关于 CoCoPIE XGen

CoCoPIE XGen 是一款专门用于优化 AI 模型以适用于终端设备的工具。XGen 的创建是为了解决 AI 应用开发人员经常遇到的问题。当他们尝试将训练好的 AI 模型部署到设备上时，开发人员通常会发现模型对实际需求来说速度太慢、体积太大或对功耗需求过高。这个问题在移动设备和嵌入式设备上最为严重，如智能手机、Raspberry Pi 和物联网设备。该版本重点支持运行 Android 和 iOS 系统的智能手机。对于其他类型的设备（例如 Raspberry Pi、RISC-V 等）和非 16 位精度，请联系 CoCoPIE 公司。

CoCoPIE XGen 通过提供易于使用的工具来自动优化 AI 模型，解决了 AI 部署问题。借助 XGen，输入的 AI 模型可以变成可直接部署的代码，体积和速度可以缩小数十倍，同时仍然保持令人满意的准确性。通过消除主要障碍并使更多的 AI 任务在终端设备上变得可行，开发人员可以充分利用原本难以触及的巨大移动 AI 市场。

CoCoPIE 现在提供 XGen 作为一款本地软件工具，可安装在用户的设备上，避免对数据/模型安全性的担忧。此特定版本是单节点版本，运行在一台单独的机器上。对于在计算机集群上运行的分布式版本，请联系 CoCoPIE 公司。

XGen 带来的好处

缩短 AI 解决方案的上市时间

通过 XGen 的自动化 AI 优化，使得将 AI 模型部署在移动应用中的过程，从几个月缩短到几天。

降低成本

大大提高的生产效率可以显著降低部署 AI 解决方案以及其维护和升级的成本。此外，由于其优越的优化使得一些基于云的 AI 任务可以在终端设备上运行，XGen 可以帮助企业降低云计算成本。

卓越的性能和紧凑的模型

CoCoPIE 的世界领先的协同优化技术使得 XGen 生成的 AI 模型和代码比现有工具生成的模型小数倍，速度更快，且功耗更高效，同时保持相同的准确性水平。

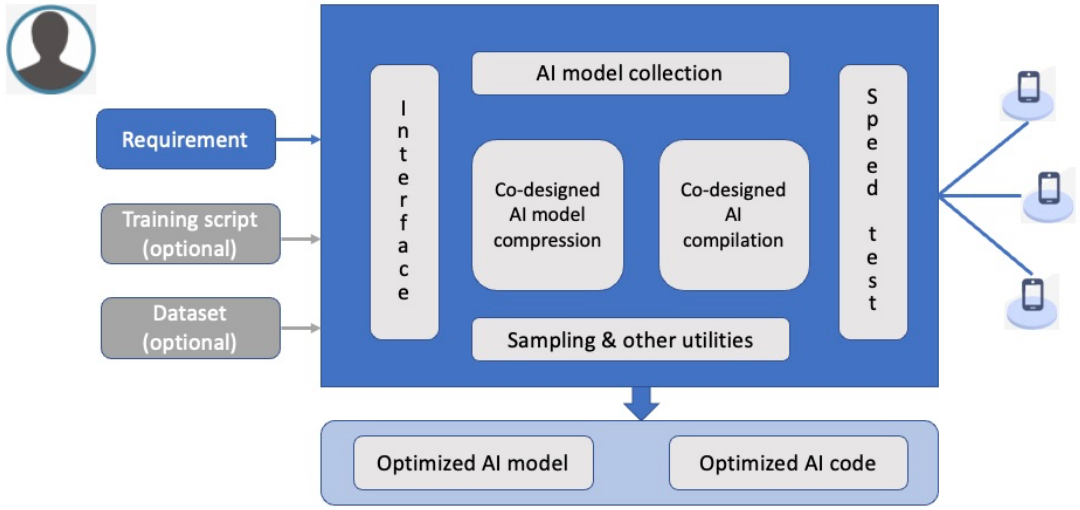
拓展业务机会

通过在终端设备上实现许多 AI 任务的实时性能，XGen 可用于开发许多由于缺乏设备 AI 能力而受阻的新业务机会。

XGen 在 AI 工作流程中的位置

用户可以使用 XGen 轻松获得满足其 AI 需求的 DNN 模型和代码。XGen 支持在公共数据集或用户自己的数据集上生成高效的 AI 模型。

XGen 的高级视图如下所示。这个版本的 XGen 运行在连接有一个或多个真实设备的一台单独的计算机上。它接收用户的需求和其他输入，优化所需的 AI 模型，并输出优化后的模型和代码。在优化过程中，它包括将设备上的速度测试纳入循环，以确保生成的模型的实际性能。它具有一些预定义的 AI 模型，同时也适用于用户自己的 AI 模型。它会重用之前运行生成的模型以节省时间，但不依赖它们来运行。



具体而言，有以下两种主要的使用场景：

场景一：用户需要一个通用的 AI 能力，满足特定的要求（例如，在某些常见设备上达到一定的速度、准确性和大小），而 XGen 已经包含的其中一个基础 AI 模型适用于一般任务，但模型太大或太慢，需要在用户的数据上进行优化以获得更好的性能。
使用方式一：从 XGen 中包含的模型集中，用户选择适合其所需 AI 能力的基础模型，并输入他们的需求；XGen 对模型进行优化，并生成满足要求的模型和代码。

场景二：用户需要一种 AI 能力，但 XGen 已经包含的基础 AI 模型不适用于该类型的任务。用户有自己的模型和训练脚本，但模型太大或太慢。
使用方式二：用户按照使用方式中的指导方针，使其训练脚本与 XGen 的要求兼容，将训练脚本和其他输入提供给 XGen，XGen 自动生成满足用户要求的模型。

除了这些主要的使用场景，XGen 还可用于一些较小的任务，例如训练 AI 模型，评估 AI 模型的质量和速度，测试 AI 训练模型与 XGen 的兼容性，管理 AI 模型等。

通用特性

- 支持广泛的 AI 模型（例如 CNN、RNN、Transformers）用于各种 AI 任务（计算机视

觉、自然语言处理等）。

- 作为本地工具，无需担忧安全和隐私问题。
- 卓越的速度和准确性。
- 可一键解决问题，同时支持完全可定制的用法。
- 灵活支持用户的 DNN 模型和数据。
- 通过世界领先的技术对 DNN 模型和代码进行协同优化。
- 采用专有的尖端 DNN 剪枝方法。
- 全新的有效 AI 编译和代码生成。
- 通过在循环中使用实际设备进行可靠的质量和速度评估。
- 灵活的电源控制 API，确保 AI 的高能效。
- 生成的代码可在所有主流的 Android 和 iOS 设备上运行；对其他设备的支持可以单独添加。

这个版本特定的功能：

- 此 XGen 是单节点版本，运行在单个 GNU/Linux 机器上。它可以使用机器上配备的一个或多个 GPU。
- 生成的 AI 模型精度为 16 位，适用于 Android 9 及 iOS 11 及更高版本。
- 当用于优化客户的 AI 模型时，必须提供 PyTorch 的训练脚本。

它支持多个用户，在一台机器上可以启动多个 XGen 实例容器。

对于特殊处理器或其他类型的设备（例如 iOS、Raspberry Pi、RISC-V 等），或者非 16 位精度（例如通过量化实现的），请联系 CoCoPIE 公司。

更多文档详情，请参考: https://xgen.cocopie.ai