# CLASSIFICATION OF ENVIRONMENTAL SOUND USING IOT SENSORS

Jon Nordby jon@soundsensing.no

November 19, 2019

soundsensing

# INTRODUCTION

# JON NORDBY

Internet of Things specialist

- B.Eng in **Electronics**
- 9 years as **Software** developer. **Embedded** + **Web**
- M. Sc in **Data** Science

Now:

- CTO at Soundsensing
- Machine Learning Consultant


soundsensing

# What we do

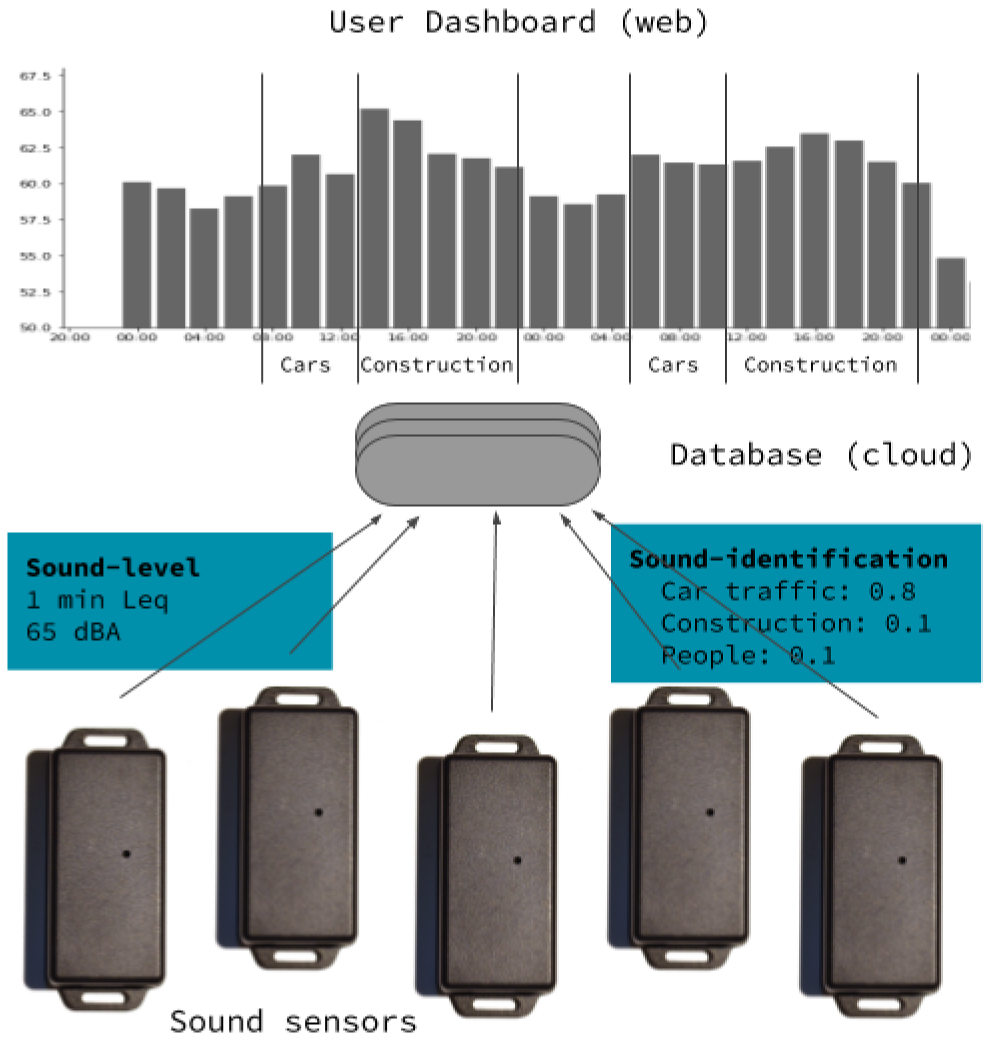
---

Noise Monitoring for outdoor and indoor environments

- Smart cities
- Workspaces/offices
- Hotels/AirBnB
- Music venues

**Designed for Privacy:**
Sound is not transmitted or stored.
- only the noise information

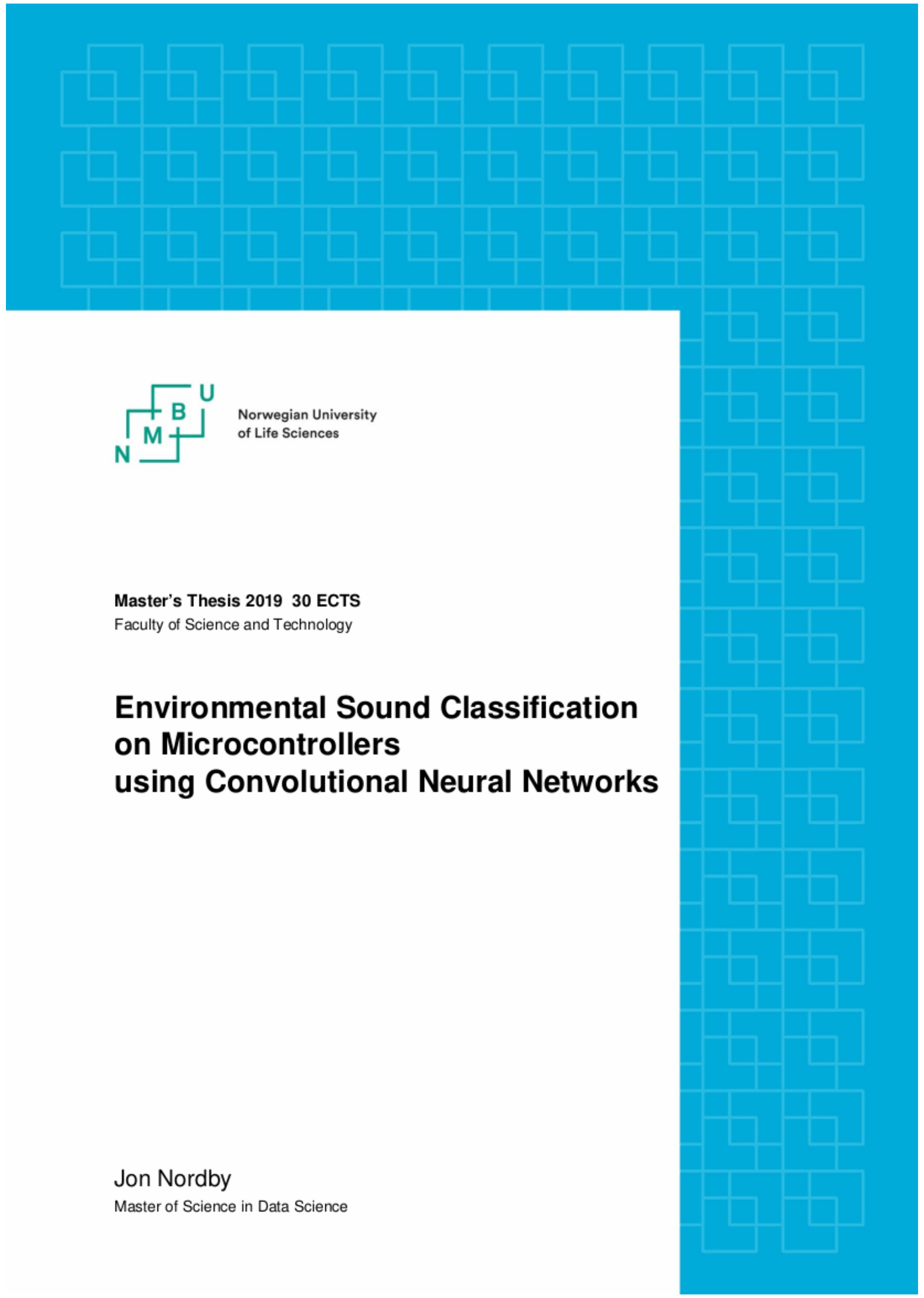
**Our R&D focus**
How to provide the best information possible about noise
using IoT sensors and machine learning




Pilot projects with customers Now - 2020

# THESIS

*Environmental Sound Classification on Microcontrollers using Convolutional Neural Networks*


Norwegian University
of Life Sciences

Master's Thesis 2019 30 ECTS
Faculty of Science and Technology

**Environmental Sound Classification on Microcontrollers using Convolutional Neural Networks**

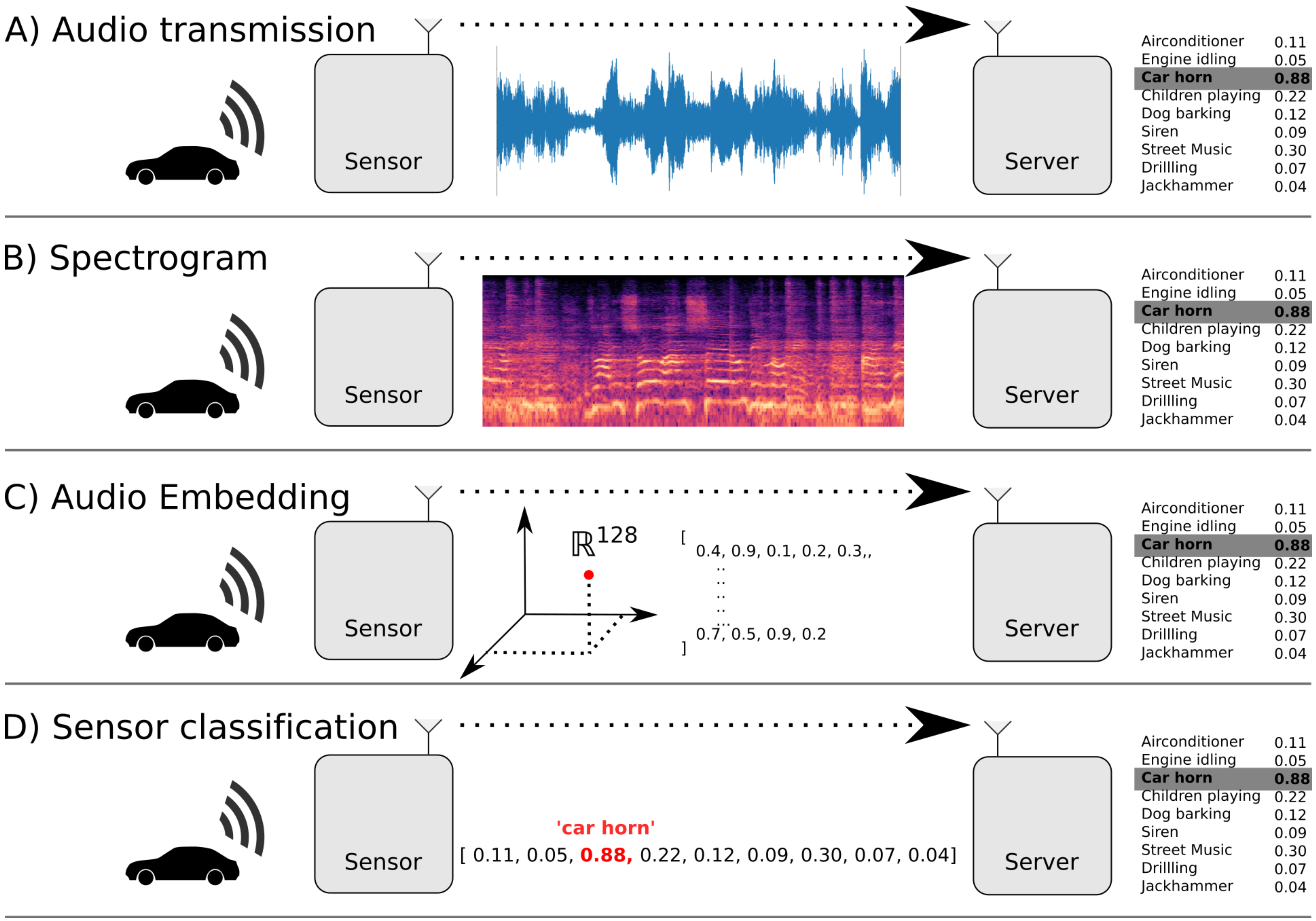
Jon Nordby
Master of Science in Data Science

Report & Code: https://github.com/jonnor/ESC-CNN-microcontroller

# WIRELESS SENSOR NETWORKS

- Want: Wide and dense coverage
- Need: Sensors need to be low-cost
- **Opportunity**: Wireless reduces costs
- **Challenge**: Power consumption

