



Implement a neural style transfer on Android with Arm NN APIs

Revision: r0p0

Guide

Non-Confidential

Copyright © 2020, 2022 Arm Limited (or its affiliates).
All rights reserved.

Issue 01

102924_0000_01_en



Implement a neural style transfer on Android with Arm NN APIs

Guide

Copyright © 2020, 2022 Arm Limited (or its affiliates). All rights reserved.

Release information

Document history

Issue	Date	Confidentiality	Change
0100-01	1 March 2020	Non-Confidential	Initial release
2208-01	9 September 2022	Non-Confidential	First release for 22.08

Proprietary Notice

This document is protected by copyright and other related rights and the practice or implementation of the information contained in this document may be protected by one or more patents or pending patent applications. No part of this document may be reproduced in any form by any means without the express prior written permission of Arm. No license, express or implied, by estoppel or otherwise to any intellectual property rights is granted by this document unless specifically stated.

Your access to the information in this document is conditional upon your acceptance that you will not use or permit others to use the information for the purposes of determining whether implementations infringe any third party patents.

THIS DOCUMENT IS PROVIDED "AS IS". ARM PROVIDES NO REPRESENTATIONS AND NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY, SATISFACTORY QUALITY, NON-INFRINGEMENT OR FITNESS FOR A PARTICULAR PURPOSE WITH RESPECT TO THE DOCUMENT. For the avoidance of doubt, Arm makes no representation with respect to, and has undertaken no analysis to identify or understand the scope and content of, patents, copyrights, trade secrets, or other rights.

This document may include technical inaccuracies or typographical errors.

TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL ARM BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF ARM HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

This document consists solely of commercial items. You shall be responsible for ensuring that any use, duplication or disclosure of this document complies fully with any relevant export laws and regulations to assure that this document or any portion thereof is not exported, directly or indirectly, in violation of such export laws. Use of the word “partner” in reference to Arm’s customers is not intended to create or refer to any partnership relationship with any other company. Arm may make changes to this document at any time and without notice.

This document may be translated into other languages for convenience, and you agree that if there is any conflict between the English version of this document and any translation, the terms of the English version of the Agreement shall prevail.

The Arm corporate logo and words marked with ® or ™ are registered trademarks or trademarks of Arm Limited (or its affiliates) in the US and/or elsewhere. All rights reserved. Other brands and names mentioned in this document may be the trademarks of their respective owners. Please follow Arm’s trademark usage guidelines at <https://www.arm.com/company/policies/trademarks>.

Copyright © 2020, 2022 Arm Limited (or its affiliates). All rights reserved.

Arm Limited. Company 02557590 registered in England.

110 Fulbourn Road, Cambridge, England CB1 9NJ.

(LES-PRE-20349|version 21.0)

Confidentiality Status

This document is Non-Confidential. The right to use, copy and disclose this document may be subject to license restrictions in accordance with the terms of the agreement entered into by Arm and the party that Arm delivered this document to.

Unrestricted Access is an Arm internal classification.

Product Status

The information in this document is Final, that is for a developed product.

Feedback

Arm® welcomes feedback on this product and its documentation. To provide feedback on the product, create a ticket on <https://support.developer.arm.com>

To provide feedback on the document, fill the following survey: <https://developer.arm.com/documentation-feedback-survey>.

Inclusive language commitment

Arm values inclusive communities. Arm recognizes that we and our industry have used language that can be offensive. Arm strives to lead the industry and create change.

This document includes language that can be offensive. We will replace this language in a future issue of this document.

To report offensive language in this document, email terms@arm.com.

Contents

1. Introduction.....	6
1.1 Product revision status.....	6
1.2 Intended audience.....	6
1.3 Conventions.....	6
1.4 Useful resources.....	8
2. Overview.....	9
3. What is neural style transfer.....	10
4. What are the Arm NN SDK and the Arm NN APIs?.....	12
5. Looking at the Android code.....	13
6. Arm NN optimization.....	15
7. Deploying the Android Arm NN driver.....	17
8. Tuning performance with OpenCL tuner.....	19
9. Next steps.....	21
10. Revisions.....	22

1. Introduction

1.1 Product revision status

The r_xp_y identifier indicates the revision status of the product described in this manual, for example, $r1p2$, where:

r_x	Identifies the major revision of the product, for example, $r1$.
p_y	Identifies the minor revision or modification status of the product, for example, $p2$.

