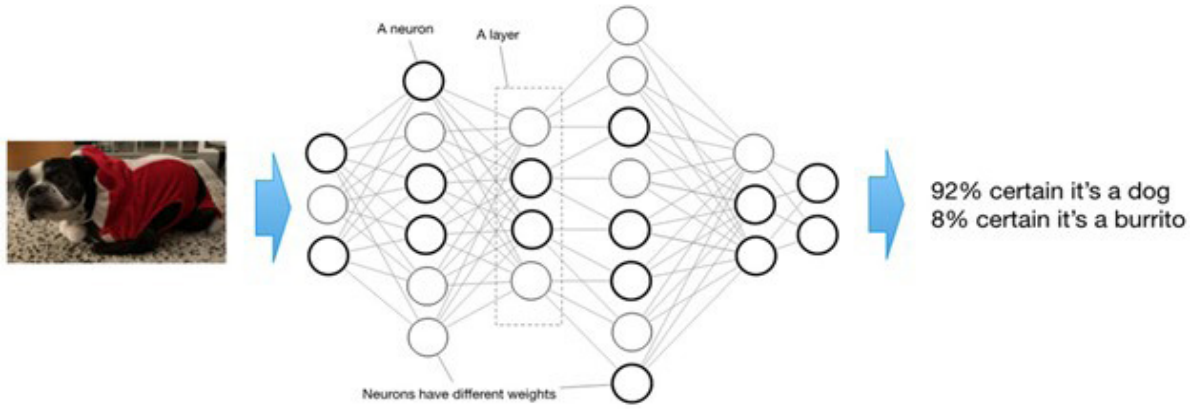
使用Ambiq的AI工具加速您的AI开发

无论您是从零开始创建模型、将模型移植到Ambiq的平台，还是优化您的核心内容，Ambiq都有工具来帮您轻松完成任务。

AI如何使用不同形态的存储器

深度学习AI模型由一系列层组成，每一层都包含许多所谓的“神经元”。单独而言，这些神经元本身都很简单：它们获取输入值并将其乘以与该特定神经元相关的“权重”。具有权重的神经元继续对该组合应用“激活函数”，然后将结果传输到下一层。对于经过训练的模型，这些权重是静态的 - 它们永远不会改变。

图 1-3: 深度学习人工智能模型



虽然这种描述过于简单化了。但是，它确实表明AI模型内存使用由两部分组成：静态部分和动态部分。静态部分代表权重，动态部分由基于这些权重流经神经元的数据组成，这也被称为“激活数据”。我们将利用这些事实来探索如何针对Apollo4 Plus的内存配置优化AI模型。

适用于MCU的TensorFlow[1] Lite是一个运行时解释器，它通过AI模型运行数据，为每次推理执行上述操作数百万次。MCU的内存架构反映了运行AI模型所需的权重和激活数据内存类型。模型的权复位义为模型各层中使用的参数（包括可训练的和不可训练的），存储在模型数组（用于存储和分析的多个模型对象的集合）中。模型的激活数据存储在所谓的“TensorFlow Arena”中。我们可以使用编译器指令控制和指定编译过程放置这些内存对象的具体位置。例如，以下就是我们如何在AmbiqSuite SDK中控制放置的方式：

```
const char weights[] = {0x03, 0x55}; // This will be placed in non-volatile MRAM
char activations[40*1024];           // This will be placed in tightly coupled memory
AM_SHARED_RW char other_activations[40*1024]; // This will be placed in SSRAM
```

[1]TensorFlow、TensorFlow 徽标及任何相关标记均为 Google Inc. 的商标。

Apollo4 Plus Soc的存储器类型

Apollo4 Plus SoC为我们提供了三种可以用于AI的存储器类型：MRAM、紧密耦合内存(TCM)和SSRAM。MRAM是一种高效的非易失性存储器，主要用于存储静态数据。顾名思义，TCM是与CPU紧密耦合的高性能读/写存储器。SSRAM是距离CPU"更远"的通用读/写存储器。使用和访问不同存储器都会对功耗和性能产生不同的影响。

实验

众所周知，TensorFlow的性能很难预测。因此，进行实验是一种更简单的方法。在我们的实验中，我们运行了MLPerf[1] Tiny Inference的关键词识别(KWS)基准测试。我们利用评估测试的复杂系统的来测量性能和功耗，来实证各种内存分配方法的影响。具体来说，我们尝试了以下配置：

表 1-1: 使用 MLPerf 对 Tiny Inference's KWS 评估测试配置

Weights Stored In...	Activations Stored In...	Comments
TCM	TCM	Turn-off SSRAM
TCM	SSRAM	Turn-off half of TCM
MRAM	TCM	Turn-off SSRAM
MRAM	SSRAM	Turn-off half of TCM
SSRAM	TCM	
SSRAM	SSRAM	Turn-off half of TCM