# SENSOR NETWORK ARCHITECTURES



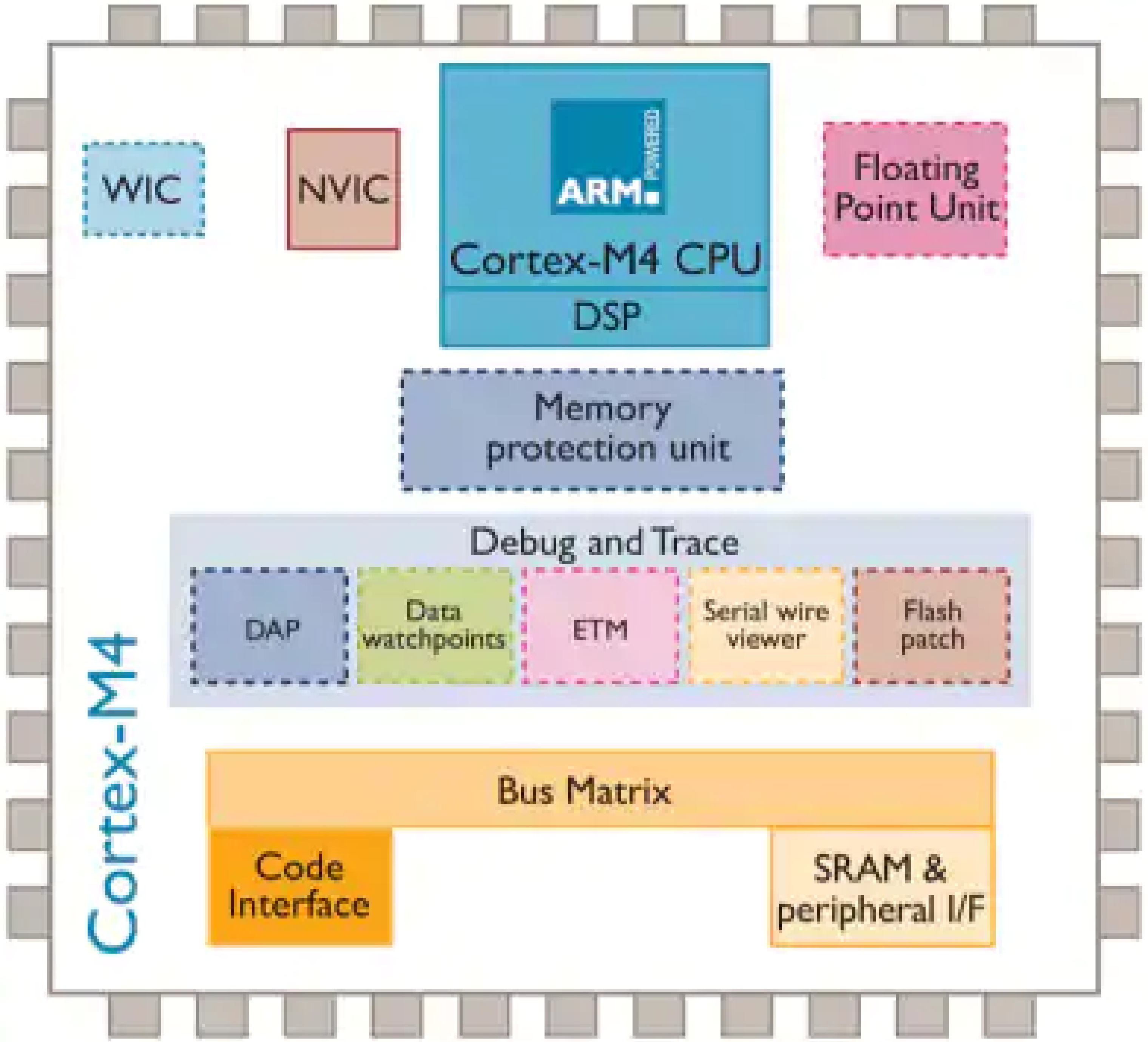
# AUDIO MACHINE LEARNING ON LOW-POWER SENSORS

# WHAT DO YOU MEAN BY LOW-POWER?

Want: 1 year lifetime for palm-sized battery

Need: `<1mW` system power

# GENERAL PURPOSE MICROCONTROLLER

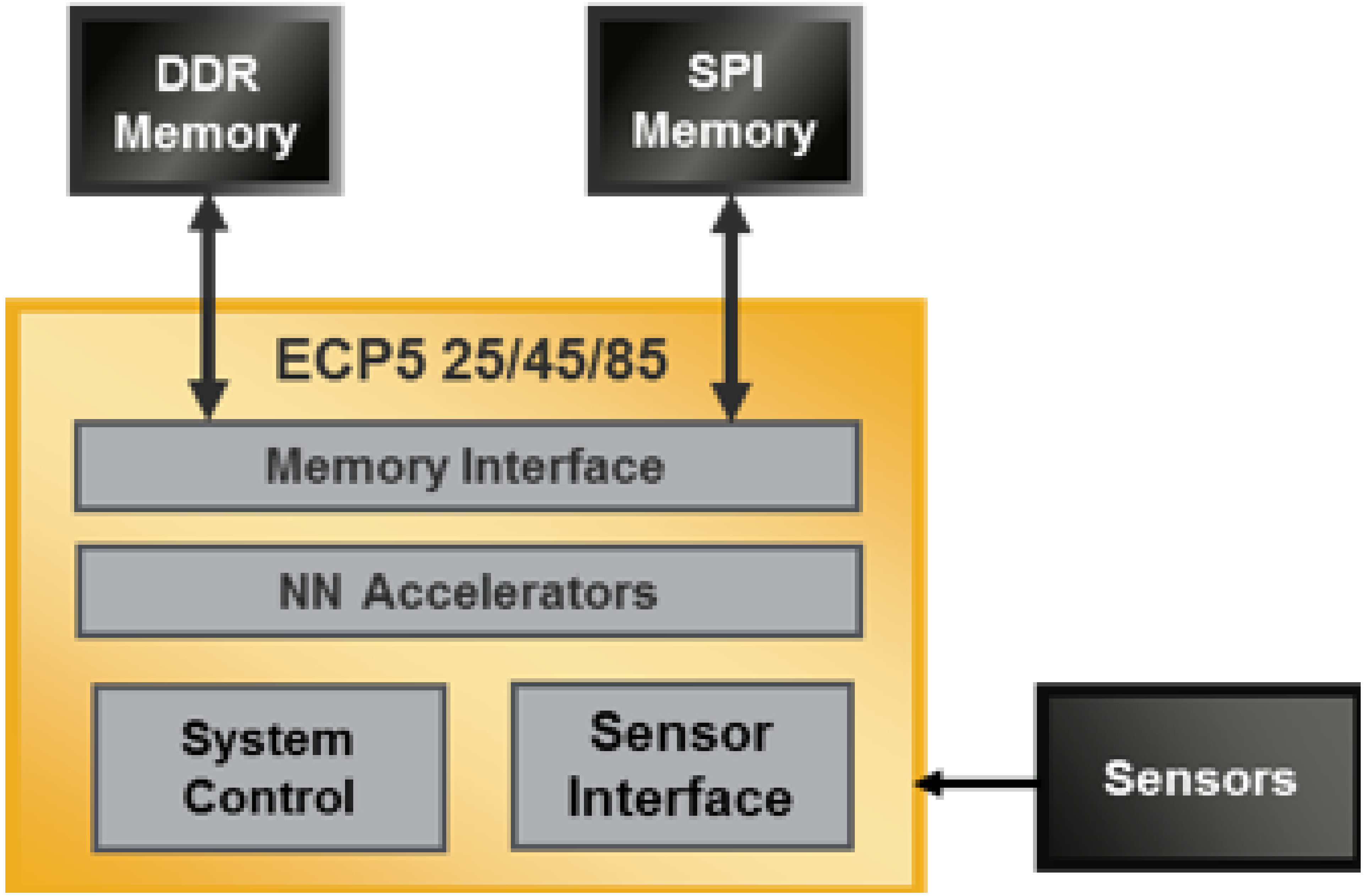



STM32L4 @ 80 MHz. Approx **10 mW**.

- TensorFlow Lite for Microcontrollers (Google)
- ST X-CUBE-AI (ST Microelectronics)

soundsensing

# FPGA



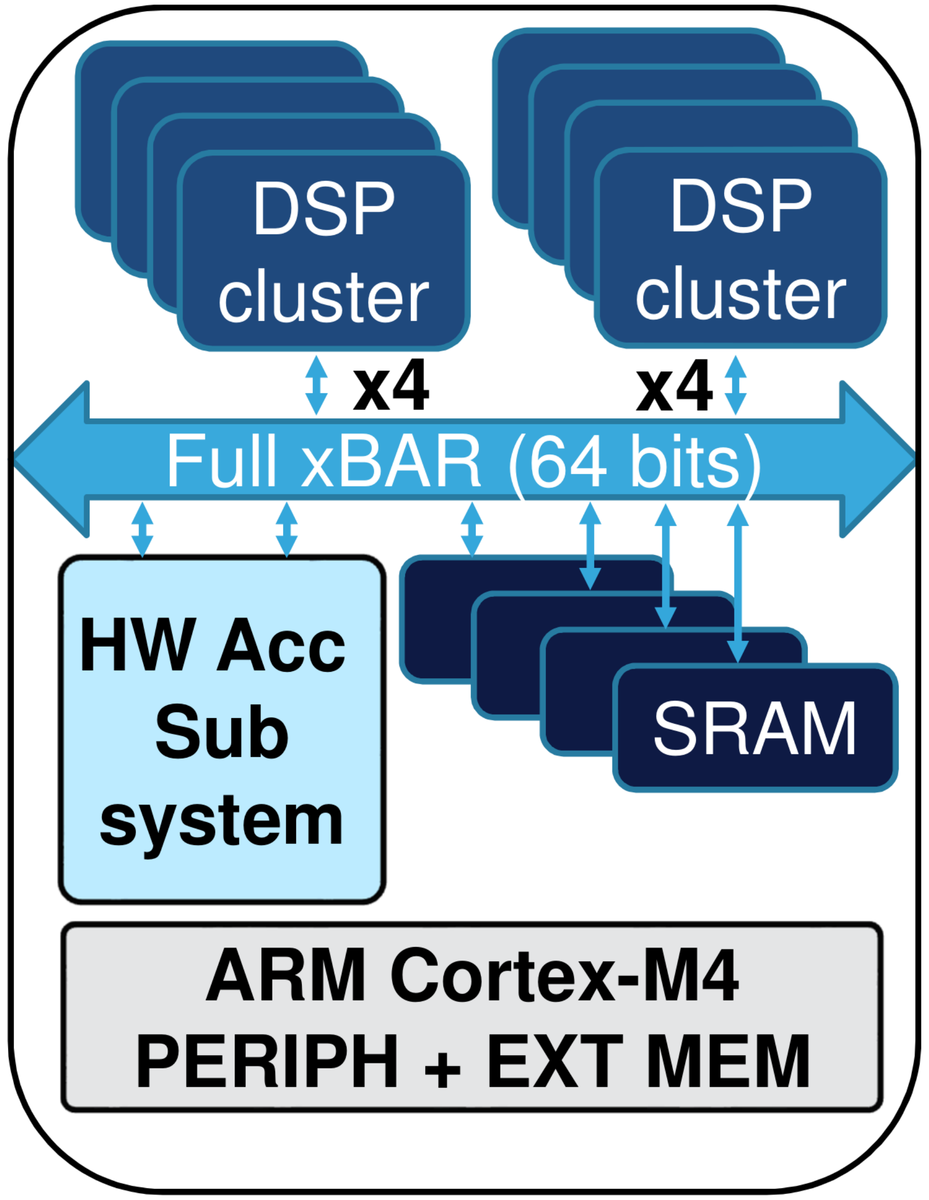

Lattice ICE40 UltraPlus with Lattice sensAI

Human presence detection. VGG8 on 64x64 RGB image, 5 FPS: 7 mW.