1.2 Intended audience

This guide is for Software Developers and Application Developers who want to implement a neural style transfer on Android with Arm NN APIs.

1.3 Conventions

The following subsections describe conventions used in Arm documents.

Glossary







The Arm Glossary is a list of terms used in Arm documentation, together with definitions for those terms. The Arm Glossary does not contain terms that are industry standard unless the Arm meaning differs from the generally accepted meaning.

See the Arm® Glossary for more information: developer.arm.com/glossary.

Typographic conventions

Arm documentation uses typographical conventions to convey specific meaning.

Convention	Use
<i>italic</i>	Citations.
bold	Terms in descriptive lists, where appropriate.
monospace	Text that you can enter at the keyboard, such as commands, file and program names, and source code.
monospace <u>underline</u>	A permitted abbreviation for a command or option. You can enter the underlined text instead of the full command or option name.

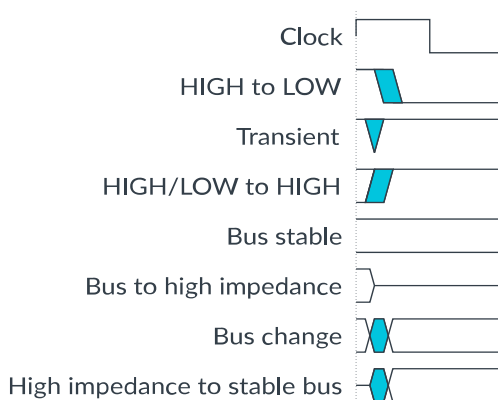
Convention	Use
<and>	Encloses replaceable terms for assembler syntax where they appear in code or code fragments. For example: <pre>MRC p15, 0, <Rd>, <CRn>, <CRm>, <Opcode_2></pre>
SMALL CAPITALS	Terms that have specific technical meanings as defined in the <i>Arm® Glossary</i> . For example, IMPLEMENTATION DEFINED , IMPLEMENTATION SPECIFIC , UNKNOWN , and UNPREDICTABLE .
 Caution	Recommendations. Not following these recommendations might lead to system failure or damage.
 Warning	Requirements for the system. Not following these requirements might result in system failure or damage.
 Danger	Requirements for the system. Not following these requirements will result in system failure or damage.
 Note	An important piece of information that needs your attention.
 Tip	A useful tip that might make it easier, better or faster to perform a task.
 Remember	A reminder of something important that relates to the information you are reading.

Timing diagrams

The following figure explains the components used in timing diagrams. Variations, when they occur, have clear labels. You must not assume any timing information that is not explicit in the diagrams.

Shaded bus and signal areas are undefined, so the bus or signal can assume any value within the shaded area at that time. The actual level is unimportant and does not affect normal operation.

Figure 1-1: Key to timing diagram conventions



Signals

The signal conventions are:

Signal level

The level of an asserted signal depends on whether the signal is active-HIGH or active-LOW. Asserted means:

- HIGH for active-HIGH signals.
- LOW for active-LOW signals.

Lowercase n

At the start or end of a signal name, n denotes an active-LOW signal.

1.4 Useful resources

This document contains information that is specific to this product. See the following resources for other useful information.

Access to Arm documents depends on their confidentiality:

- Non-Confidential documents are available at developer.arm.com/documentation. Each document link in the following tables goes to the online version of the document.
- Confidential documents are available to licensees only through the product package.

Arm product resources	Document ID	Confidentiality
A Neural Algorithm of Artistic Style	-	Non-Confidential
Arm NN SDK	-	Non-Confidential
Fast neural style keras pre-trained models	-	Non-Confidential
Intuitive Guide to Neural Style Transfer	-	Non-Confidential



Arm tests its PDFs only in Adobe Acrobat and Acrobat Reader. Arm cannot guarantee the quality of its documents when used with any other PDF reader.

Adobe PDF reader products can be downloaded at <http://www.adobe.com>

2. Overview

In this guide, we will show you how to build a style transfer Android application with Arm NN APIs.

