Leveraging the power of C++ for efficient machine learning on embedded devices

Adrian Stanciu
adrian.stanciu.pub@gmail.com

NDC TechTown, 2023

About me

- I am a software engineer from Romania
- I have a bachelor's degree in computer science from Politehnica University of Bucharest
- I have a master's degree in computer and network security from Politehnica University of Bucharest
- I have 12 years of professional experience in Linux system programming
- My main programming languages are C and C++
- I am interested in algorithms, data structures and operating systems
- I work at Bitdefender, where I develop security software solutions that run on routers

Disclaimer

- This presentation is based on a personal project I worked on in my spare time and it is not directly related to the work I do at Bitdefender
- The outcome of this project is just a proof of concept
- All mistakes are mine

Agenda

- Motivation
- Image classification
- Hand gesture recognition
- Best practices
- Summary

Motivation

Machine Learning (ML)

- Subfield of Artificial Inteligence (AI)
- Enables computers to learn from data and then use that knowledge to make predictions
- Applications:
  - Computer vision
  - Information retrieval
  - Medicine
  - Recommender systems
  - Search engines

Embedded devices

- Computing devices designed to perform specific tasks within larger systems
- Applications:
  - Consumer electronics (smart TVs, mobile phones)
  - Home automation (thermostats, lighting control systems)
  - Medical equipment (pacemakers, insulin pumps)
- Characteristics:
  - Limited computer hardware resources
  - Low power consumption
  - May have real-time performance constraints

Machine learning on embedded devices

- Alternative to cloud-based machine learning
- Advantages:
  - Real-time processing
  - Low latency
  - Reduced bandwidth usage
  - Offline operation
  - Improved privacy
- Disadvantages:
  - Compatibility with various hardware and software platforms
  - Maintenance and updates

Using C++ for machine learning on embedded devices

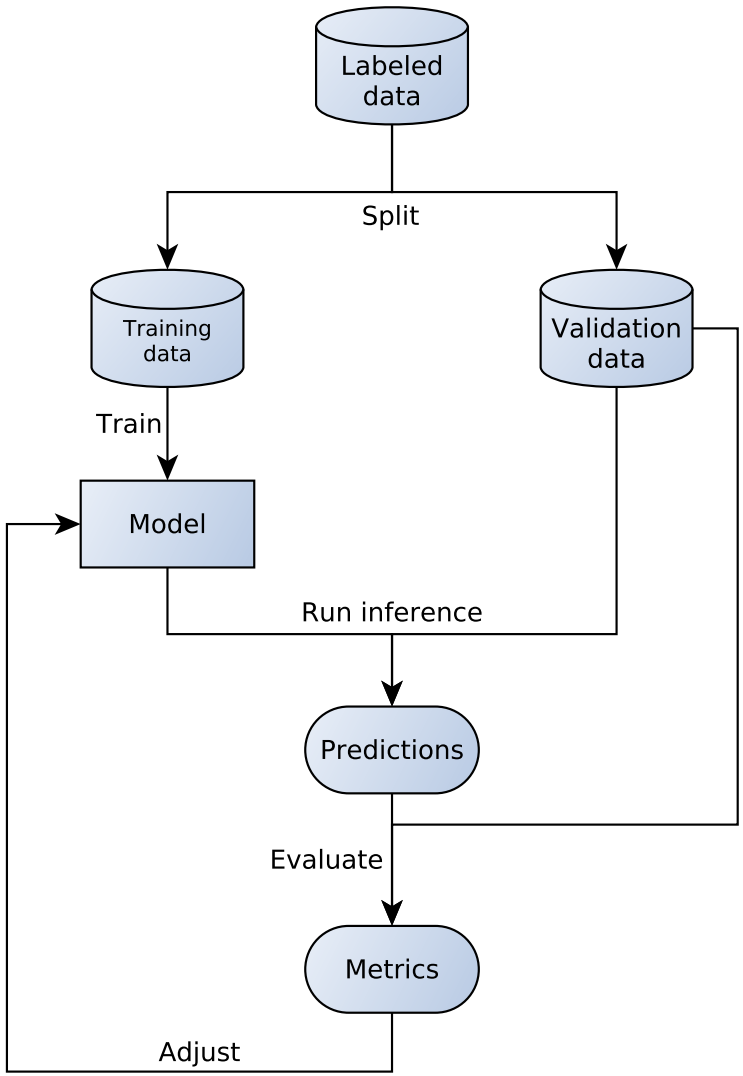
C++ is widely used in embedded systems:

- Was designed with efficiency in mind
- Offers low-level access to hardware resources
- Provides high-level abstractions
- Is supported on most hardware and software platforms

