移动端案例丨使用 TFLite 在移动设备上优化与部署风格转化模型

（原文链接）

风格转化 (Style Transfer) 是一种优化技术，用于采集两张图像，一张内容图像（如建筑物），一张风格图像（如著名画家的作品），并将其融合交织在一起，使输出图像看起来就像是以参考风格图像中的风格“画出”了内容图像。

- 风格转化

现在，我们很高兴和大家分享一个用 TensorFlow Lite 针对移动设备优化的预训练风格转化模型，以及在 Android 和 iOS 上的示例应用，可用来为任何图像转换风格。

- 模型
- Android
- iOS

本文中，我们将向您介绍如何优化大型 TensorFlow 模型以进行移动部署，以及如何通过 TensorFlow Lite 在移动应用中高效使用该模型。我们希望您可在您的应用中使用我们的预训练风格转化模型，或受此启发，创建更加有趣的应用。

背景

（风格转化的示例）

风格转化在《一种艺术风格的神经网络算法》*(A Neural Algorithm of Artistic Style)* 中首次发布。但是初始技术的计算量相当大，即使采用高端 GPU，也需要几秒钟才能转换一张图像的风格。接下来几位作者的工作（如：fast-style-transfer ）展示了如何加速风格转化。

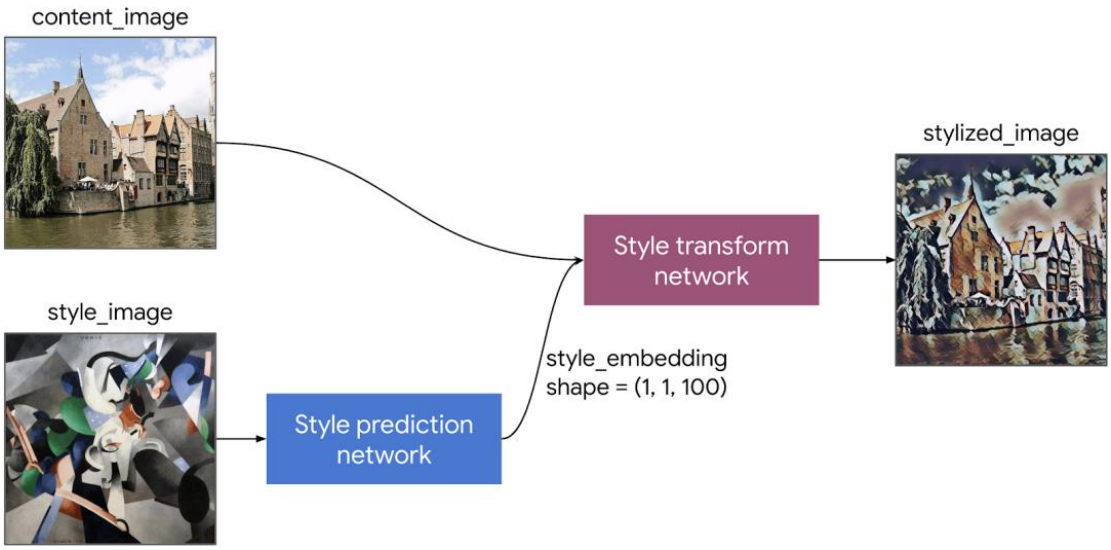
- 一种艺术风格的神经网络算法

- fast-style-transfer

评估几种模型架构后，我们决定一开始先在样本应用中采用来自 Magenta 的预训练的自由风格转化模型。该模型将内容和风格图像作为输入，然后使用前馈神经网络生成风格化的输出图像。与 Gatys 论文中的技术相比，此模型的风格转化速度明显提升，但模型参数量仍然较大 (44 MB)，且速度仍然偏慢（Pixel 4 CPU 上为 2340 毫秒）。因此，我们需要继续优化模型，在移动应用中也适合使用。本文将会分享我们的优化经验，并提供一些资源供您在工作中使用。