Audio ML approx **1 mW**

soundsensing

# NEURAL NETWORK CO-PROCESSORS




Project Orlando (ST Microelectronics), expected 2020

2.9 TOPS/W. AlexNet, 1000 classes, 10 FPS. 41 mWatt

Audio models probably **< 1 mWatt**.

soundsensing

# ON-EDGE CLASSIFICATION OF NOISE

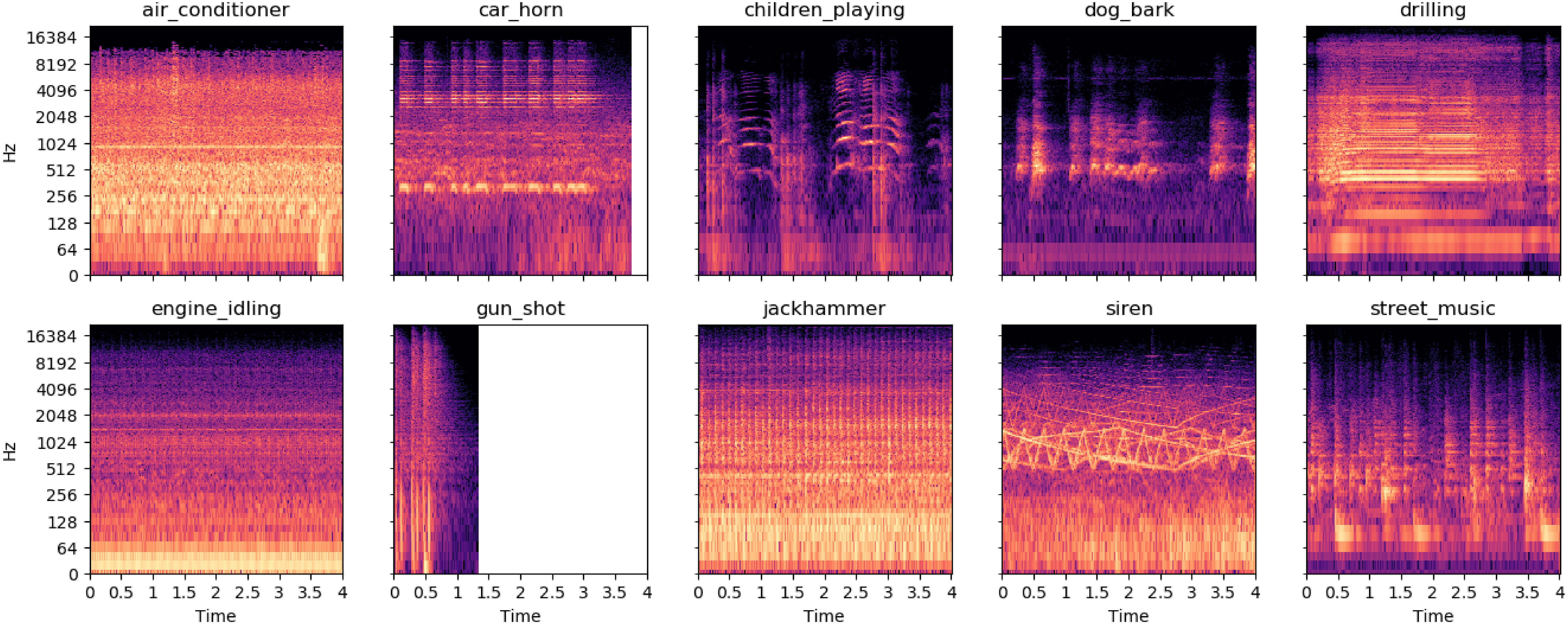
# ENVIRONMENTAL SOUND CLASSIFICATION

*Given an audio signal of environmental sounds,*

*determine which class it belongs to*

- Widely researched. 1000 hits on Google Scholar
- Datasets. Urbansound8k (10 classes), ESC-50, AudioSet (632 classes)
- 2017: Human-level performance on ESC-50

# URBANSOUND8K




soundsensing

# EXISTING WORK

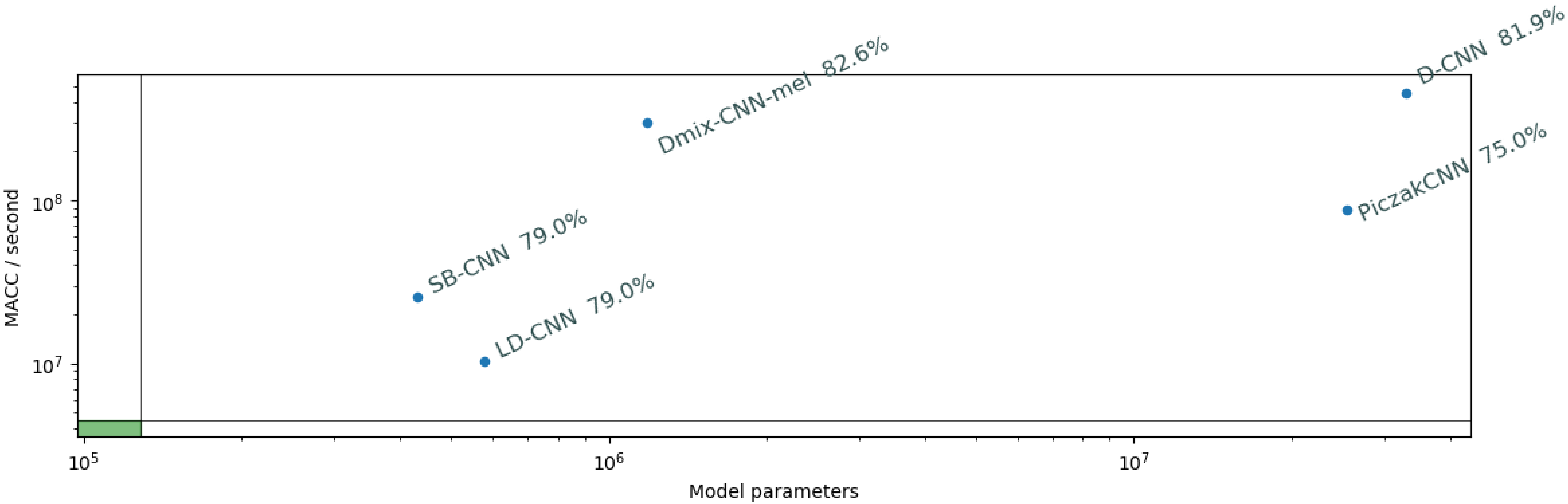
- Convolutional Neural Networks dominate
- Techniques come from image classification
- Mel-spectrogram input standard
- End2end models: getting close in accuracy
- “Edge ML” focused on mobile-phone class HW
- “Tiny ML” (sensors) just starting

# MODEL REQUIREMENTS

With 50% of STM32L476 capacity:

- 64 kB RAM
- 512 kB FLASH memory
- 4.5 M MACC/second

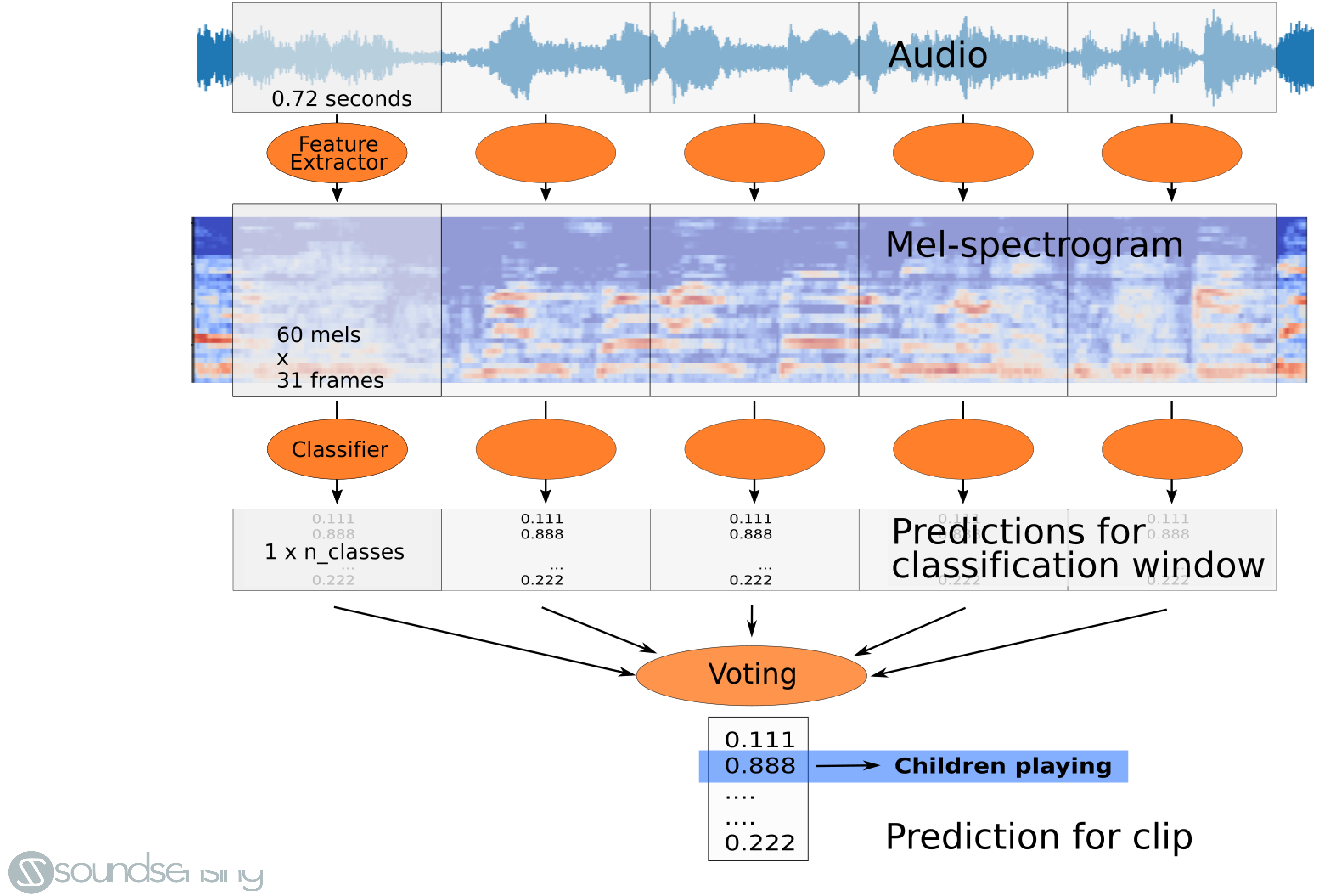
# EXISTING MODELS



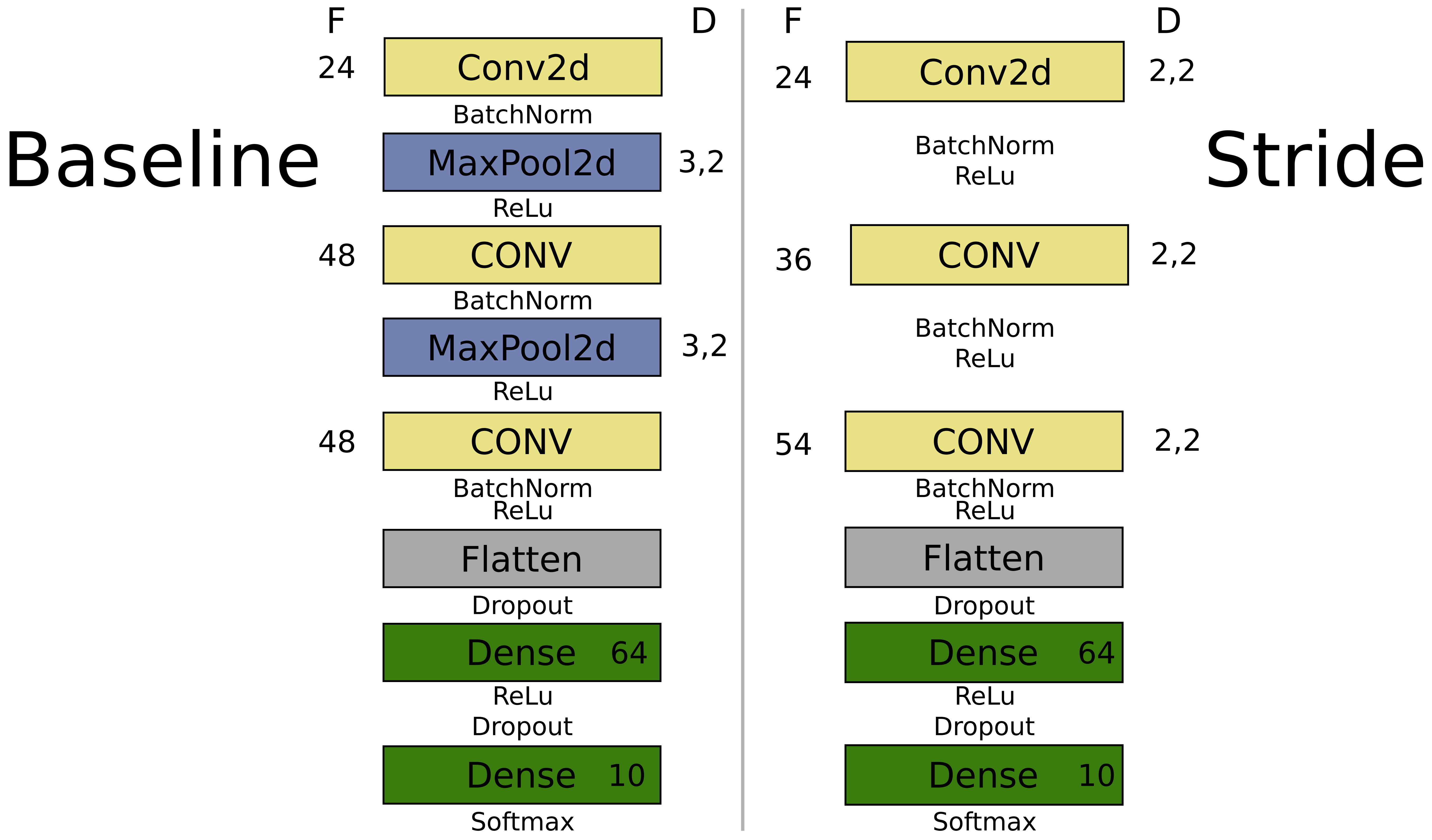

Green: Feasible region

eGRU: running on ARM Cortex-M0 microcontroller, accuracy 61% with **non-standard** evaluation

soundsensing


Audio
0.72 seconds
Feature Extractor
Mel-spectrogram
60 mels
x
31 frames
Classifier
Predictions for classification window
1 x n_classes
0.111
0.888
...
0.222
Voting
0.111
0.888
....
....
0.222
Children playing
Prediction for clip
soundsensing

