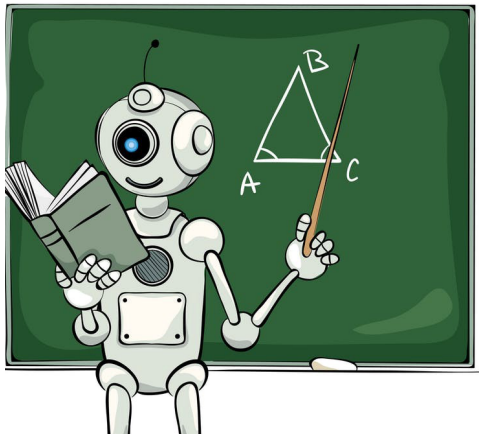

Supervised ML on
ULP Devices
Ingy ElSayed-Aly

# Problem Statement

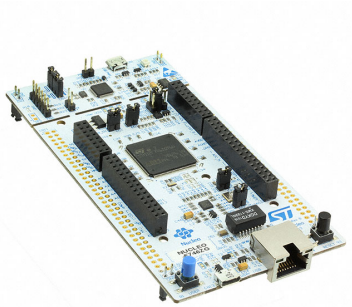
Deploy a Convolutional Neural Network (CNN) inference model for sign language recognition to a Cortex-M processor.