Image classification

Supervised learning

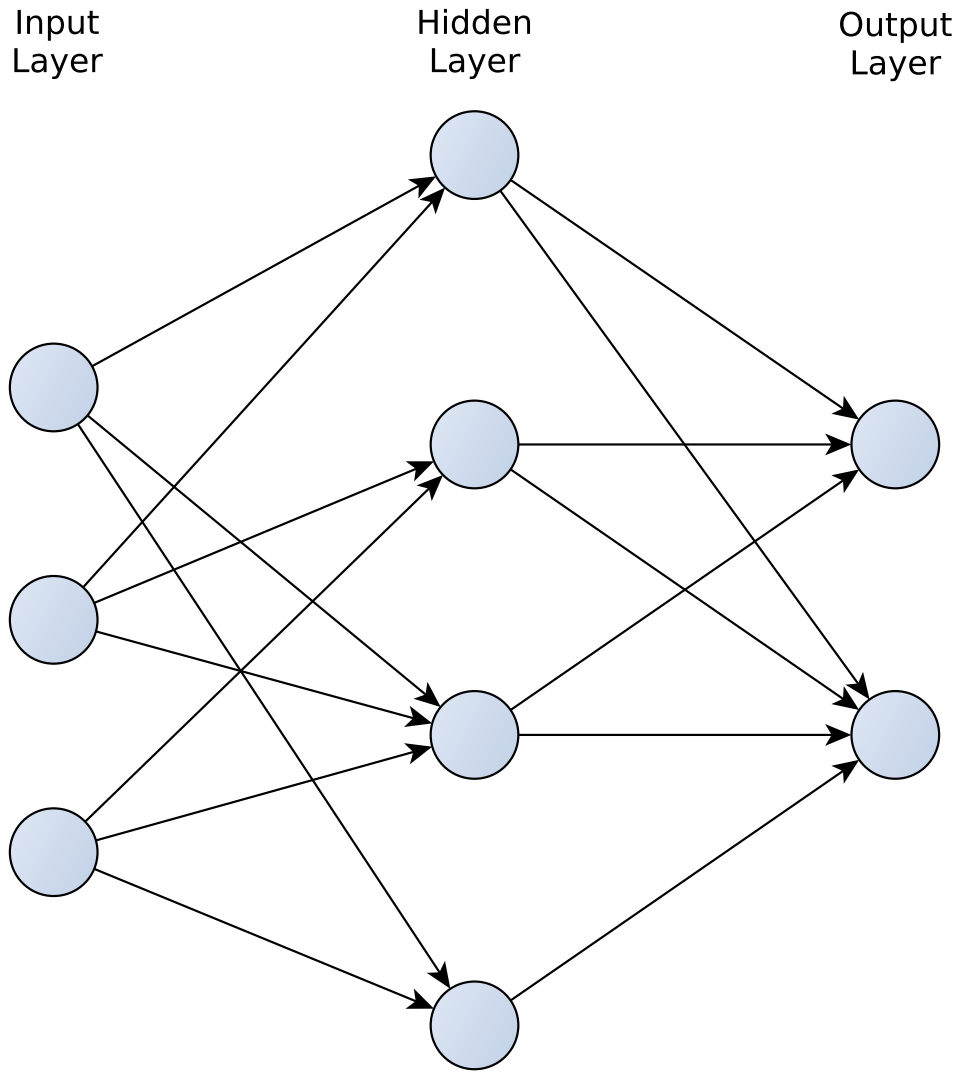
Machine learning paradigm in which an algorithm learns from labeled training data to make predictions