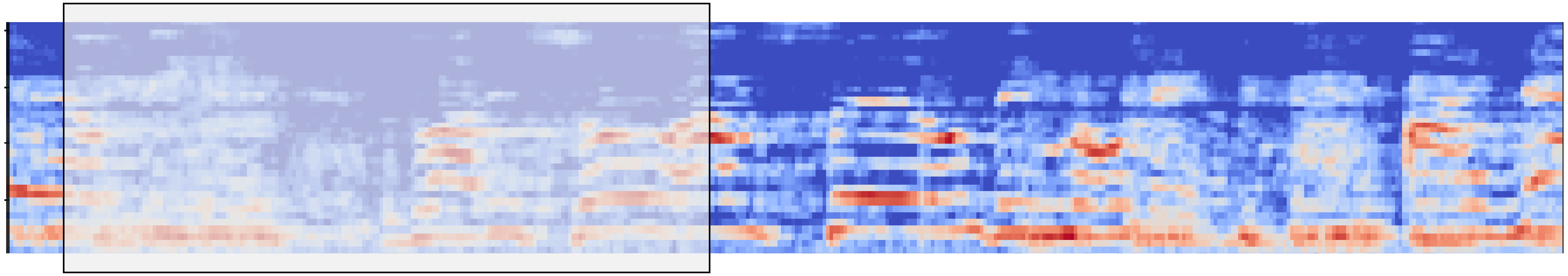
# MODELS


Baseline
F
D
24
Conv2d
BatchNorm
MaxPool2d
3,2
ReLu
48
CONV
BatchNorm
MaxPool2d
3,2
ReLu
48
CONV
BatchNorm
ReLu
Flatten
Dropout
Dense
64
ReLu
Dropout
Dense
10
Softmax
Stride
F
D
24
Conv2d
2,2
BatchNorm
ReLu
36
CONV
2,2
BatchNorm
ReLu
54
CONV
2,2
BatchNorm
ReLu
Flatten
Dropout
Dense
64
ReLu
Dropout
Dense
10
Softmax

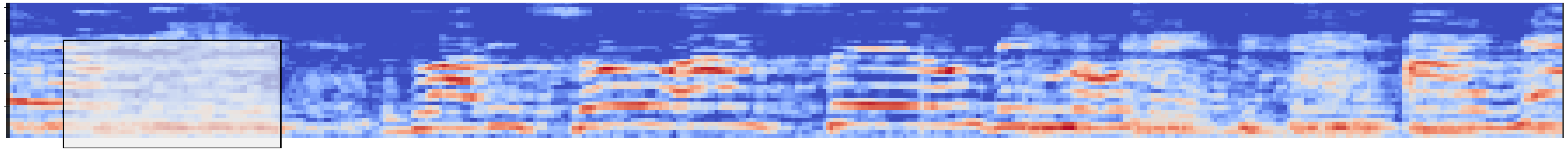
# STRATEGIES FOR SHRINKING CONVOLUTIONAL NEURAL NETWORK

# REDUCE INPUT DIMENSIONALITY

44.1kHz, 2 seconds, 128x128



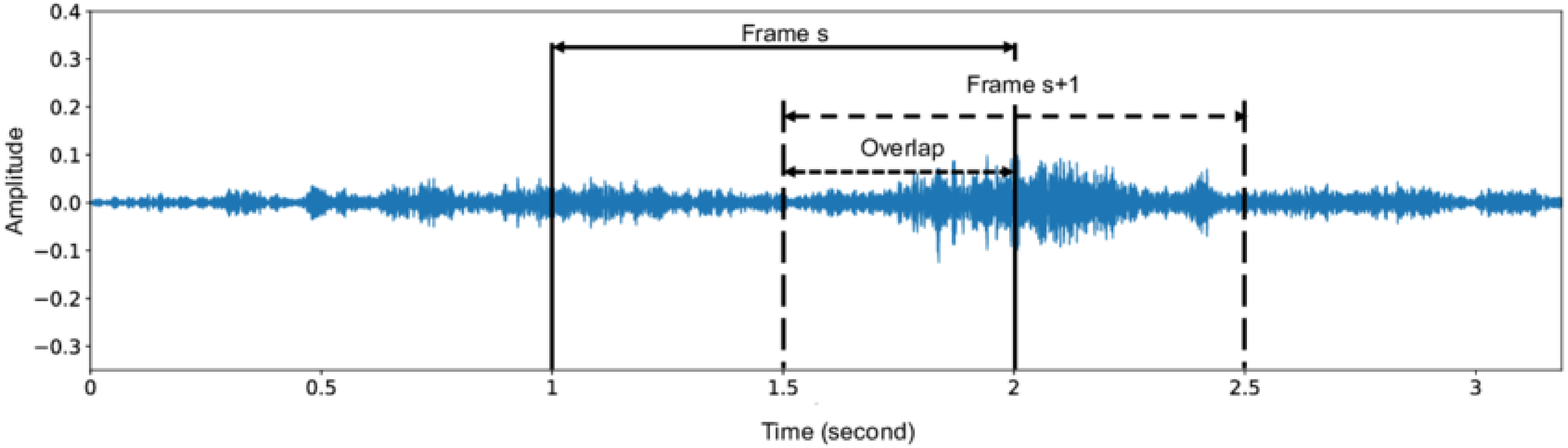
16kHz, 0.75 seconds, 32x32



- Lower frequency range
- Lower frequency resolution
- Lower time duration in window
- Lower time resolution

soundsensing

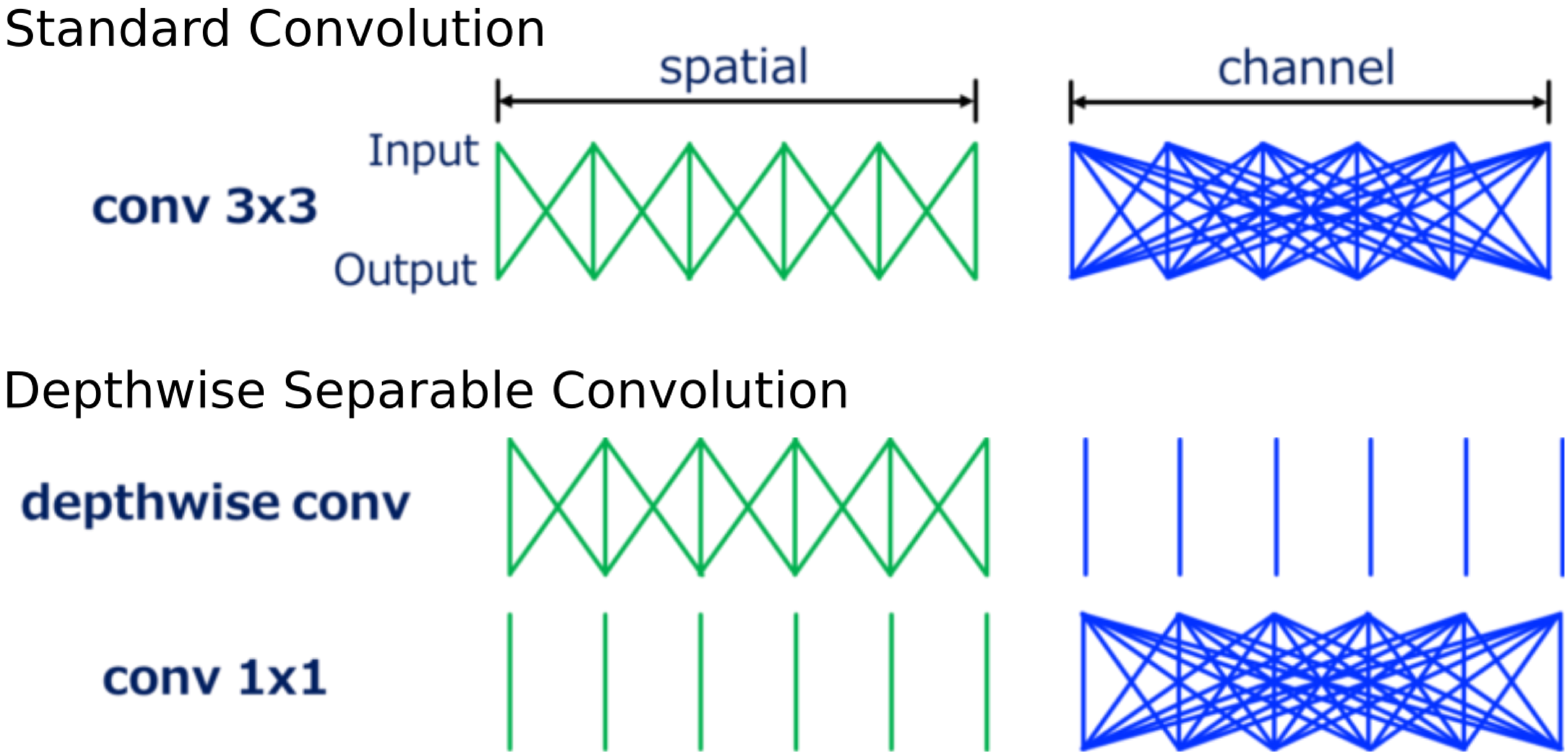
# REDUCE OVERLAP




Models in literature use 95% overlap or more. 20x penalty in inference time!