- Magenta
- Gatys 论文

优化模型架构



风格转化模型的结构

Magenta 的自由风格转化模型由两个子网组成：

- 风格预测网络：将风格图像转换为风格嵌入矢量。
- 风格转换网络：对内容图像应用风格嵌入矢量，以生成风格化的图像。

Magenta 的风格预测网络采用的是 InceptionV3 骨干网，我们可以将其替换为 MobileNetV2 骨干网，以此来对移动设备进行优化。风格转换网络包含几个卷积层。我们运用 MobileNet 中宽度缩放因子的思路，将所有卷积层的输出通道数缩小为原来的 1/4。

然后，我们必须决定如何训练模型。我们试验了几种方案：从头开始训练移动模型，或者从预训练的 Magenta 模型中提取参数。我们发现：在固定 MobileNetV2 宽度的同时，从头开始优化其他参数得到的结果最好。

这样能达到与原模型相近的效果，而模型的大小显著缩小，速度也大幅提升。

	Original Model	Mobile Model	Improvement
Size	44 MB	10.1 MB	**4.4X smaller**
Latency	2340 ms	459 ms	**5.1X faster**

* 基于 Pixel 4 CPU 的 2 线程 TensorFlow Lite 的基准测试，2020 年 4 月。
* 请参阅此论文 (https://arxiv.org/abs/1705.06830)，了解此风格转化模型中所用的损失函数定义的详情。

量化

敲定模型架构后，我们使用 TensorFlow 模型优化工具包，通过量化来进一步缩小移动模型。量化是适用于大多数 TensorFlow 模型移动部署的一项重要技术，在本例中，它可将模型大小缩小为原来的 1/4，在大幅加速模型推理的同时，对质量的影响很小。

- 量化

在 TensorFlow 提供的多个量化选项中，我们决定使用训练后整型量化，因其能做到简单性和模型质量二者兼顾。在将 TensorFlow 模型转换为 TensorFlow Lite 时，我们只需提供一小部分训练数据集即可。

- 训练后整型量化

与初始模型相比，量化后，我们的模型大小不止缩小了一个量级，速度也不止提升了一个量级，同时将风格和内容损失程度维持在同等水平。

	Original Model	Quantized Mobile Model	Improvement
Size	44 MB	3.1 MB	**14X smaller**
Latency	2340 ms	221 ms	**11X faster**

* 基于 Pixel 4 CPU 的 2 线程 TensorFlow Lite 的基准测试，2020 年 4 月。

移动部署

我们通过一款 Android 应用来展示如何使用风格转化模型。此应用通过采集一张风格图像与一张内容图像，输出将输入图像的风格和内容相融合的图像。

通过手机摄像头的 Camera2 API 拍摄内容图像后，应用提供了一系列名画作为风格图像的可选项。如上所述，通过两个步骤将风格应用于内容图像。首先，我们利用风格预测网络将风格提取为浮点数组。然后，我们利用风格转换网络对内容图像应用此风格。

- Camera2 API

为了在 CPU 和 GPU 上都能达到最佳性能，我们创建了针对每种芯片进行了优化的两组 TensorFlow Lite 模型。我们用 int8 量化模型进行 CPU 推理，用 float16 量化模型进行 GPU 推理。GPU 通常能比 CPU 达到更好的性能，但 GPU 目前仅支持浮点模型，获得的模型 size 比经 int8 量化的模型稍大。以下是 int8 和 float16 模型的表现：

	Int8 model *(CPU, 2 thread)*	Float16 model *(GPU, OpenCL)*	Comparison
Size	3.1 MB	5.1 MB	1.7X larger
Latency	221 ms	34 ms	**6.3X faster**

* 基于 Pixel 4 的 TensorFlow Lite 基准测试，2020 年 4 月。

- int8 量化
- float16 量化

另一种可能提升性能的方式是：缓存风格预测网络的结果，如果您的移动应用仅计划支持一组固定的风格图像。这将进一步缩小您的应用，无需再包含风格预测网络（占总网络大小的 91%）。这是此流程分为两个模型，而不仅仅是一个模型的主要原因。