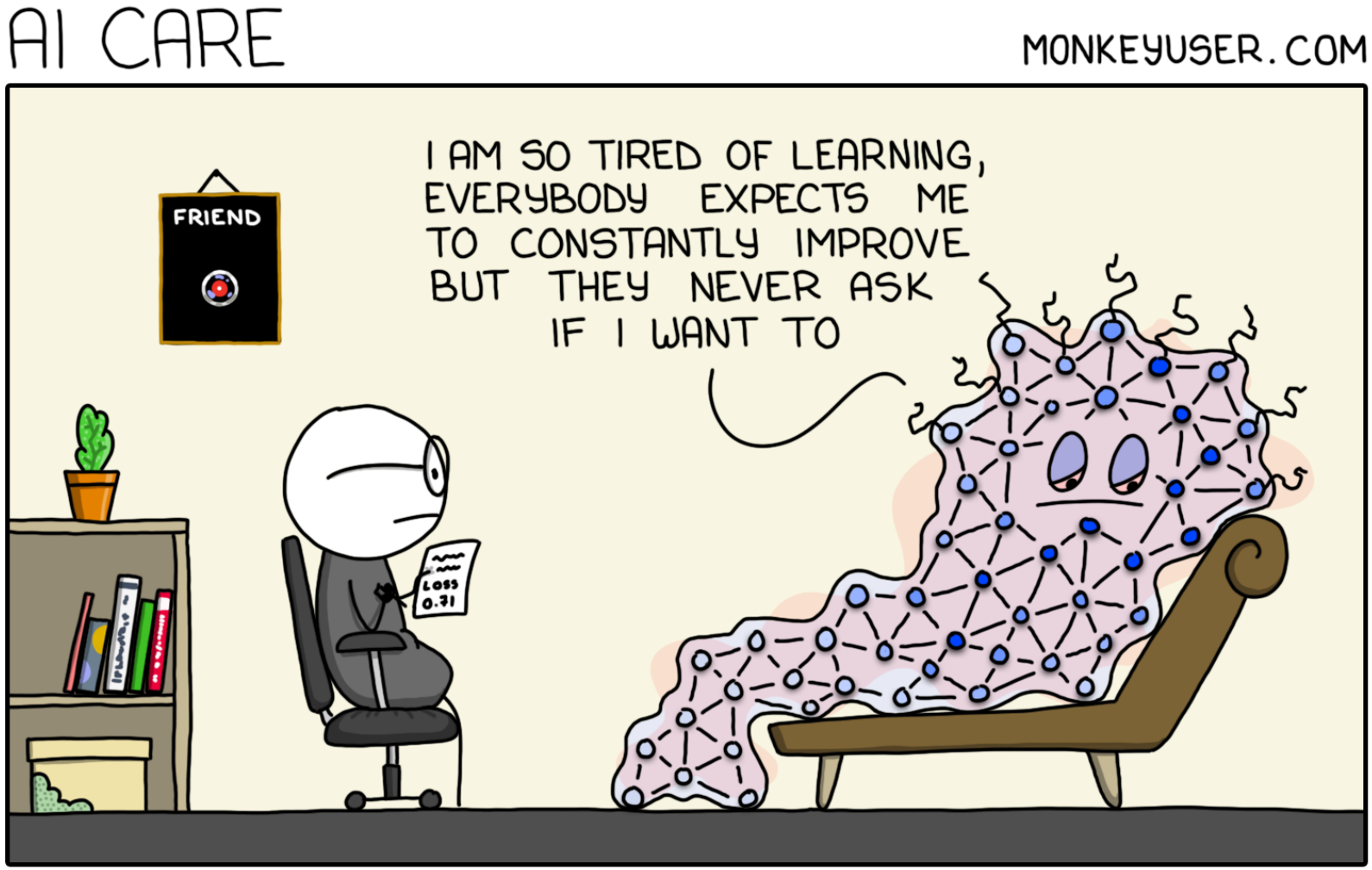
Supervised learning

Source: xkcd.com

Neural network (NN)



Neural network (NN)



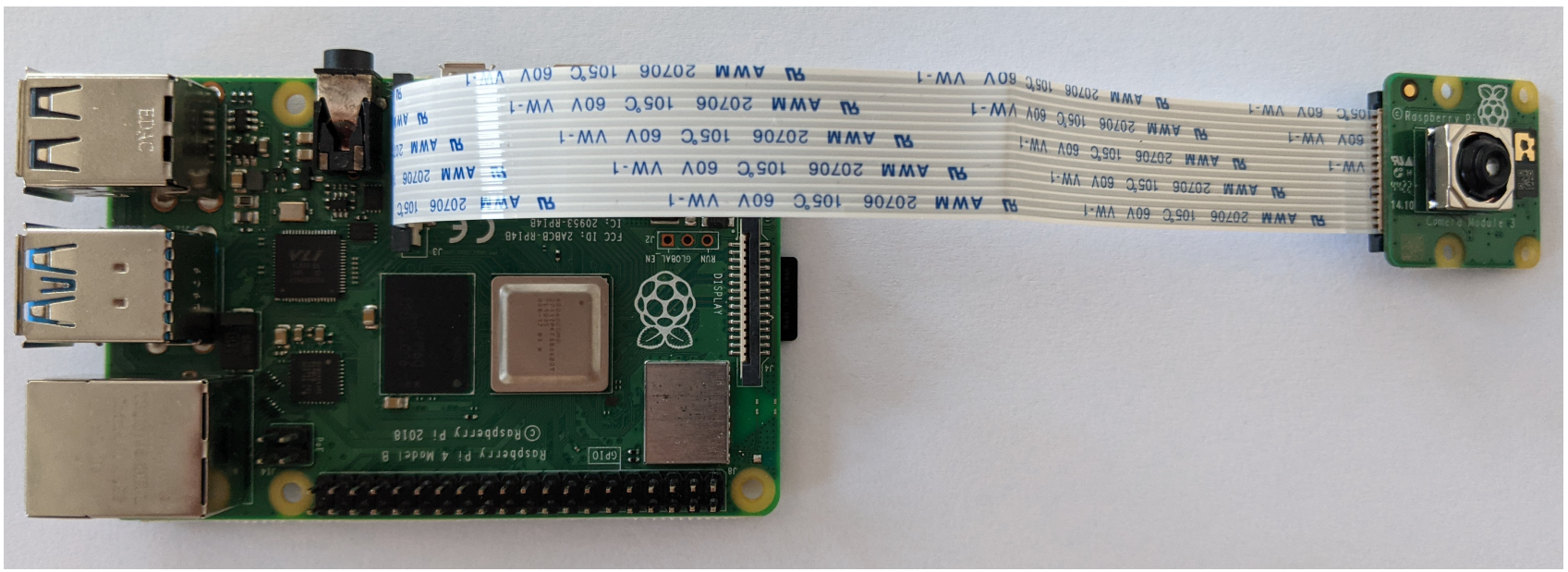
Source: monkeyuser.com

Convolutional neural network (CNN)

- Efficient in image classification
- A convolutional layer can apply filters to detect:
  - Edges
  - Shapes
  - Objects

Hardware setup

- Raspberry Pi 4 model B:
  - Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.8GHz
  - 4GB RAM
- 8GB microSD card
- Raspberry Pi Camera module 3



Software

- 64-bit Raspberry Pi OS
- gcc with C++20 support
- TensorFlow Lite
- OpenCV

TensorFlow Lite

- Optimized pre-trained models
- C++ and Python APIs

Data organization

- A tensor is an array in which data (both input and output) is organized
- For 2D images the size of the tensor is computed as `width * height * channels * sizeof(type)`

MobileNet pre-trained model

- CNN architecture created by Google
- 1000 classes (labels)
- Labels are stored separate from the model:
  - `labels_mobilenet_quant_v1_224.txt` (11KB)
  - `mobilenet_v1_1.0_224_quant.tflite` (4.1MB)
- Tensor dimensions:
  - 224x224 pixels images
  - 3 color channels per pixel (RGB)
  - Type is `uint8_t`

Algorithm

1. Load model and labels
2. Build interpreter
3. Allocate input and output tensors
4. Read image
5. Resize image
6. Copy resized image to input tensor
7. Run inference
8. Extract results from output tensor

Steps 4-8 can be repeated multiple times

1. Load model and labels

```
1  // const char *model_path;
2  // const char *labels_path;
3
4  std::unique_ptr<tflite::FlatBufferModel> model{
5    tflite::FlatBufferModel::BuildFromFile(model_path)
6  };
7
8  std::ifstream labels_ifs{labels_path};
9  std::string label;
10 std::vector<std::string> labels;
11 while (getline(labels_ifs, label)) {
12   labels.push_back(std::move(label));
13 }
```