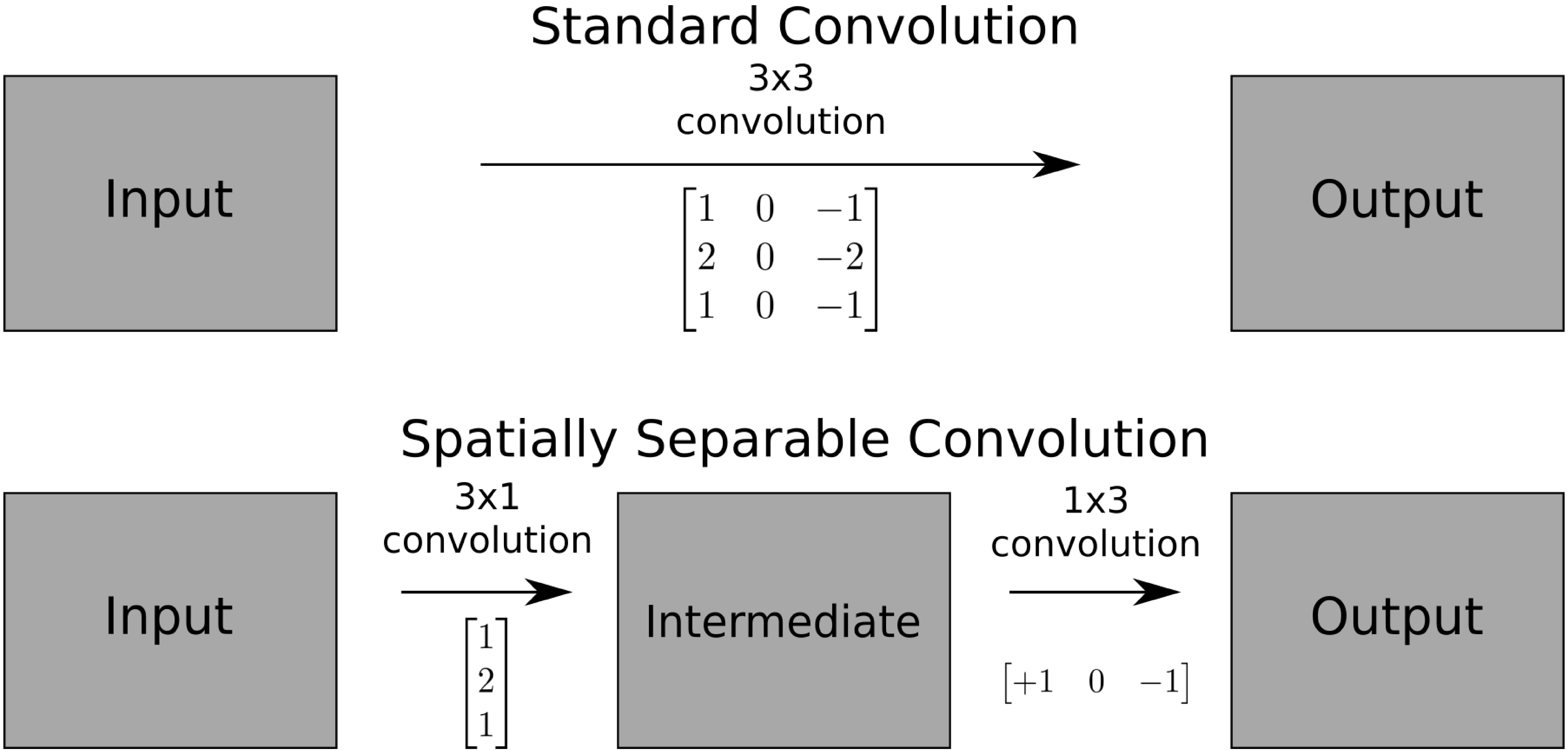
Often low performance benefit. Use 0% (1x) or 50% (2x).

# DEPTHWISE-SEPARABLE CONVOLUTION



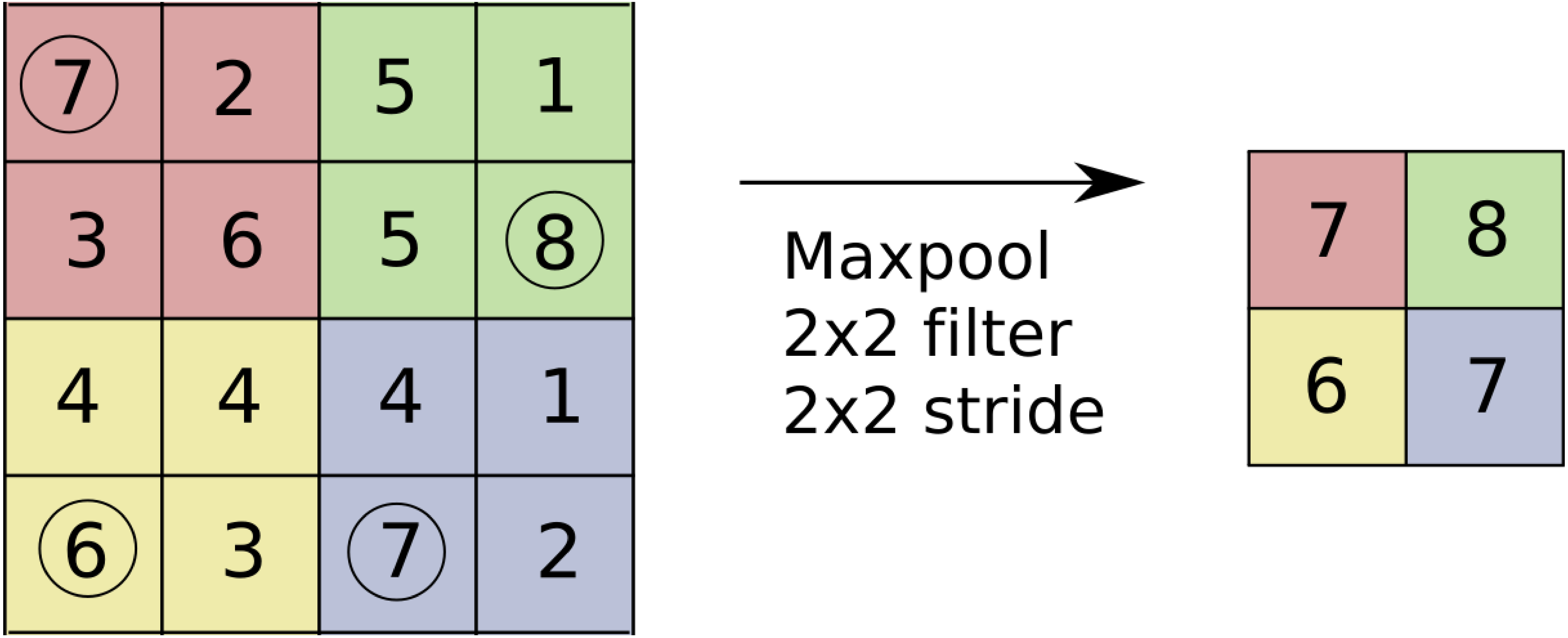

MobileNet, “Hello Edge”, AclNet. 3x3 kernel,64 filters: 7.5x speedup

# SPATIALLY-SEPARABLE CONVOLUTION



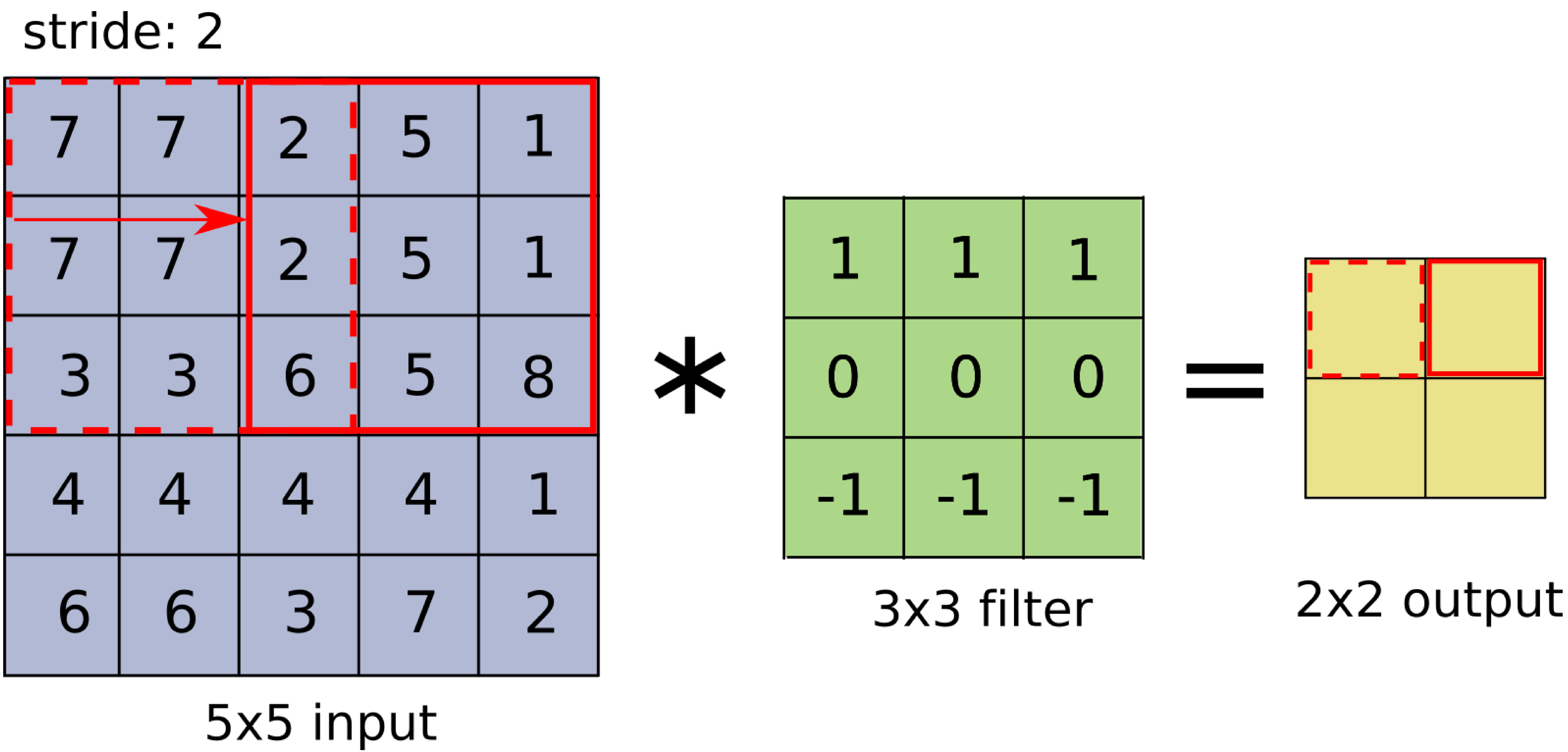

EffNet, LD-CNN. 5x5 kernel: 2.5x speedup

# DOWNSAMPLING USING MAX-POOLING



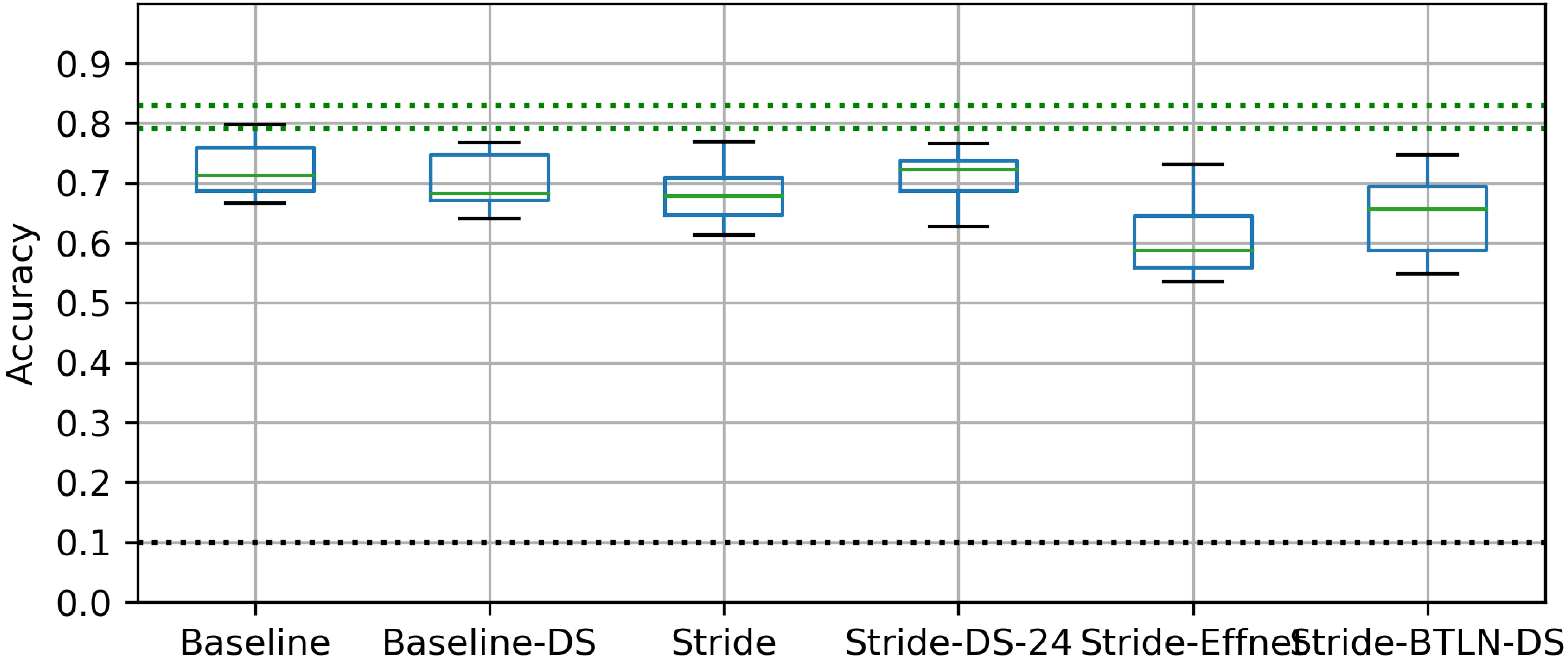

Wasteful? Computing convolutions, then throwing away 3/4 of results!