# Motivation

- Machine Learning applications
- Robustness
- Privacy Concerns



# Related Work



**Research**

- Quantization
- Model compression
- Jetson TX1

**CMSIS NN**

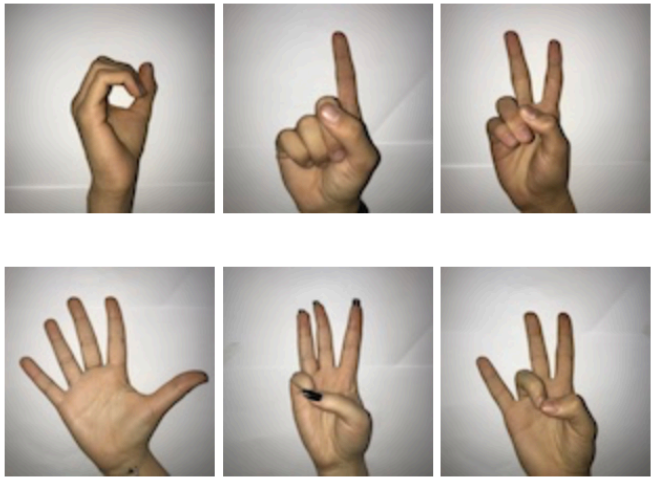
- CIFAR 10 problem
- Cortex-M7




# Workflow for deployment

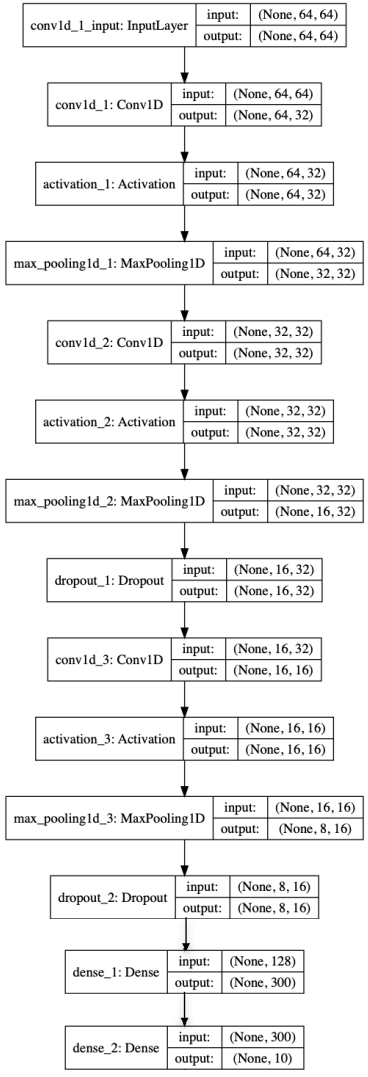
# Dataset

Original Dataset: 64 x 64 x 3

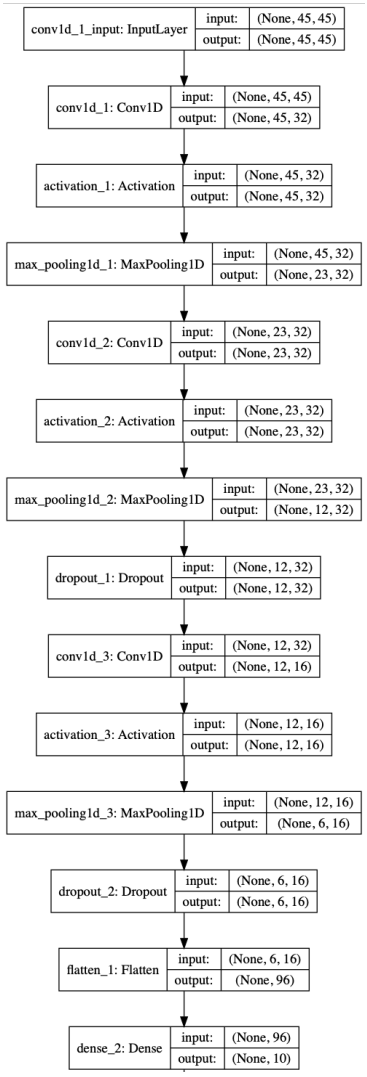


Transformed Dataset: 45 x 45 x 1

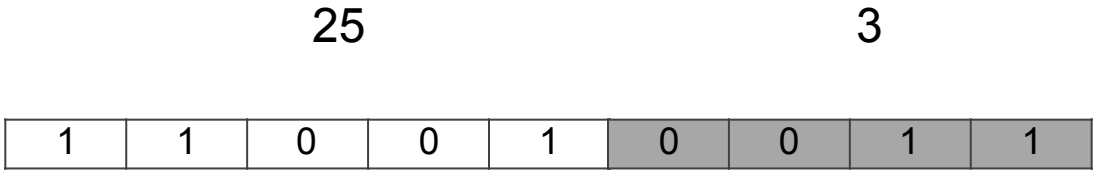
# Model Architecture




| Layer | Type | Kernel Shape | Output Shape |
|---|---|---|---|
| Input | - | - | $45 \times 45 \times 1$ |
| Layer 1 | Convolution | $5 \times 5 \times 32$ | $45 \times 45 \times 32$ |
| Layer 2 | Max Pooling | $2 \times 2$ | $23 \times 23 \times 32$ |
| Layer 3 | Convolution | $5 \times 5 \times 32$ | $23 \times 23 \times 32$ |
| Layer 4 | Max Pooling | $2 \times 2$ | $12 \times 12 \times 32$ |
| Layer 5 | Convolution | $5 \times 5 \times 16$ | $12 \times 12 \times 16$ |
| Layer 6 | Max Pooling | $2 \times 2$ | $6 \times 6 \times 16$ |
| Layer 7 | Dense | $6 \times 6 \times 16 \times 10$ | 10 |



# Fixed Point Quantization




200.305 ~ 25 x 2 ^ 3

# Conversion to C/C++

arm_sign_nb_inputs.h

arm_sign_nb_parameter.h

arm_sign_nb_weights.h

arm_sign_nb.cpp

```
#define CONV1_IM_DIM 45
#define CONV1_IM_CH 1
#define CONV1_KER_DIM 5
#define CONV1_PADDING 2
#define CONV1_STRIDE 1
#define CONV1_OUT_CH 32
#define CONV1_OUT_DIM 45
```

## Demo

```
0: 0
1: 0
2: 0
3: 0
4: 42
5: 0
6: 0
7: 42
8: 42
9: 0
```

# Limitations

- Unclear quantization method
- Unclear weight ordering
- Runs but classification is wrong

# Thank you

# References

**Images:**


- http://parneetk.github.io/images/2017-01-23-cnn-cifar10_files/2017-01-23-cnn-cifar10_8_0.png
- https://media.digikey.com/photos/STMicro%20Photos/NUCLEO-F746ZG_Top.jpg
- https://images.anandtech.com/doci/9779/Jetson_TX1_Press_Deck_Final-page-006.jpg
- https://towardsdatascience.com/the-4-machine-learning-skills-you-wont-learn-in-school-or-moocs-c641cae24f1f
- https://start.lesechos.fr/actu-entreprises/technologie-digital/comprendre-l-internet-des-objets-iot-en-5-questions-12253.php
- https://www.kaggle.com/ardamavi/sign-language-digits-dataset/home

# References


- [1] F. N. Iandola and K. Keutzer, “Keynote: Small neural nets are beautiful: Enabling embedded systems with small deep-neural-network architectures,” CoRR, vol. abs/1710.02759, 2017. [Online]. Available: http://arxiv.org/abs/1710.02759
- [2] “Krebsonsecurity,” https://krebsonsecurity.com/tag/mirai-botnet/, accessed: 2018-09-20.
- [3] M. Malik and H. Homayoun, “Big data on low power cores: Are low power embedded processors a good fit for the big data workloads?” in 2015 33rd IEEE International Conference on Computer Design (ICCD), Oct 2015, pp. 379–382.
- [4] B. Wu, F. N. Iandola, P. H. Jin, and K. Keutzer, “Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving,” CoRR, vol. abs/1612.01051, 2016. [Online]. Available: http://arxiv.org/abs/1612.01051
- [5] “Jetson/graphics performance,” https://elinux.org/Jetson/Graphics PerformancePower Use - Overview, accessed: 2018-09-20.
- [6] “nRF52840 board specification,” http://infocenter.nordicsemi.com/pdf/nRF52840 PS v1.0.pdf, accessed: 2018-09-20.
- [7] N. S. Liangzhen Lai and V. Chandra, “Cmsis-nn: Efficient neural network kernels for arm cortex-m cpus,” Jan 2018. [Online]. Available: https://arxiv.org/pdf/1801.06601.pdf
- [8] “NUCLEO-F746ZG digi-key overview page,” https://www.digikey.com/product- detail/en/stmicroelectronics/NUCLEO-F746ZG/497-16282-ND/5806779, accessed: 2018-09-20.
- [9] “Sign language digits dataset,” https://www.kaggle.com/ardamavi/sign-language-digits- dataset, accessed: 2018-09-30.
- [10] L. Lai, N. Suda, and V. Chandra, “Deep convolutional neural network inference with floating-point weights and fixed-point activations,” CoRR, vol. abs/1703.03073, 2017. [Online]. Available: http://arxiv.org/abs/1703.03073