2. Build interpreter

```
// std::unique_ptr<tflite::FlatBufferModel> model;
// int num_threads;

tflite::ops::builtin::BuiltinOpResolver resolver;
tflite::InterpreterBuilder builder{*model, resolver};

builder.SetNumThreads(num_threads);

std::unique_ptr<tflite::Interpreter> interpreter;
builder(&interpreter);
```

3. Allocate input and output tensors

```
1  // std::unique_ptr<tflite::Interpreter> interpreter;
2
3  interpreter->AllocateTensors();
```

4. Read image

```
1  // const char *image_path;
2
3  cv::Mat image{cv::imread(image_path)};
```

5. Resize image

```
// cv::Mat image;
// int required_image_width;
// int required_image_height;

cv::Mat resized_image;
cv::resize(
  image,
  resized_image,
  cv::Size{required_image_width, required_image_height}
);
```

6. Copy resized image to uint8_t-typed input tensor

```
1  // std::unique_ptr<tflite::Interpreter> interpreter;
2  // cv::Mat resized_image;
3
4  std::memcpy(
5    interpreter->typed_input_tensor<uint8_t>(0),
6    resized_image.data,
7    resized_image.total() * resized_image.elemSize()
8  );
```

6. Copy resized image to float-typed input tensor - Wrong!

```
1  // std::unique_ptr<tflite::Interpreter> interpreter;
2  // cv::Mat resized_image;
3
4  std::memcpy(
5    interpreter->typed_input_tensor<float>(0),
6    resized_image.data,
7    resized_image.total() * resized_image.elemSize()
8  );
```

6. Copy resized image to float-typed input tensor - Correct

```
// std::unique_ptr<tflite::Interpreter> interpreter;
// cv::Mat resized_image;

auto idx = 0;
for (auto row = 0; row < resized_image.rows; ++row) {
  for (auto col = 0; col < resized_image.cols; ++col) {
    auto pixel = resized_image.at<cv::Vec3b>(row, col);
    for (auto ch = 0; ch < pixel.channels; ++ch) {
      tensor_data[idx++] = pixel.val[ch];
    }
  }
}
```

7. Run inference

```
1  // std::unique_ptr<tflite::Interpreter> interpreter;
2
3  interpreter->Invoke();
```

8. Extract results from uint8_t-typed output tensors

```
// std::unique_ptr<tflite::Interpreter> interpreter;
// std::vector<std::string> labels;

std::span<uint8_t> outputs{
  interpreter->typed_output_tensor<uint8_t>(0),
  labels.size()
};

std::vector<float> probabilities;
probabilities.reserve(labels.size());
for (auto output : outputs) {
  probabilities.push_back(
    1.0f * output / std::numeric_limits<uint8_t>::max()
  );
}
```

8. Extract results from float-typed output tensors

```
1  // std::unique_ptr<tflite::Interpreter> interpreter;
2  // std::vector<std::string> labels;
3
4  std::span<float> probabilities{
5    interpreter->typed_output_tensor<float>(0),
6    labels.size()
7  };
```

Demo

```
$ libcamera-jpeg --width=640 --height=480 -o out.jpeg
```

Demo - Good results


```
0.96 | tennis ball
```

Demo - Bad results

```
0.30 | vase
0.25 | bell pepper
ambiguous results
```

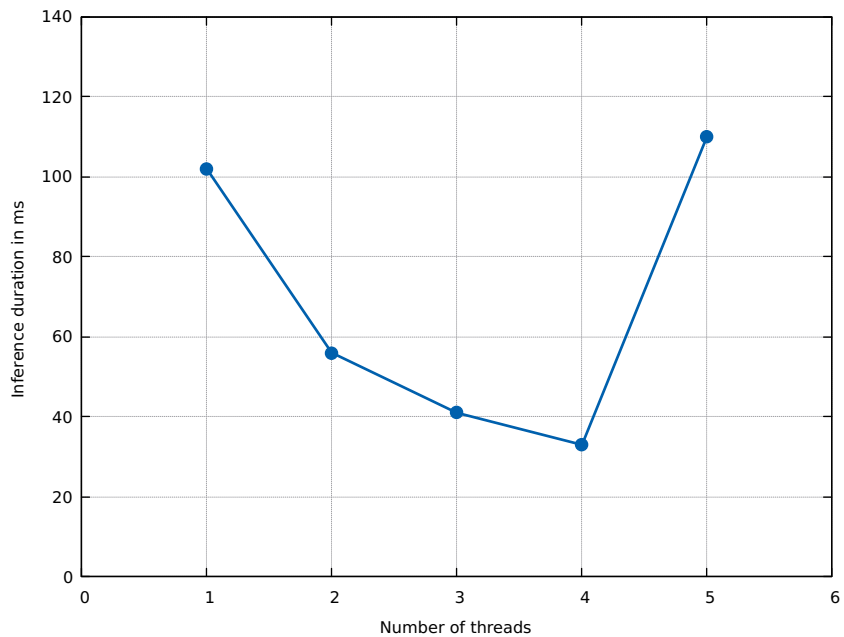
Demo - Bad results explanation

```
$ grep "tomato" labels_mobilenet_quant_v1_224.txt
$
```