To work through this guide, you will need the following resources:

- An Android device running Android 9 or later
- Android Studio, including v24 or higher of the SDK build tools

3. What is neural style transfer

Neural style transfer is a technique that uses two images: A content image and a style image. The style image might be, for example, an artwork by a famous painter. A neural style transfer copies the texture, color, and other aspects of the style image and applies them to the content image. You can see an example of a content image, a style image, and a generated image in the following image:

Figure 3-1: An example of a neural style image



Neural style transfer uses a pre-trained Convolutional Neural Network (CNN). Using your content image and your style image, you generate a new image that blends the content image and the style image. You begin with a simple white noise image, or use the content image or style image for optimization efficiency. Then you process the content image, style image, and generated images through the pre-trained neural network. Finally, you calculate loss functions at different layers.

Useful resources includes a link to an article called [Intuitive Guide to Neural Style Transfer](#), which includes more detail.

There are three types of loss functions:

- Content loss function
- Style loss function
- Total loss function

The content loss function ensures that the content present in the content image is captured in the generated image. In a multiple layer CNN, lower layers are more focused on individual pixel values. Higher layers capture information about content. This means that we use the top CNN layer to define the content loss function in our illustration.

The style loss function ensures that the correlation of activations in all the layers is similar between the style image and the generated image.

The total loss function is the weighted sum of the content cost function and the style loss functions for the generated image. The weights are user-defined hyperparameters that control the amount of content and style that is injected into the generated image. Once the loss is calculated,

it can be minimized using backpropagation. Backpropagation optimizes the randomly generated image into a piece of art.

4. What are the Arm NN SDK and the Arm NN APIs?

Arm NN SDK is a set of open-source Linux software tools that enables machine learning workloads on power-efficient devices. In the application that we use in this guide, we use Arm NN to enhance performance on the Arm architecture. The inference engine provides a bridge between existing neural network frameworks and Arm Cortex-A CPUs, Arm Mali GPUs, and NPUs.

The Android Neural Networks API (NNAPI) is an Android C API that is designed to run computationally intensive operations for Machine Learning on mobile devices.

NNAPI supports various hardware accelerations. NNAPI uses TensorFlow as a core technology. If you build your mobile app with TensorFlow, your app gets the benefits of hardware acceleration through the NNAPI. NNAPI abstracts the hardware layer for ML inference. Arm NN works with Android NNAPI to target Arm processors, enabling exponential performance boosts.

A TensorFlow Lite delegate is a way to delegate part or all of graph execution to another executor. Running computationally intensive NN models on mobile devices is resource demanding on mobile CPUs. This means that devices with hardware accelerators provide better performance and higher energy efficiency through NNAPI.

TensorFlow Lite uses an NNAPI delegate to access NNAPI. You can access their [open-source code](#).

5. Looking at the Android code

In this section of the guide, we will explore the Android source code.

Before you begin

You can look at the [Android source code](#) for this guide.

We used pre-trained models and made changes to the model architecture, so that the architecture is fully compatible with Arm NN operators. The changes that we made to the model include:

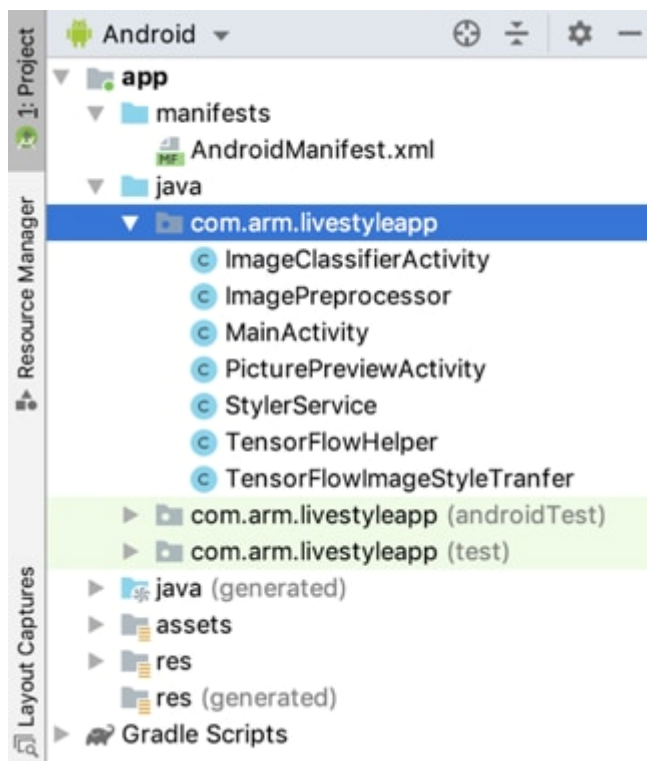
- Replacing all reflection padding with same padding
- Replacing all instance normalization layers with batch normalization layers
- In unpooling layers, using bilinear resize operation instead of the nearest neighbor resize operation
- In the first Conv2D layer, using valid padding instead of same padding