# DOWNSAMPLING USING STRIDED CONVOLUTION



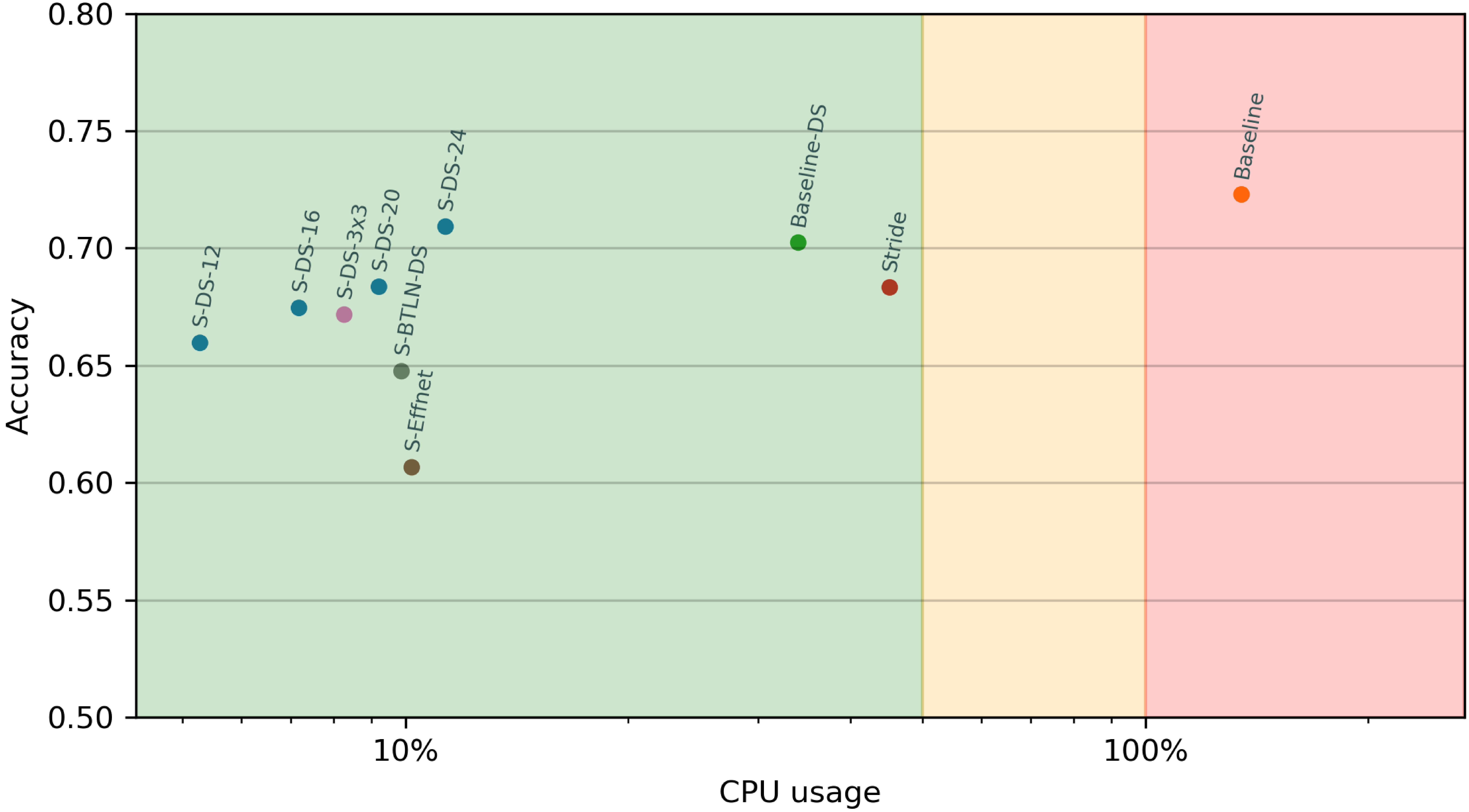

Striding means fewer computations and "learned" downsampling

# MODEL COMPARISON




soundsensing

# PERFORMANCE VS COMPUTE



Accuracy
CPU usage
0.80
0.75
0.70
0.65
0.60
0.55
0.50
10%
100%
S-DS-12
S-DS-16
S-DS-3x3
S-DS-20
S-BTLN-DS
S-DS-24
S-Effnet
Baseline-DS
Stride
Baseline

# QUANTIZATION

Inference can often use 8 bit integers instead of 32 bit floats

- 1/4 the size for weights (FLASH) and activations (RAM)
- 8bit **SIMD** on ARM Cortex M4F: 1/4 the inference time
- Supported in X-CUBE-AI 4.x (July 2019)

# CONCLUSIONS

- Able to perform Environmental Sound Classification at ~ `10mW` power,
- Using *general purpose microcontroller*, ARM Cortex M4F
- Best performance: 70.9% mean accuracy, under 20% CPU load
- Highest reported Urbansound8k on microcontroller (over eGRU 62%)
- Best architecture: Depthwise-Separable convolutions with striding
- Quantization enables 4x bigger models (and higher perf)
- With dedicated Neural Network Hardware

# FURTHER RESEARCH

# WAVEFORM INPUT TO MODEL

- Preprocessing. Mel-spectrogram: **60** milliseconds
- CNN. Stride-DS-24: **81** milliseconds
- With quantization, spectrogram conversion is the bottleneck!
- Convolutions can be used to learn a Time-Frequency transformation.

Can this be faster than the standard FFT? And still perform well?

# ON-SENSOR INFERENCE CHALLENGES

- Reducing power consumption. Adaptive sampling
- Efficient training data collection in WSN. Active Learning?
- Real-life performance evaluations. Out-of-domain samples

# WRAPPING UP

# SUMMARY

- Noise pollution is a growing problem
- Wireless Sensor Networks can used to quantify
- Noise Classification can provide more information
- Want high density of sensors. Need to be low cost
- On-sensor classification desirable for power/cost and privacy

# MORE RESOURCES

Machine Hearing. ML on Audio

- github.com/jonnor/machinehearing

Machine Learning for Embedded / IoT

- github.com/jonnor/embeddedml

Thesis Report & Code

- github.com/jonnor/ESC-CNN-microcontroller

# QUESTIONS

?

Email: jon@soundsensing.no

# COME TALK TO ME!

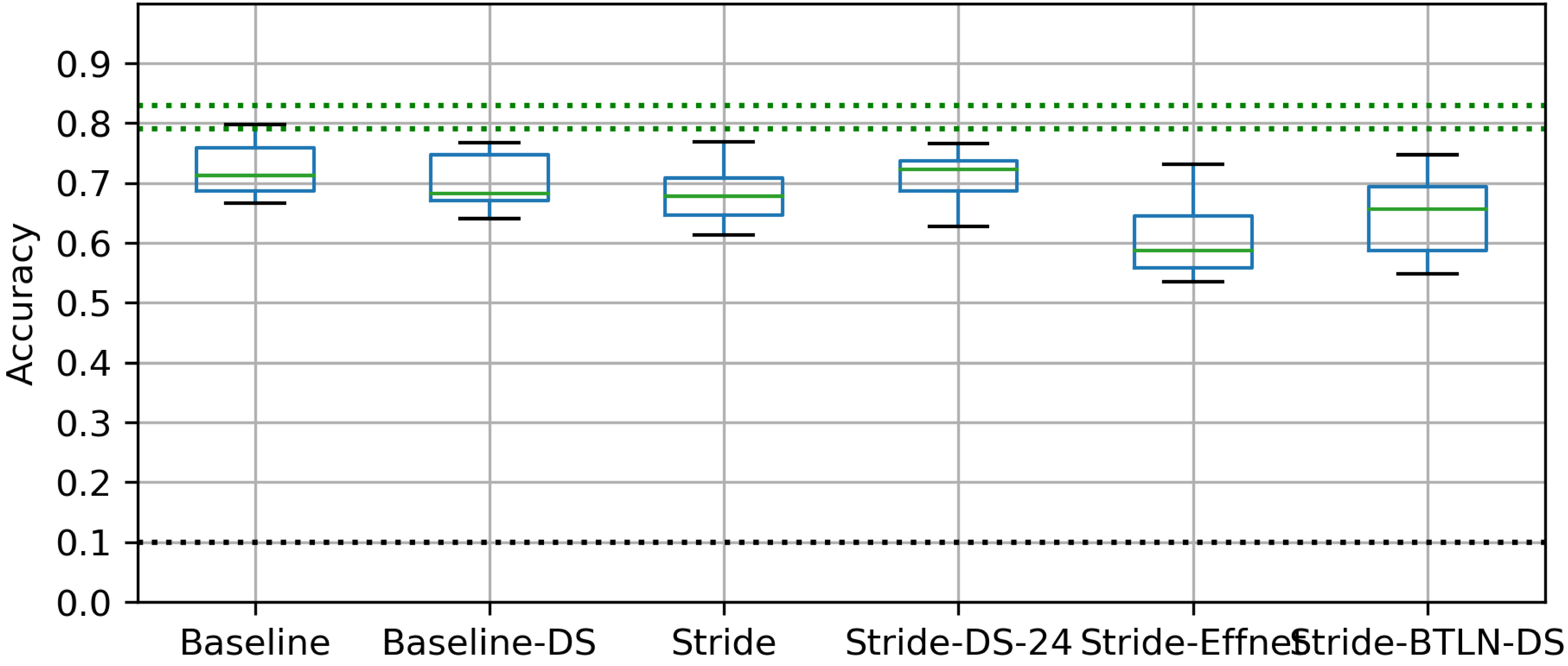
- Noise Monitoring sensors. Pilot projects for 2020?
- Environmental Sound, Wireless Sensor Networks for Audio. Research partnering?
- “On-edge” / Embedded Device ML. Happy to advise!