Performance

- Compilation duration: 35s
- Binary size: 67KB
- Running duration: 1s

Number of threads	Inference duration (ms)
1	102
2	56
3	41
4	33
5	110



- Memory consumption with 4 threads: 93MB

Comparision with Python

Comparision made with TensorFlow Lite's `label_image.py`

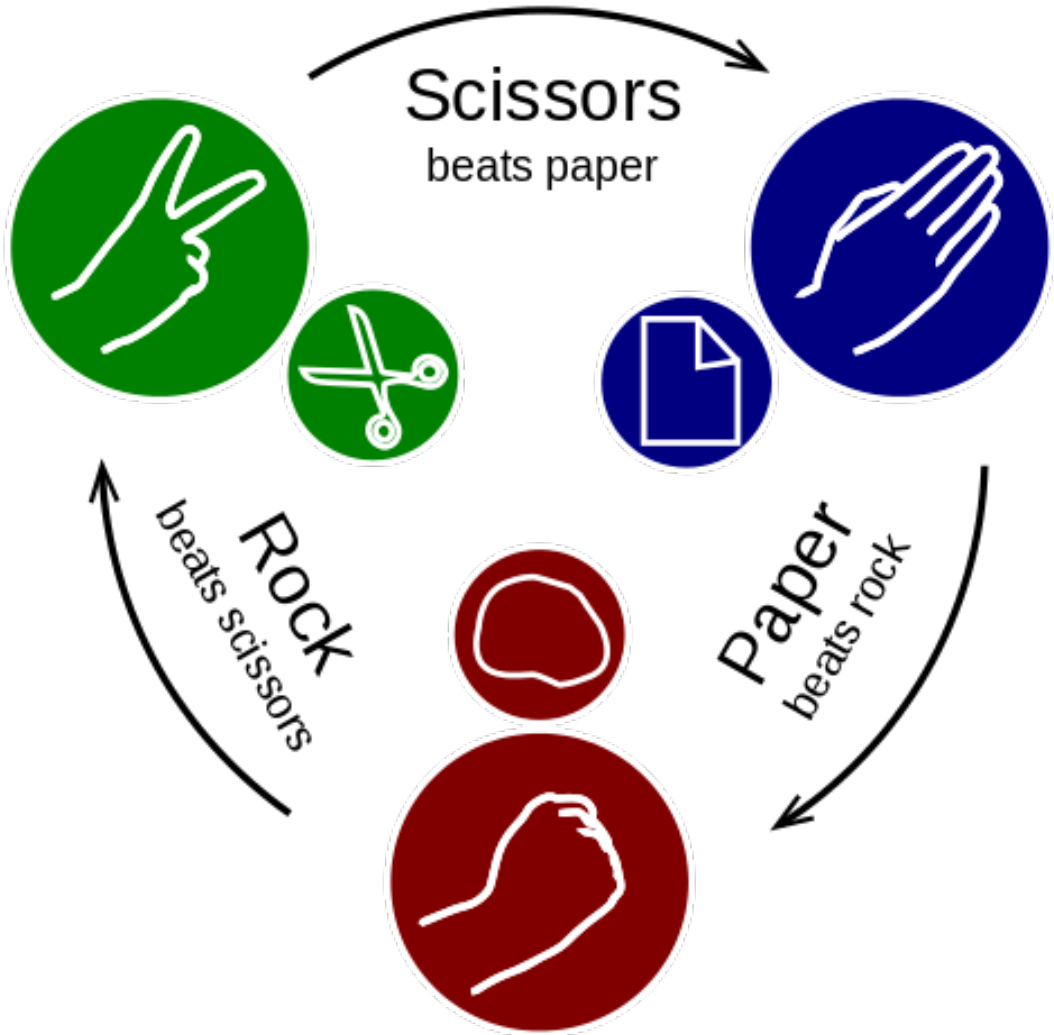
	C++	**Python**
Running duration (s)	1	7

Number of threads	**Inference duration (ms)**	
	C++	**Python**
1	102	118
2	56	62
3	41	42
4	33	33
5	110	170

	C++	**Python**
Memory consumption with 4 threads (MB)	93	320

Hand gesture recognition

Rock-Paper-Scissors



Source: wikipedia.org

Data

- Google MediaPipe's Rock-Paper-Scissors dataset for hand gesture recognition
  - Has a "none" class
  - Contains 125 images for each class
  - Images have various sizes
  - Images have 3 color channels per pixel (RGB)
- Laurence Moroney's Rock-Paper-Scissors dataset (published by Sani Kamal on Kaggle)
  - Contains 964 images for each class split into training set (840) and testing set (124)
  - Images are 300x300 pixels
  - Images have 3 color channels per pixel (RGB)

Train a CNN model - Prerequisites

- To be run on host
- Python
- TensorFlow Lite
- OpenCV

Train a CNN model - Algorithm

1. Build the model
2. Configure the model
3. Load the data
4. Train the model
5. Convert to lite model

Train a CNN model - 1. Build the model

```
base_model = tf.keras.applications.ResNet50(
  include_top = False,
  input_shape = (96, 96, 3),
  pooling = "avg"
)

base_model.trainable = False

model = tf.keras.models.Sequential()

model.add(base_model)
model.add(tf.keras.layers.Dropout(rate = 0.5))
model.add(
  tf.keras.layers.Dense(
    units = 3,
    activation = tf.keras.activations.softmax
  )
)
```

Train a CNN model - 2. Configure the model

```
model.compile(
  optimizer = tf.keras.optimizers.SGD(),
  loss = "categorical_crossentropy",
  metrics = ["accuracy"]
)
```

Train a CNN model - 3. Load the data

```
1  LABELS = ["rock", "paper", "scissors"]
2
3  training_data = tf.keras.utils.image_dataset_from_directory(
4    "data/training",
5    image_size = (96, 96),
6    label_mode = "categorical",
7    class_names = LABELS
8  )
9
10 validation_data = tf.keras.utils.image_dataset_from_directory(
11   "data/validation",
12   image_size = (96, 96),
13   label_mode = "categorical",
14   class_names = LABELS
15 )
```