Procedure

To implement the style transfer code, follow these steps:

1. Import the Live Style project into Android Studio. The following screenshot shows the project structure:

Figure 5-1: Screenshot of the project structure



Our style transfer code is implemented in the `doStyleTransfer()` function in `TensorFlowImageStyleTransfer.java`.

1. Convert the image in this function to data that the model can understand, as you can see in the following code:

```
TensorFlowHelper.convertBitmapToByteBuffer(image, intValues, imgData);
```

2. Run inference. The TF Lite interpreter runs the model that is assigned to it. The following line of code runs a neural model without exposing its complexity:

```
tfLite.run(imgData, outputVector);
```

3. Convert the output data of the interpreter to an image, as you can see in the following code:

```
outputImage = Bitmap.createBitmap(mInputImageWidth, mInputImageHeight,  
    Bitmap.Config.ARGB_8888);
```

6. Arm NN optimization

Arm NN uses Arm Compute Library (ACL) to provide a set of optimized operators, for example convolution and pooling, that target Arm-specific accelerators like the DSP (Neon) or the Mali GPU. ACL also provides a GPU tuner tool called CLTuner. CLTuner tunes a set of hardware knobs to fully utilize all the computational horsepower that the GPU provides.

Because Arm NN implements the Android NNAPI interface, developers only need to install the driver. Your Android application will seamlessly interact with the driver to exploit these accelerations.

This part of the code is illustrated in the `TensorFlowImageStyleTransfer()` function in `TensorFlowImageStyleTransfer.java`. To install the driver, the code performs the following steps:

1. Check the Android OS version.
2. Determine whether NNAPI can be enabled on the device.
3. Create a delegate, if NNAPI can be supported:

```
if (enableNNAPI && TensorFlowHelper.canNNAPIEnabled()) {
    delegate = new NnApiDelegate();
    this.tfLiteOptions.addDelegate(delegate);
    this.tfLite = new
Interpreter(TensorFlowHelper.loadModelFile(context, mModelFile), tfLiteOptions);
} else {
    this.tfLite = new
Interpreter(TensorFlowHelper.loadModelFile(context, mModelFile));
}
```

Arm NN implements the Android NNAPI interface. This means that, when developers have the driver installed, your Android application will seamlessly interact with the underlying APIs. This will allow you to exploit the accelerators.

Toggle the NNAPI checkbox to experience the performance enhancement that NNAPI provides.

The Arm NN driver is not bundled with Android releases. Instead, the Arm NN driver is shipped by OEMs like Samsung, HiSilicon, and MTK. For example, all Samsung devices with Android O MR1 or later firmware releases have pre-installed the Arm NN driver.

If your Android device does not have an Arm NN driver pre-installed, or if you want to build your own Arm NN driver, [Deploying the Android Arm NN driver](#) provides information on how to manually install the driver.

Use the Android app to see whether you can create your own art piece. The following generated image of Cambridge is created in La Muse style and built with Arm NN:

Figure 6-1: Generated neural image



7. Deploying the Android Arm NN driver

If you do not see a significant performance acceleration when Arm NN is enabled on your Android phone, you must upgrade your Android NN driver to the latest version.

To use the latest Arm NN Android driver, you must start the driver service. The LiveStyle app creates a delegate that automatically uses the service.

