Fitting AI for Embedded Deployment

Bernhard Suhm, MathWorks
Emelie Andersson, MathWorks
Edge AI innovates many industries!
Hardware Constraints

Memory Footprint / Compute

Memory: 32GB @ 10TFLOPs

Memory: 8GB @ GFLOPs

Memory: 1kB~1MB @ MFLOPs

Power Consumption
What is “Edge” (Embedded) AI?

The chip has only 500 KB memory – make that smaller

How can I make my AI smaller?

Data Scientist

Embedded Software Engineer
Why is Edge AI (Model Compression) difficult?

AI is often big

Knowledge Gap
Model Compression Workflow

1. Determine Hardware Constraints
   - Select Model
2. Simplify Model
3. Quantize Model Parameters
   - Deploy & Integrate
Compressing Machine Learning
Step 1: Size aware model selection

Accuracy on Complex tasks

Size / Execution Time

Deep Neural Net

Shallow Nets

Ensembles

Kernel SVM

Gaussian Process

Decision Tree

Linear Model
Step 2: Simplify the structure of your model

1. Fewer features
2. Tune size-relevant hyperparameters
3. Maximize accuracy given size constraint
Step 3 Quantize your model

Model

Convert in Fixed-Point Designer
Demo: Embedding AI in an intelligent Hearing Aid

Directional

All Around

0.5 to 256 kB on-chip memory
Functionality for Compressing Machine Learning in MATLAB

1. Classification / Regression Learner
   - Determine Hardware Constraints
   - Select (Initial) Model
   - Simplify Model
     - Select Features
     - Tune Hyperparameters

2. In-App Feature Selection
   - Bayesopt
   - Quantize Model Parameters

3. Fixed Point Designer / Native Simulink Block
   - Deploy & Integrate
Demo: Fit Machine Learning for Intelligent Hearing Aid

Fitting Machine Learning onto Memory-limited Hardware

In the context of building an intelligent hearing aid, this script demonstrates the various methods available to fit machine learning onto memory-limited hardware.

Chips on hearing aids range between a few hundred down to below one kB. We’ll take 50 kB as target for our example.

Load Data

As a starting point, we train an initial machine learning model to classify acoustic scenes, using a subset of the data used in the original example


We are just using the first 100 examples from the training set, with 15 scenes resulting in 1500 data points.

```matlab
% load the subset of acoustic scene data we're using here
load('AcousticScenes-SmallTrain.mat');
c = cvpartition(trainLabels,'HoldOut',0.2);
trainSmall = xTrain(c.training,:);
testSmall = xTrain(c.test,:);
```
Machine Learning Demo

Size Reduction by factor 20

Target Hardware
Size: 50 kB
Compressing Deep Learning
Step 1  Size aware model selection
Step 2 Smart pruning

Remove *unimportant* parts of the network

**Pruning process**

1. Evaluate importance of weights
2. Remove the least important weights
3. Fine Tuning (training)
4. Retrain

**Trained Network**

**Pruned + Retrained Network**
Step 3 Quantize your model
Deep Learning Demo: Scene classification

Classify 10 classes
More difficult problem $\rightarrow$ more complex model
Functionality for Compressing Deep Neural Nets

1. Deep Network Designer

2. Taylor Pruning

3. Deep Network Quantizer

- Determine Hardware Constraints
- Select Model
- Prune Deep Neural Network
- Quantize Model Parameters
- Deploy & Integrate
Step 1: Select Model

Load original trained CNN model and dataset

```matlab
load('trained10classNetwork');
load('data')
```

Note: Sounds have been converted to spectrograms
Deep Learning Demo  Size Reduction by factor 5

- Initial: 4.48 MB
- Pruned: 3.57 MB
- Pruned & Quantized: 0.89 MB
One Codebase – Many Embedded Deployment targets
One Codebase – Many Embedded Deployment targets

Zynq is a registered trademark of Xilinx Corporation
Intel logo is a registered trademark of Intel Corporation
Conclusions

You can fit AI for many applications onto limited hardware

MathWorks tools make fitting AI models on constrained hardware a lot easier

Same high-level Workflow for any type of AI

Which constraints are most challenging for your application?
Learn More

To get your started:

- Learn about Embedded Deployment
- Quantization of classification SVM (Doc)
- Deploy Hand-Gesture Classifier onto Arduino (Doc)
- Generate C/C++ Code from Simulink (Video)
- Quantizing a Deep Neural Network (Video)
- Card to Classify Blood Type (IDNEO)
Smaller Models are often Better!

www.linkedin.com/in/emelie-andersson-ai
www.linkedin.com/in/bernhard-suhm-ai