Train a CNN model - 4. Train the model

```
model.fit(
  training_data,
  epochs = 10,
  validation_data = validation_data
)
```

Train a CNN model - 5. Convert to lite model

```
converter = tf.lite.TFLiteConverter.from_keras_model(model)
converter.optimizations = [tf.lite.Optimize.DEFAULT]

lite_model = converter.convert()
```

Train on small dataset

- Random shuffle and split the images into:
  - 80% for training (300 images)
  - 20% for validation (75 images)

```
Epoch 1/10
9s - loss: 2.0545 - accuracy: 0.5700 - val_loss: 3.0788 - val_accuracy: 0.4800
Epoch 2/10
5s - loss: 1.4671 - accuracy: 0.7300 - val_loss: 1.5156 - val_accuracy: 0.7467
Epoch 3/10
5s - loss: 0.6304 - accuracy: 0.8567 - val_loss: 0.1994 - val_accuracy: 0.9467
Epoch 4/10
5s - loss: 0.3750 - accuracy: 0.9100 - val_loss: 0.3066 - val_accuracy: 0.9067
Epoch 5/10
5s - loss: 0.4212 - accuracy: 0.8933 - val_loss: 0.2647 - val_accuracy: 0.9333
Epoch 6/10
5s - loss: 0.3357 - accuracy: 0.8933 - val_loss: 0.1780 - val_accuracy: 0.9200
Epoch 7/10
5s - loss: 0.1435 - accuracy: 0.9533 - val_loss: 0.1747 - val_accuracy: 0.9733
Epoch 8/10
5s - loss: 0.2573 - accuracy: 0.9267 - val_loss: 0.1982 - val_accuracy: 0.9333
Epoch 9/10
5s - loss: 0.1846 - accuracy: 0.9367 - val_loss: 0.1352 - val_accuracy: 0.9867
Epoch 10/10
5s - loss: 0.1590 - accuracy: 0.9433 - val_loss: 0.1135 - val_accuracy: 0.9600
```

- Duration: 2m15s
- Model size: 23MB

Test on small dataset

Using the testing dataset of the big dataset (124 images per class)

Gesture	**Correct predictions**	**Accuracy (%)**
Rock	108	87.09
Paper	124	100
Scissors	8	6.45
All	240 out of 372	64.51

Train on big dataset

- Random shuffle and split the training set:
  - 80% for training (2016 images)
  - 20% for validation (504 images)

```
Epoch 1/10
37s - loss: 1.0362 - accuracy: 0.8284 - val_loss: 0.1478 - val_accuracy: 0.9385
Epoch 2/10
35s - loss: 0.1554 - accuracy: 0.9444 - val_loss: 0.0386 - val_accuracy: 0.9861
Epoch 3/10
35s - loss: 0.1092 - accuracy: 0.9608 - val_loss: 0.0256 - val_accuracy: 0.9901
Epoch 4/10
35s - loss: 0.0918 - accuracy: 0.9668 - val_loss: 0.0133 - val_accuracy: 0.9980
Epoch 5/10
35s - loss: 0.0589 - accuracy: 0.9797 - val_loss: 0.0325 - val_accuracy: 0.9861
Epoch 6/10
35s - loss: 0.0500 - accuracy: 0.9826 - val_loss: 0.0091 - val_accuracy: 0.9980
Epoch 7/10
35s - loss: 0.0361 - accuracy: 0.9896 - val_loss: 0.0070 - val_accuracy: 1.0000
Epoch 8/10
35s - loss: 0.0410 - accuracy: 0.9856 - val_loss: 0.0071 - val_accuracy: 1.0000
Epoch 9/10
35s - loss: 0.0310 - accuracy: 0.9886 - val_loss: 0.0074 - val_accuracy: 1.0000
Epoch 10/10
35s - loss: 0.0395 - accuracy: 0.9866 - val_loss: 0.0046 - val_accuracy: 1.0000
```

- Duration: 7m10s
- Model size: 23MB

Test on big dataset

Using its own testing dataset (124 images per class)

Gesture	**Correct predictions**	**Accuracy (%)**
Rock	124	100
Paper	122	98.38
Scissors	97	78.22
All	343 out of 372	92.20

Inference using the trained CNN model

- Same hardware setup
- Same software dependencies
- Same C++ code

Demo

Demo - Small dataset results

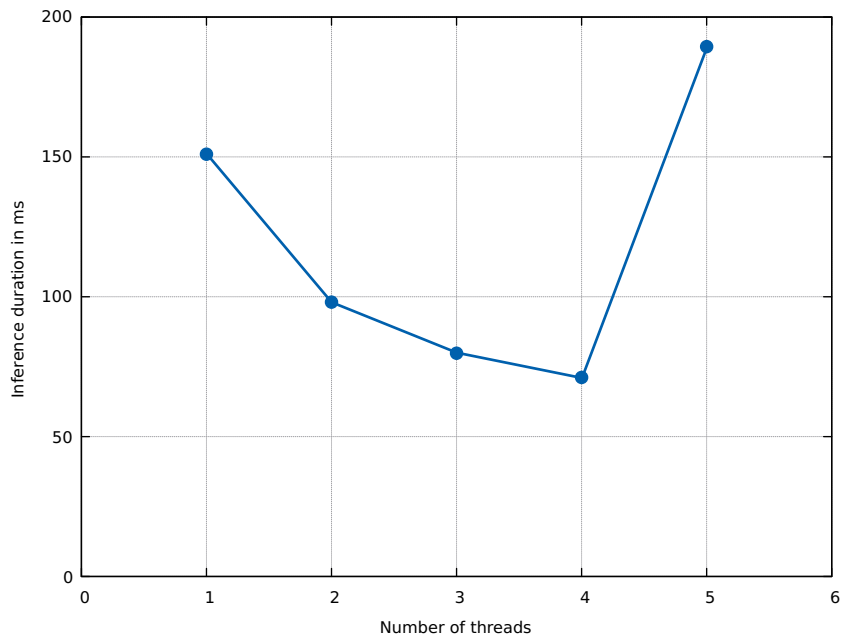

```
0.54 | scissors
0.31 | paper
0.15 | rock
ambiguous results
```