Email: jon@soundsensing.no

soundsensing

# THESIS RESULTS

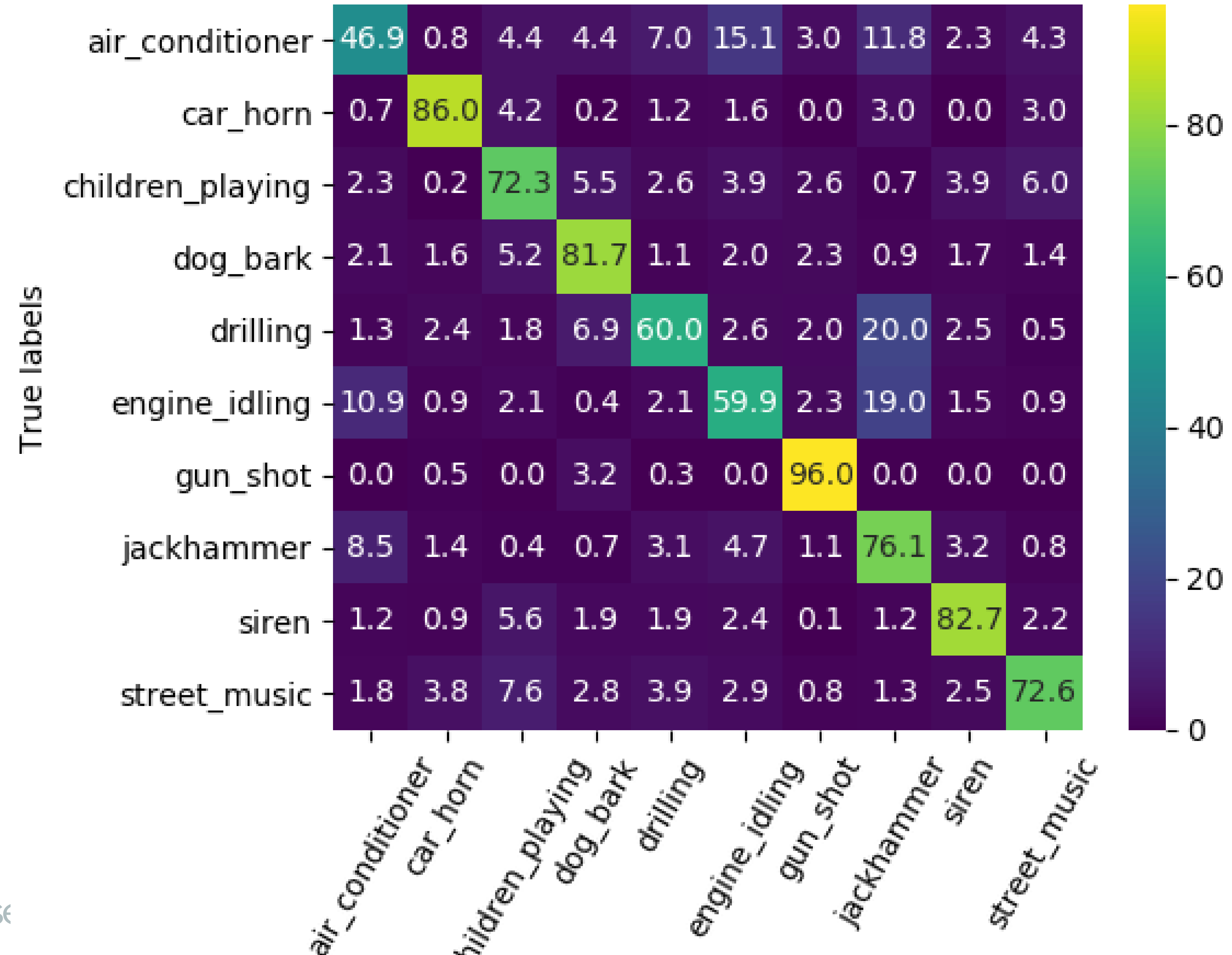
# MODEL COMPARISON




soundsensing

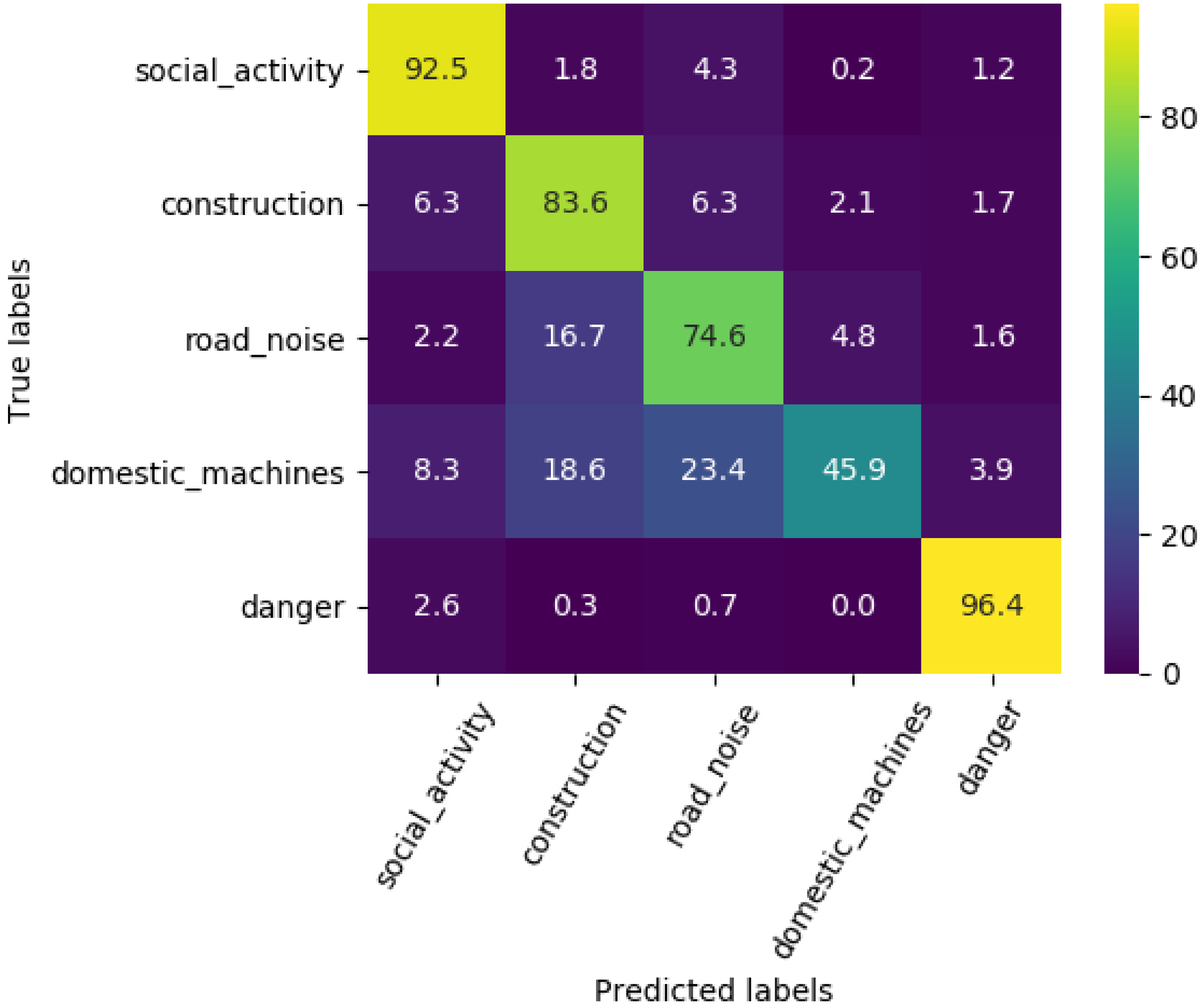
# LIST OF RESULTS

| Model | CPU use | Accuracy | FG Accuracy | BG Accuracy |
|---|---|---|---|---|
| Baseline | 971 ms | 72.3% ±4.6 | 78.3% ±7.1 | 60.5% ±7.7 |
| Baseline-DS | 244 ms | 70.2% ±4.7 | 76.1% ±7.5 | 58.6% ±8.2 |
| Stride | 325 ms | 68.3% ±5.2 | 74.1% ±6.6 | 56.6% ±8.0 |
| Stride-BTLN-DS | 71 ms | 64.8% ±7.1 | 69.5% ±8.2 | 55.3% ±8.9 |
| Stride-DS-12 | 38 ms | 66.0% ±6.0 | 72.6% ±6.5 | 53.3% ±9.1 |
| Stride-DS-16 | 51 ms | 67.5% ±5.6 | 73.3% ±7.7 | 56.2% ±8.3 |
| Stride-DS-20 | 66 ms | 68.4% ±5.2 | 75.0% ±7.4 | 55.2% ±10.0 |
| Stride-DS-24 | 81 ms | 70.9% ±4.3 | 75.8% ±6.3 | 61.8% ±6.8 |
| Stride-DS-3x3 | 59 ms | 67.2% ±6.5 | 73.0% ±7.4 | 55.8% ±9.1 |
| Stride-Effnet | 73 ms | 60.7% ±6.6 | 66.9% ±7.9 | 48.7% ±8.3 |

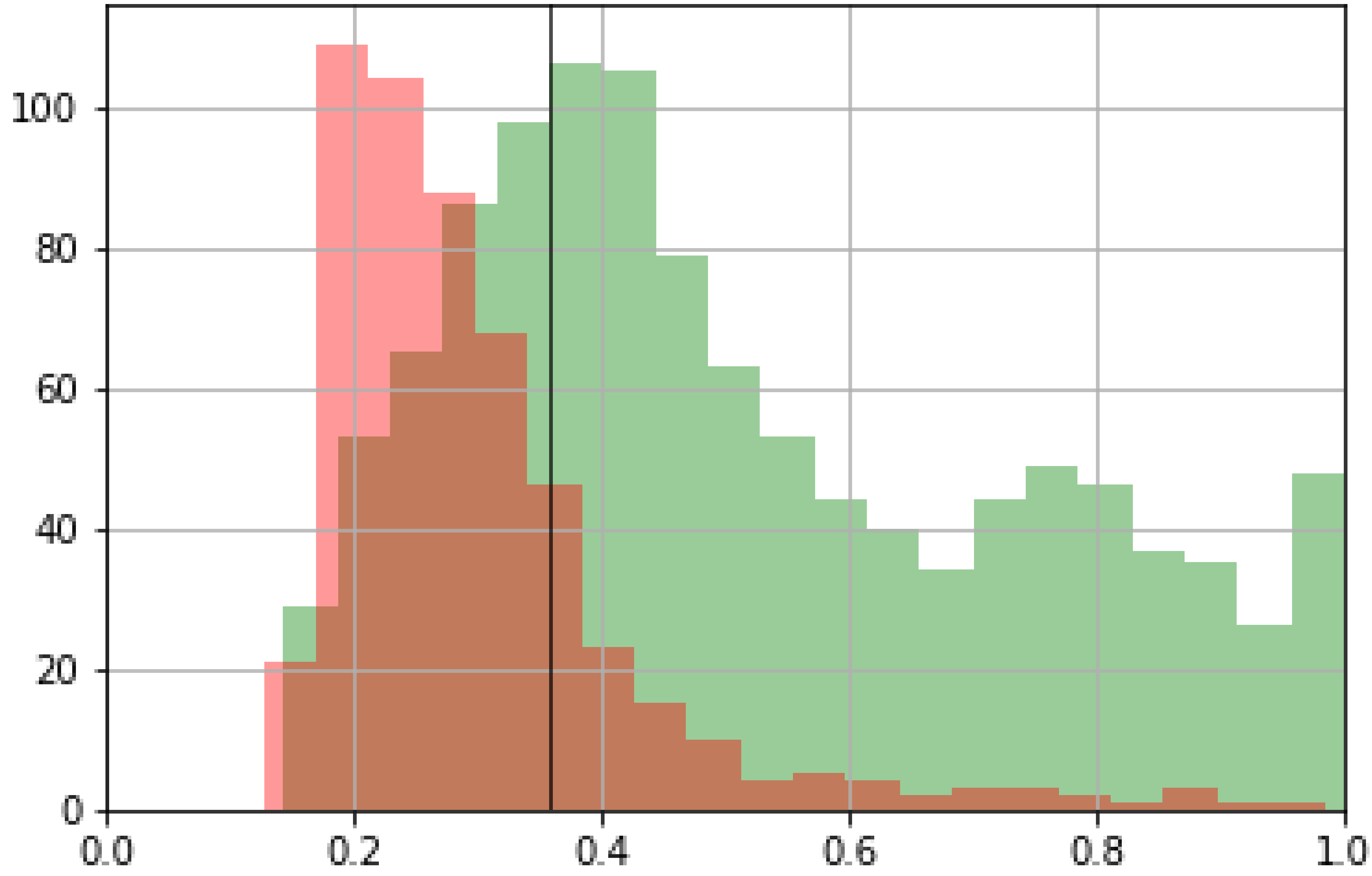

True labels
air_conditioner 46.9 0.8 4.4 4.4 7.0 15.1 3.0 11.8 2.3 4.3
car_horn 0.7 86.0 4.2 0.2 1.2 1.6 0.0 3.0 0.0 3.0
children_playing 2.3 0.2 72.3 5.5 2.6 3.9 2.6 0.7 3.9 6.0
dog_bark 2.1 1.6 5.2 81.7 1.1 2.0 2.3 0.9 1.7 1.4
drilling 1.3 2.4 1.8 6.9 60.0 2.6 2.0 20.0 2.5 0.5
engine_idling 10.9 0.9 2.1 0.4 2.1 59.9 2.3 19.0 1.5 0.9
gun_shot 0.0 0.5 0.0 3.2 0.3 0.0 96.0 0.0 0.0 0.0
jackhammer 8.5 1.4 0.4 0.7 3.1 4.7 1.1 76.1 3.2 0.8
siren 1.2 0.9 5.6 1.9 1.9 2.4 0.1 1.2 82.7 2.2
street_music 1.8 3.8 7.6 2.8 3.9 2.9 0.8 1.3 2.5 72.6
air_conditioner
car_horn
children_playing
dog_bark
drilling
engine_idling
gun_shot
jackhammer
siren
street_music
80
60
40
20
0