我们在 GitHub 中提供了示例，应用风格的主类为 StyleTransferModelExecutor。

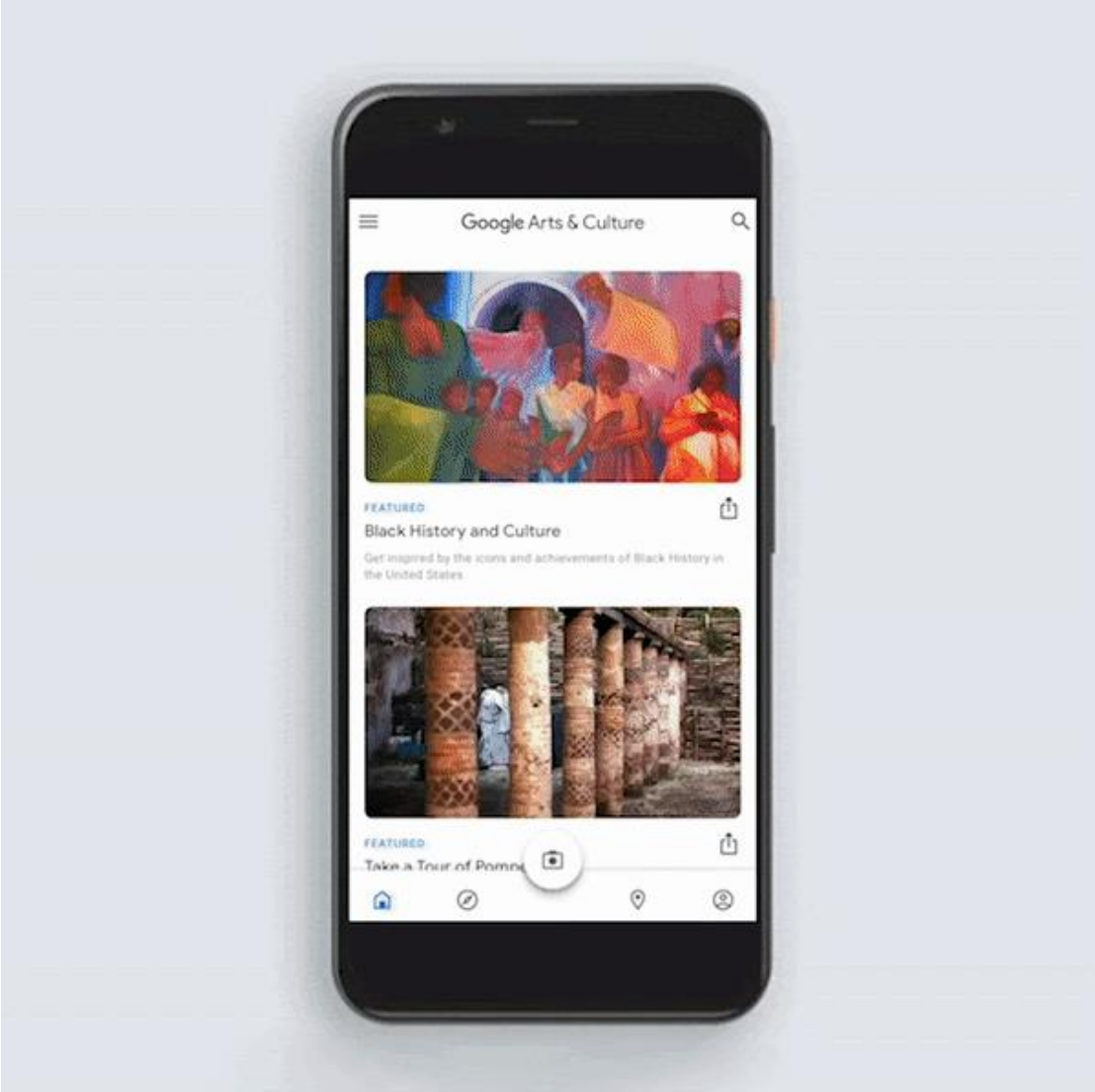
- GitHub
- StyleTransferModelExecutor

因为计算量庞大，我们不会对界面线程运行风格转化，这一点很重要。我们改为使用 AndroidX 中的 ViewModel 类和 Coroutine 来对专用的后台线程运行风格转化，并轻松更新视图。此外，在使用 GPU delegate 运行模型时，TF Lite 解释器初始化、GPU 代理初始化和推理必须在同一线程上运行。

- GPU delegate

生产中的风格转化

Google Arts & Culture 应用中最近添加了 Art Transfer，将利用 TensorFlow Lite 在设备上运行风格转化。所用模型与上述模型相似，但相较于速度和模型大小，该模型更注重质量。对生产环境中的风格转化模型有兴趣的话，您可以试用一下该应用。



- Art Transfer

轮到您了

如果要在自己的应用中添加 Style Transfer，可以来下载移动示例模型。TensorFlow Hub 中提供了 float16（预测网络、转换网络）和 int8 量化版本（预测网络、转换网络）两种模型版本。我们迫不及待地想要看看您的作品！不要忘了与我们分享您的创作。

资源

在设备上运行机器学习模型具有以下优势：保护用户数据隐私，且功能启用时延迟较低。

本文中，我们已经展示了如何将 TensorFlow 模型直接转换为 TensorFlow Lite 模型，但这可能只是迈出的第一步。若要获得良好的性能，开发者应通过量化来优化模型，并权衡好模型质量、模型大小和推理时间之间的关系。

我们通过以下资源来创建模型，也许也适用于您的设备端机器学习用例：

- Magenta 模型库
  Magenta 是一个由 TensorFlow 支持的开源项目，使用机器学习来创作音乐和绘画作品。许多模型均可转换为 TensorFlow Lite，如风格转化模型。