Demo - Big dataset results

Performance

- Compilation duration: 35s
- Binary size: 67KB
- Running duration: 1s

Number of threads	Inference duration (ms)
1	151
2	98
3	80
4	71
5	189



- Memory consumption with 4 threads: 118MB

Live camera stream

- libcamera backend
- gstreamer middleware
- OpenCV frontend

Live camera stream

```
 1  static constexpr const char *GstreamerPipeline{R"(
 2      libcamerasrc !
 3      video/x-raw,
 4      width=(int)640,
 5      height=(int)480,
 6      framerate=(fraction)10/1 !
 7      videoconvert !
 8      appsink
 9  )"};
10
11  cv::VideoCapture camera{
12    GstreamerPipeline,
13    cv::CAP_GSTREAMER
14  };
15
16  cv::Mat image;
17  camera >> image;
```

Live inference - Algorithm

1. Create image classifier
2. Open camera
3. Do in a loop:
   - 3.1 Read image from camera
   - 3.2 Use classifier to run inference on image
   - 3.3 Show predictions

Live inference - Demo

Live inference - Performance

Stats for 4 threads:

- Inference duration: 71ms
- Loop duration: 96ms
- Memory consumption: 123MB

Rock-Paper-Scissors game mode

On each round the AI:

1. Chooses Rock, Paper or Scissors uniformly at random
2. Infers the human's hand gesture
3. Computes the outcome

Rock-Paper-Scissors game mode - Demo

Best practices

Best practices

- Integer quantization optimization
- Cross-compilation
- Binary size analysis
- Updates

Integer quantization optimization

- The process of converting 32-bit floating-point numbers to the 8-bit integer numbers
- Reduces the size of the model and decreases inference duration at the expense of accuracy

Cross-compilation

- The process to compile code for one system, named target (e.g. aarch64), on a different system, named host (e.g. x86_64)
- Enables faster builds

Binary size analysis

Using Google's Bloaty McBloatface

```
$ bloaty image_classifier -d symbols
```

```
 1       FILE SIZE        VM SIZE
 2    --------------  --------------
 3     44.1%  29.5Ki  48.2%  29.5Ki    [section .text]
 4      8.3%  5.53Ki   9.0%  5.53Ki    [section .dynstr]
 5      7.7%  5.12Ki   8.4%  5.12Ki    [section .eh_frame]
 6      7.3%  4.90Ki   8.0%  4.90Ki    [section .rela.dyn]
 7      5.4%  3.59Ki   0.0%       0    [Unmapped]
 8      5.0%  3.38Ki   5.5%  3.38Ki    [section .dynsym]
 9      4.8%  3.18Ki   5.2%  3.18Ki    [section .rodata]
10      2.9%  1.97Ki   3.2%  1.97Ki    [section .rela.plt]
11      2.7%  1.81Ki   0.0%       0    [ELF Headers]
12      2.0%  1.37Ki   2.2%  1.37Ki    [section .data.rel.ro]
13      2.0%  1.34Ki   2.2%  1.34Ki    [section .plt]
14      1.6%  1.07Ki   1.7%  1.07Ki    [section .eh_frame_hdr]
15      1.1%     760   1.2%     760    [section .gcc_except_table]
16      1.0%     696   1.1%     696    [section .got.plt]
17      0.9%     624   1.0%     624    [section .dynamic]
18      0.9%     598   1.0%     598    [LOAD #2 [RX]]
19      0.8%     530   0.9%     579    [11 Others]
20      0.4%     288   0.5%     288    [section .gnu.version]
21      0.4%     274   0.0%       0    [section .shstrtab]
22      0.4%     244   0.4%     244    [section .gnu.hash]
23      0.3%     208   0.3%     208    [section .gnu.version_r]
24    100.0%  66.9Ki 100.0%  61.3Ki    TOTAL
```

Binary size analysis

If compiled with debug info (-g)

```
$ bloaty image_classifier -d symbols
```

```
     FILE SIZE        VM SIZE
 --------------  --------------
  42.0%  1.76Mi   0.0%       0    [section .debug_info]
  40.8%  1.71Mi   0.0%       0    [section .debug_str]
   8.8%   377Ki   0.0%       0    [section .debug_loc]
   3.4%   145Ki   0.0%       0    [section .debug_ranges]
   1.9%  82.6Ki   0.0%       0    [section .debug_line]
   1.3%  53.9Ki  53.1%  32.6Ki    [87 Others]
   0.5%  22.3Ki   0.0%       0    [section .debug_abbrev]
   0.1%  5.25Ki   7.5%  4.61Ki    std::__introsort_loop<>()
   0.1%  5.07Ki   0.0%       0    [section .strtab]
   0.1%  4.85Ki   0.0%       0    [section .symtab]
   0.1%  4.09Ki   6.7%  4.09Ki    [section .dynstr]
   0.1%  3.76Ki   5.1%  3.11Ki    std::__adjust_heap<>()
   0.1%  3.63Ki   4.8%  2.97Ki    std::__insertion_sort<>()
   0.1%  3.59Ki   0.0%       0    [Unmapped]
   0.1%  3.56Ki   3.6%  2.18Ki    cvflann::anyimpl::big_any_policy<>
   0.1%  3.34Ki   0.0%       0    [section .debug_aranges]
   0.1%  3.15Ki   3.4%  2.07Ki    cvflann::anyimpl::small_any_policy<>
   0.1%  3.06Ki   4.9%  2.99Ki    ic::ImageClassifier::run()
   0.1%  2.97Ki   2.0%  1.23Ki    cvflann::anyimpl::typed_base_any_policy<>
   0.1%  2.83Ki   4.4%  2.71Ki    ic::ImageClassifier::ImageClassifier()
   0.1%  2.78Ki   4.5%  2.75Ki    main
 100.0%  4.19Mi 100.0%  61.3Ki    TOTAL
```