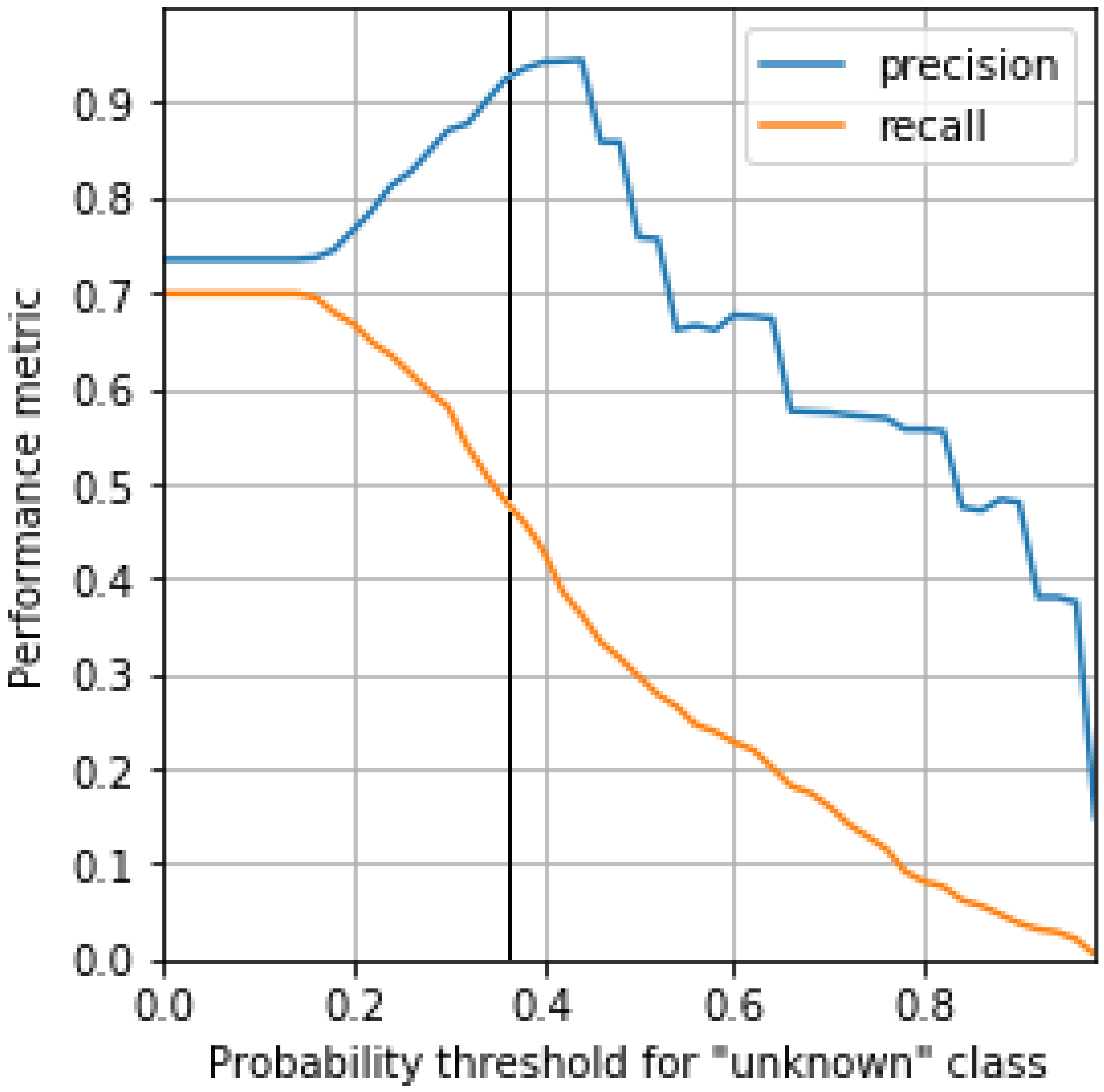
Predicted labels

# GROUPED CLASSIFICATION


social_activity
construction
road_noise
domestic_machines
danger
True labels
92.5
1.8
4.3
0.2
1.2
6.3
83.6
6.3
2.1
1.7
2.2
16.7
74.6
4.8
1.6
8.3
18.6
23.4
45.9
3.9
2.6
0.3
0.7
0.0
96.4
80
60
40
20
0
social_activity
construction
road_noise
domestic_machines
danger
Predicted labels

# UNKNOWN CLASS




soundsensing

# EXPERIMENTAL DETAILS

# ALL MODELS

| Model | Downsample | Convolution | L | F | MACC | RAM | FLASH |
|---|---|---|---|---|---|---|---|
| Baseline | maxpool 3x2 | standard | 3 | 24 | 10185 K | 35 kB | 405 kB |
| Baseline-DS | maxpool 3x2 | DS | 3 | 24 | 1567 K | 55 kB | 96 kB |
| Stride | stride 2x2 | standard | 3 | 22 | 2980 K | 55 kB | 372 kB |
| Stride-BTLN-DS | stride 2x2 | BTLN-DS | 3 | 22 | 445 K | 47 kB | 80 kB |
| Stride-DS-12 | stride 2x2 | DS | 3 | 12 | 208 K | 27 kB | 88 kB |
| Stride-DS-16 | stride 2x2 | DS | 3 | 16 | 291 K | 36 kB | 118 kB |
| Stride-DS-20 | stride 2x2 | DS | 3 | 20 | 380 K | 45 kB | 149 kB |
| Stride-DS-24 | stride 2x2 | DS | 3 | 24 | 477 K | 54 kB | 180 kB |
| Stride-DS-3x3 | stride 2x2 | DS | 4 | 24 | 318 K | 54 kB | 95 kB |
| Stride-Effnet | stride 2x2 | Effnet | 3 | 22 | 468 K | 47 kB | 125 kB |

# METHODS

Standard procedure for Urbansound8k

- Classification problem
- 4 second sound clips
- 10 classes
- 10-fold cross-validation, predefined
- Metric: Accuracy

# TRAINING SETTINGS

| | |
|---|---|
| Samplerate (Hz) | 22050 |
| Melfilter bands | 60 |
| FFT length (samples) | 1024 |
| FFT hop (samples) | 512 |
| Classification window | 31 |
| Minibatch size | 400 |
| Epochs | 100 |
| Training samples/epoch | 30000 |
| Validation samples/epoch | 5000 |
| Learning rate | 0.005 |
| Nesterov momentum | NaN |

# TRAINING

- NVidia RTX2060 GPU 6 GB
- 10 models x 10 folds = 100 training jobs
- 100 epochs
- 3 jobs in parallel
- 36 hours total

# EVALUATION

For each fold of each model

1. Select best model based on validation accuracy
2. Calculate accuracy on test set

For each model

- Measure CPU time on device

# YOUR MODEL WILL TRICK YOU

And the bugs can be hard to spot

# FAIL: INTEGER TRUNCATION

```
    features: Fix integer truncation

    Especially combined with meanstd normalization
    which makes range (-3,3),
    this accidentally removed a lot of details

--------------------------- microesc/features.py ---------------------------
index 9d16d2a..6d9ca59 100644
@@ -152,7 +154,7 @@ def load_sample(sample, settings, feature_dir, window_frames,
     if window_frames is None:
         padded = mels
     else:
-        padded = numpy.full((n_mels, window_frames), 0)
+        padded = numpy.full((n_mels, window_frames), 0.0, dtype=float)
         inp = mels[:, 0:min(window_frames, mels.shape[1])]
         padded[:, 0:inp.shape[1]] = inp
```

soundsensing

# FAIL. DROPOUT LOCATION

```
models: Fix Dropout location

Suprised this was able to train at all before,
as whole classes must have been dropped
Training and validation loss now follow eachother much more closely
```

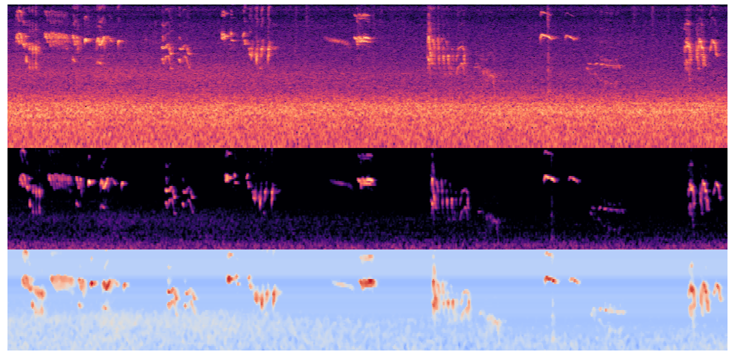
```
---------------------------- microesc/models/sbcnn.py ----------------------------
index 2e95242..3986bcf 100644
@@ -98,12 +98,12 @@ def backend_dense1(x, n_classes, fc=64, regularization=0.001, dropout=0.5):
     """

     x = Flatten()(x)
+    x = Dropout(dropout)(x)
     x = Dense(fc, kernel_regularizer=l2(regularization))(x)
     x = Activation('relu')(x)
-    x = Dropout(dropout)(x)

-    x = Dense(n_classes, kernel_regularizer=l2(regularization))(x)
     x = Dropout(dropout)(x)
+    x = Dense(n_classes, kernel_regularizer=l2(regularization))(x)
     x = Activation('softmax')(x)
     return x
```

# BACKGROUND

# MEL-SPECTROGRAM



# NOISE POLLUTION

Reduces health due to stress and loss of sleep

In Norway

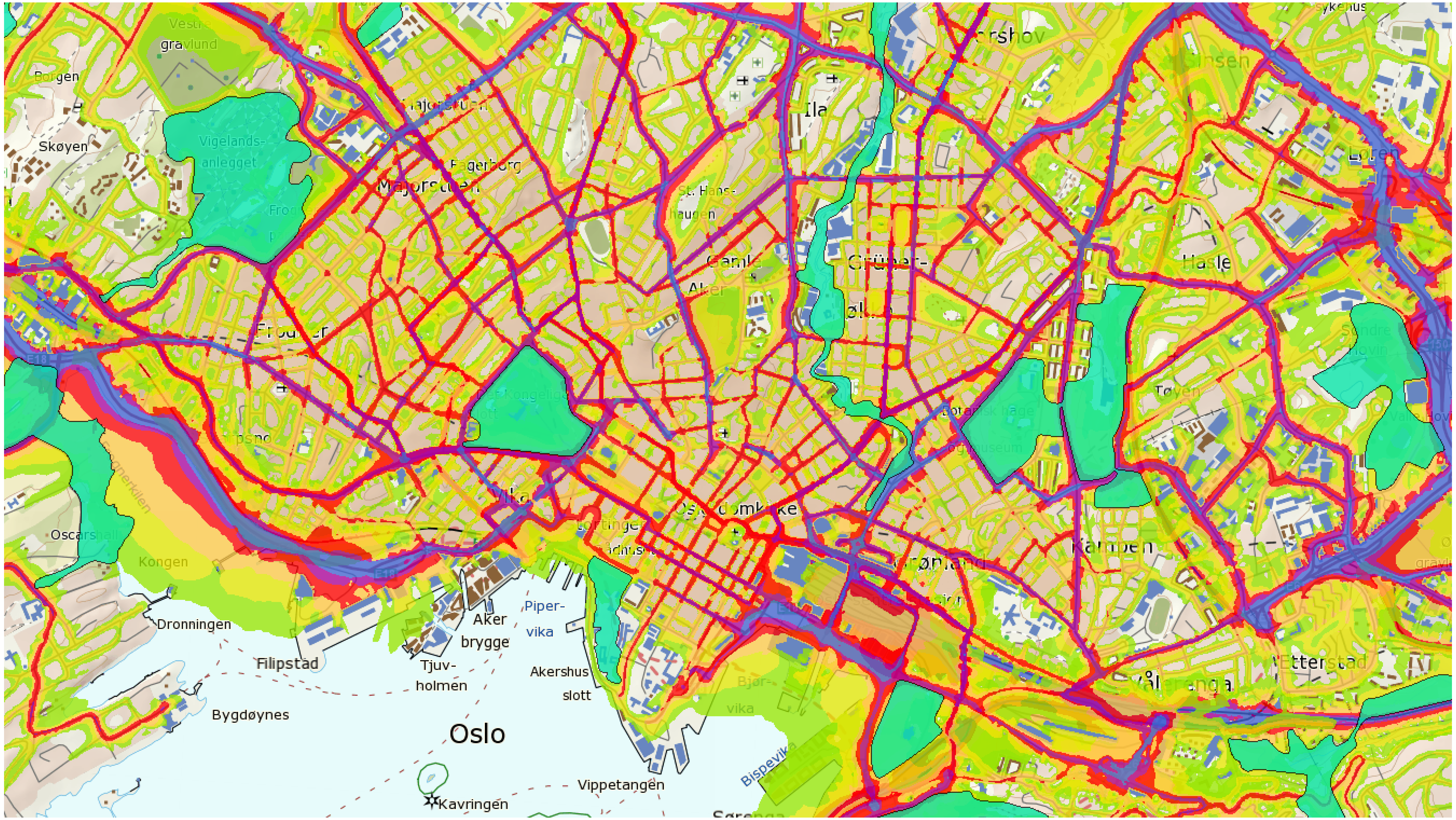
- 1.9 million affected by road noise (2014, SSB)
- 10’000 healty years lost per year (Folkehelseinstituttet)

In Europe

- 13 million suffering from sleep disturbance (EEA)
- 900’000 DALY lost (WHO)

# NOISE MAPPING

Simulation only, no direct measurements




soundsensing