- TensorFlow 模型优化工具包
  模型优化工具包提供多种方法来优化模型，包括量化和剪枝。

- TensorFlow Lite delegate
TensorFlow Lite 可利用设备上提供的多种不同类型的硬件加速器（包括 GPU 和 DSP）来加速模型推理。

如果您想详细了解 *TensorFlow* 的相关内容，请参阅以下文档。这些文档深入探讨了这篇文章中提及的许多主题：

- 移动示例模型

- TensorFlow Hub

- float16
  预测网络
  转换网络

- int8
  预测网络
  转换网络

- **除了上述案例，你还可以阅读 TensorFlow 的其他案例：**

想了解 TensorFlow 的前端应用，可以阅读《前端案例丨零基础也能在小程序上实现机器学习》，了解机器学习如何与小程序进行结合。

希望了解 TensorFlow 如何为企业赋能，可以阅读《企业级案例丨深度学习在网易严选智能客服中的应用》，了解到网易严选如何通过 TensorFlow 为业务的落地开展提供了一整套的方案，包含模型构建、训练并部署至线上使用，从中获得更多灵感启发。

同时，也欢迎大家用微信端打开 TensorFlow 案例库，了解更多精彩案例。
配合官网阅读，体验更佳：**https://tensorflow.google.cn/**

阅读案例以后，你可以通过以下方式持续进阶：

- **你还可以加入 TFUG 社区，认识更多优秀开发者，在社区中进步。**
  TFUG，欢迎你的加入！

  我们为专业的 TensorFlow 开发者提供正式认证和证书，它不仅能够证明你的学习能力，同时也助力你的职业发展点亮 LinkedIn 技能。

- **关注 TensorFlow 官方微信公众号，回复“认证”，即可获得《 TensorFlow 开发者认证候选人手册》，助你在机器学习道路上更进一步：**

期待你顺利踏上 TensorFlow 之旅！