Updates

The model can be updated:

- Together with the application (use `#embed` from C23 and replace `tflite::FlatBufferModel::BuildFromFile()` with `tflite::FlatBufferModel::BuildFromBuffer()`)
- Separately from the application

Summary

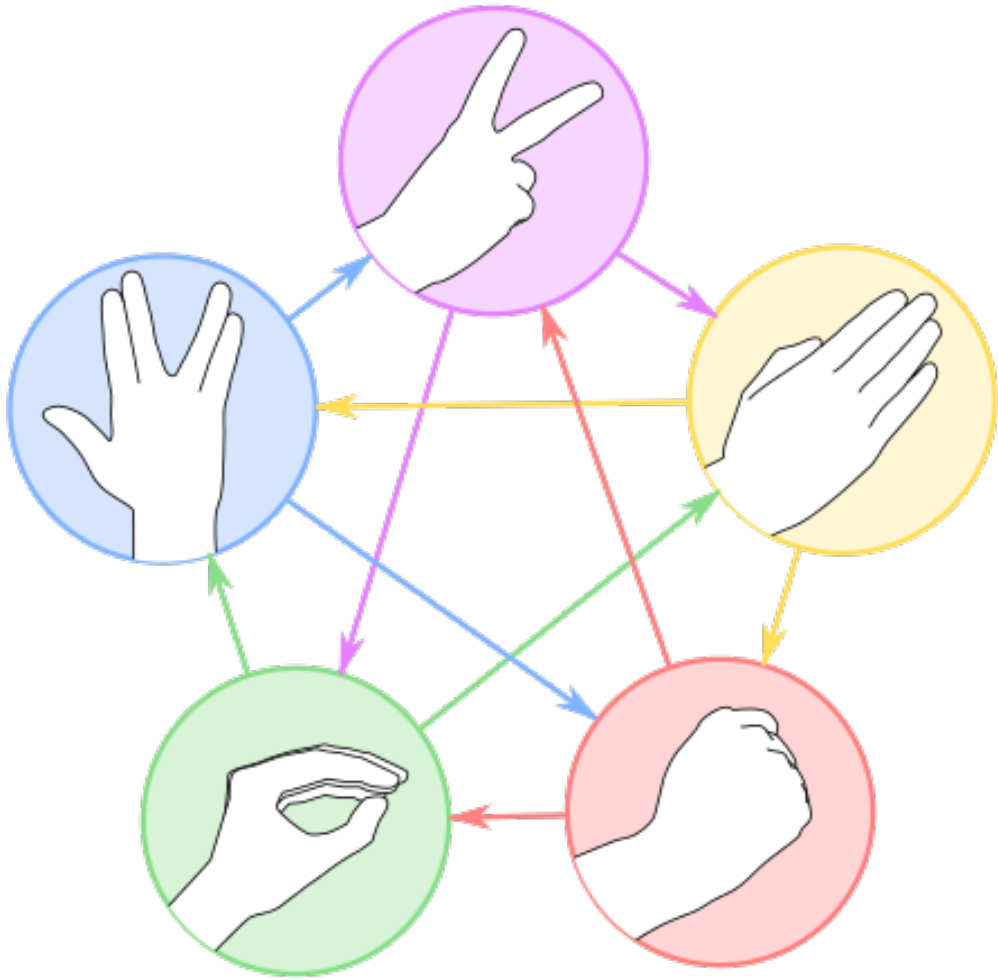
Recap

Task	**Model**	**Inference**
Image classification	pre-trained MobileNet	on-device in C++
Rock-Paper-Scissors hand gesture recognition	Resnet trained on host in Python	

Future work

- Introduce the "none" class to filter out ambiguities
- Reduce the size of the model and decrease inference duration by:
  - Enabling integer quantization optimization
  - Converting images to gray scale
- Use `libcamera`'s API directly to reduce dependencies

Future work



Source: wikipedia.org

Conclusions

- Code isn't enough... data matters
- More diverse data leads to better models
- Building accurate models is an expert job
- Running on-device inference is here to stay

Repository

`https://github.com/adrian-stanciu/cpp-embedded-ml`

Resources

- https://github.com/google/bloaty
- https://github.com/tensorflow/tensorflow/tree/master/tensorflow/lite/examples/python
- https://kaggle.com/datasets/sanikamal/rock-paper-scissors-dataset
- https://raspberrypi.com/products/camera-module-3
- https://storage.googleapis.com/mediapipe-tasks/gesture_recognizer/rps_data_sample.zip
- https://tensorflow.org/lite/examples

Feedback is welcome

https://forms.gle/BvSieuMRZbfKFpcq8

Q&A

Thank you!