需要注意的几个要点：

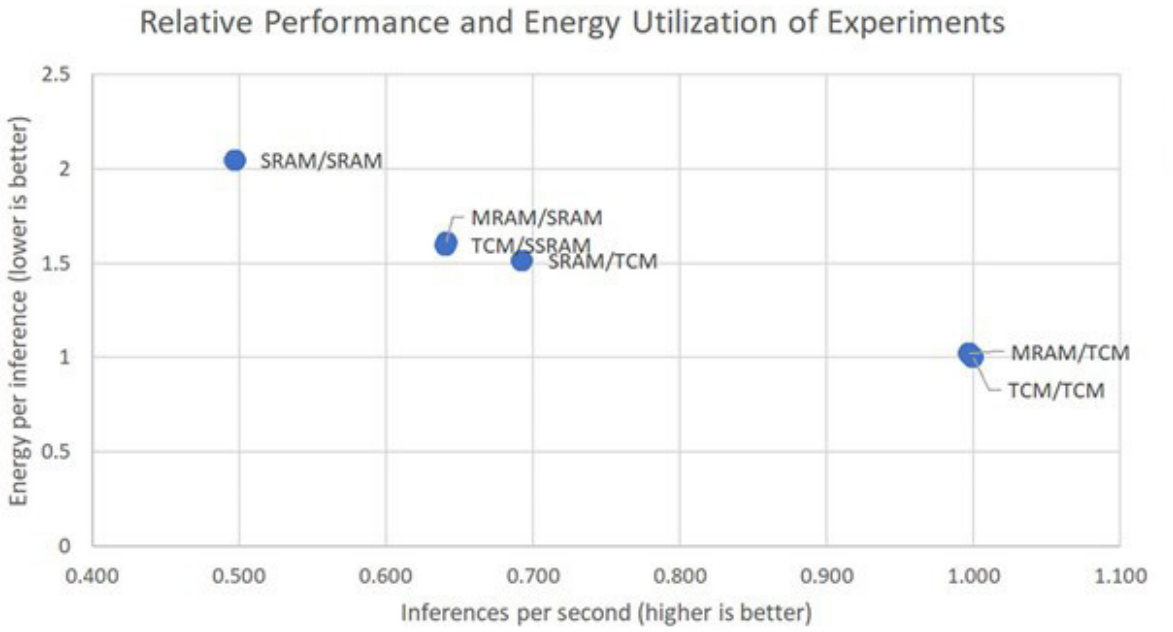
- 我们不要在MRAM中存储激活数据，因为激活数据是动态的，MRAM更适于存储静态数据。
- 我们需要关闭所有不使用的内存。

评估结果

以下图表显示了相对于在TCM中运行所有内容的各种存储方案的评估结果，我们使用轴刻度来放大实验之间的差异。实际上，这些组合中的任何一种都非常适合在IoT端点设备上运行关键词检测。

[1]MLPerf 是一个由学术界、研究实验室和行业的 AI 领导者组成的联盟，其使命是 “ 构建公正且有用的基准测试 ” ，为硬件、软件和服务提供在规定条件下进行的无偏见的训练和推理性能评估。https://mlcommons.org/en/policies/

图 1-4: 相对性能和能效比实验



我们可以看到，当MRAM与TCM或SSRAM配合使用时，可提供更为出色的能效比。

结论

AI需要大量的内存，包括静态和动态的。然而，在实际应用中，AI必须与其他应用共享内存。Apollo4 Plus无论在内存类型还是内存配置方面，都为AI开发者提供了多种选项。在上述实例中，希望提供最优化能效的开发者可以将权重放在Apollo4宽裕的2MB MRAM中，激活数据放置在TCM中，几乎不会产生任何影响。然而，无论开发者选择哪种配置，我们支持SPOT的平台[1]都将始终如一、可靠地提供出色的性能和卓越的功耗效率。

[1] SPOT®，即 Subthreshold Power Optimized Technology，是 Ambiq® 的专有技术平台。它通过提供市场上最节能的解决方案，革命性地拓展了 #EndpointAI 的可能性。

关于Ambiq

Ambiq的使命是通过开发最低功耗的半导体解决方案，使智能设备无处不在，从而推动一个更节能、可持续和数据驱动的世界。Ambiq是基于专有的亚阈值功耗优化技术(SPOT®) 平台的超低功耗半导体解决方案的领导者。SPOT®为我们最终客户的电子产品提供了颠覆性的、成倍的能效提升。Ambiq帮助全球领先的制造商开发出一次充电可运行数周（而不是数天）的产品，同时在紧凑的工业设计中提供最大的功能集。Ambiq的目标是利用其先进的超低功耗SoC解决方案，将人工智能(AI)推向移动和便携式设备领域前所未有的高度。截止到2023年3月，Ambiq的SoC出货量已超过2亿颗。欲了解更多信息，请访问www.ambiq.com。

作者

Carlos Morales

Carlos Morales是Ambiq人工智能部门的副总裁，拥有超过30年从芯片到云的研发经验。除了AI，他的背景专业知识还包括在基于云的后端应用、网络安全、工作负载调度、编排和隔离以及高效网络等方面。

6500 River Place Boulevard, Building 7, Suite 200, Austin, TX 78730
www.ambiq.com
sales@ambiq.com
+1 (512) 879-2850

A-SOCA4P-WPGA01CN v1.0
2023年11月