Before you begin

Before you begin, you need the following:

- An Android phone with root access
- A host machine that supports adb
- Adb on host
- A latest [pre-built Arm NN android-nn driver](#) on host
- The built Lifestyle app on your device

Procedure

Follow these steps:

1. Transfer the driver from the host to a local folder on the phone. You must put the driver in a directory with read/write permissions. Here is an example of transferring the driver to the data/local/tmp folder of your phone:

```
user@host: adb push <ArmNN android-nn driver> /data/local/tmp
```

2. Log in to the Android shell to start the driver service, as you can see in the following code:

```
user@host: adb shell
* daemon not running; starting now at tcp:5037
* daemon started successfully
```

3. Log in as root by typing su and cd into the directory containing the latest driver we just pushed, as you can see here:

```
user@android: su
root@android: cd /data/local/tmp
```

4. Start the driver in the background. To make it explicit, pass the -c GpuAcc to enable GPU (OpenCL) acceleration.

```
root@android: ./ -c GpuAcc &
```

5. Open another shell on your host machine, and do a log cat on the most recent kernel messages, as you can see in the following code. This verifies whether the app is using our driver or not:

```
user@host: adb log cat -T 10 ' grep Arm NN
```

6. Run the app on your phone. Choose a style and click TAKE A PICTURE. You should see output containing Arm NN driver information on `log cat`. The output of `log cat` should look something like what you can see in this code:

```
<Time stamp><PID><PID>V ArmnnDriver:  
ArmnnPreparedModel::execute(): 1 input(s), 37 operation(s), 1 output(s), 185  
operand(s)  
...  
<Time stamp><PID><PID>V Arm NN Driver:
```

8. Tuning performance with OpenCL tuner

You can use the Compute Library OpenCL tuner to find optimum values for GPU acceleration tuning parameters. This involves running the driver twice:

- Once in tuning mode to find a set of good parameters
- Once in normal mode, with the tuned parameters

For more information, please refer to Using the GPU tuner in Arm NN android-nn driver. We assume that the Android NN driver is already transferred onto your phone at `/data/local/tmp`.

Procedure

Follow these steps:

1. Run the driver in tuning mode. Tuning could take several minutes, particularly for deeper networks. This means that the OpenCL tuner is difficult to deploy in real-world applications. To overcome this issue, we have introduced three levels of tuning in OpenCL tuner:
 - EXHAUSTIVE - This level offers peak performance with a high tuning time.
 - RAPID - This level offers the shortest tuning time with reduced performance uplift.
 - NORMAL - This level offers the balance between tuning time and a good approximation of the optimal performance.

We use NORMAL level in this example. The tuned parameter output file must be at a writable location.

2. Write the output to a file called `tuned_params` at `/data/local/tmp`:

```
root@android: cd /data/local/tmp
```

3. Create the `tuned_params` file:

```
root@android: touch tuned_params
```

When the service starts in the background, you will see the PID of the service printed in the `stdout`. Make a note of the PID. We need to kill the service with this PID when the tuning process finishes.

4. Start your app as normal. Because the tuning parameters are being selected, you will notice a delay, between one minute up to five minutes, when processing the frame. After the first frame is processed and displayed in the app, the tuning process is finished.
5. Terminate the driver with the PID that you wrote down in step 3, using this code:

```
root@android: kill <PID>
```

If you need to find the PID again, run the following command, followed by "kill <PID>":

```
root@android: ps -Af ' grep <driver>
```

You can also inspect the tuned parameters with:

```
root@android: cat tuned_params
```

6. Restart the driver with the tuned parameters that were created in the `tuned_params` file, as you can see in the following code:

```
root@android: ./<driver> --cl-tuned-parameters-file tuned_params -c GpuAcc &
```

7. Restart the app as usual. You should see a noticeable performance boost.

9. Next steps

Neural style transfer is a useful Machine Learning technique that you can use to turn your photo into a piece of art.

A useful next step is to train your own style transfer model. If you want to train your own model, or a model with a different style, you can follow the training steps detailed in this paper or other open-source projects. You will need to apply the four changes that we mentioned in [Looking at the Android code](#).

10. Revisions

This appendix describes the technical changes between released issues of this book.

Table 10-1: First release for version 1.01

Change	Location
First release	—

Table 10-2: First release for version 22.08

Change	Location
Changed required version of Android to 9 or later. Removed mention of Camera 2 SDK.	Overview