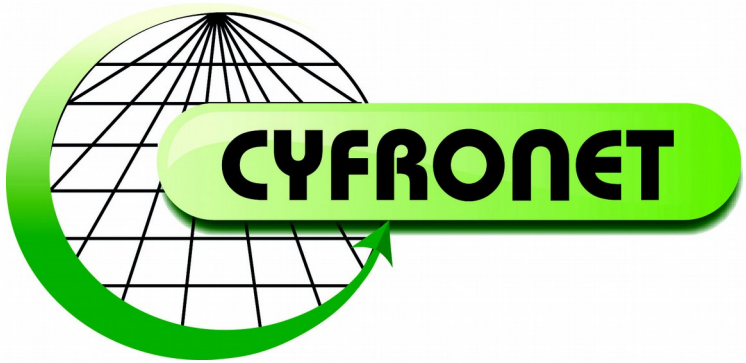
AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

AGH UNIVERSITY OF SCIENCE
AND TECHNOLOGY



Hardware aware neural network compression

Krzysztof Wróbel, Marcin Pietroń, Maciej Wielgosz,
Michał Karwatowski, Kazimierz Wiatr

Kraków 22-24.10.2018

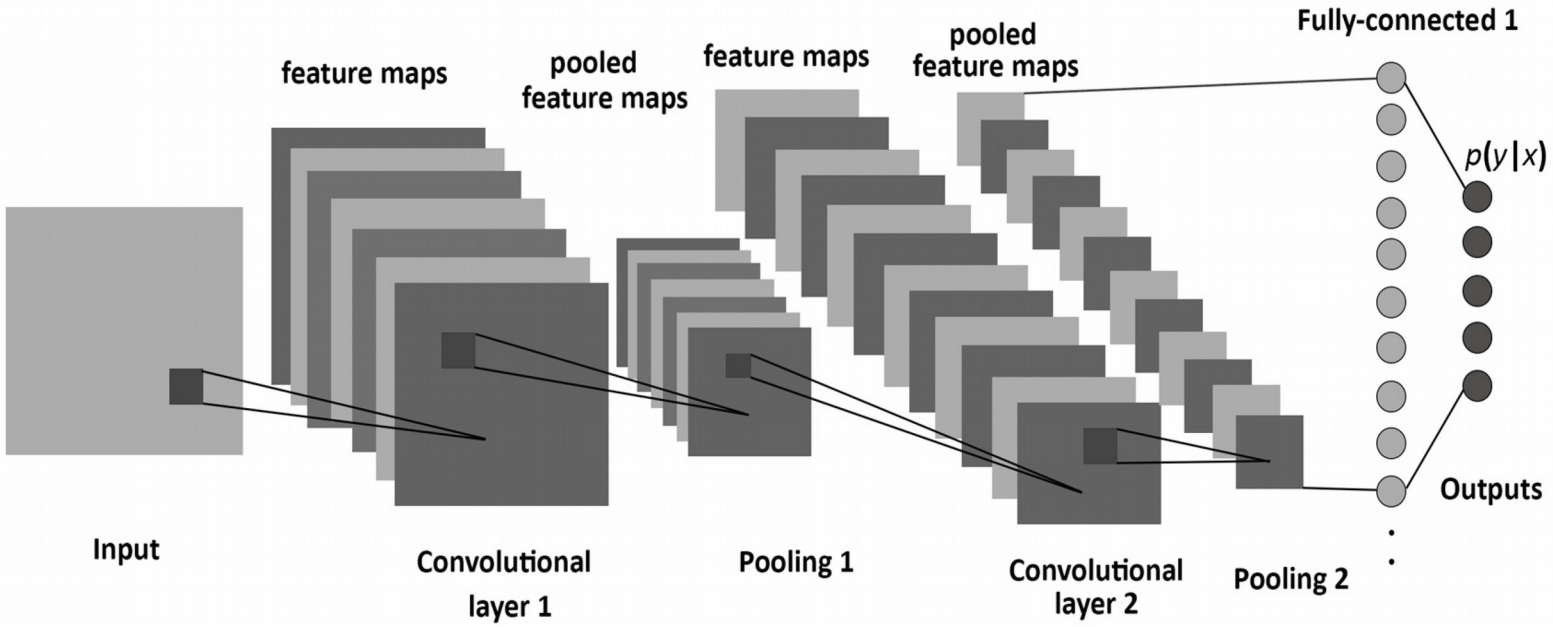
Agenda

- Natural Language Processing
- Compression techniques
- Implementation results

Natural Language Processing

- machine translation
- voice typing
- sentiment analysis
- question answering
- automatic summaryzation
- ...

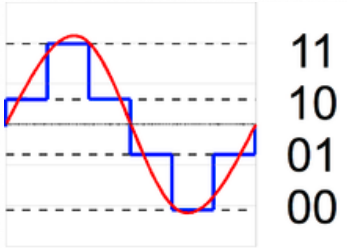
Convolutional Neural Network



$$y_i = b_i + \sum_{j=1}^{M} F_{ij} * x_j$$

Quantization

$$\mathrm{flp}_{single} : \{\pm 2^{-126}, \ldots, \pm(2 - 2^{-23}) \times 2^{127}\}$$

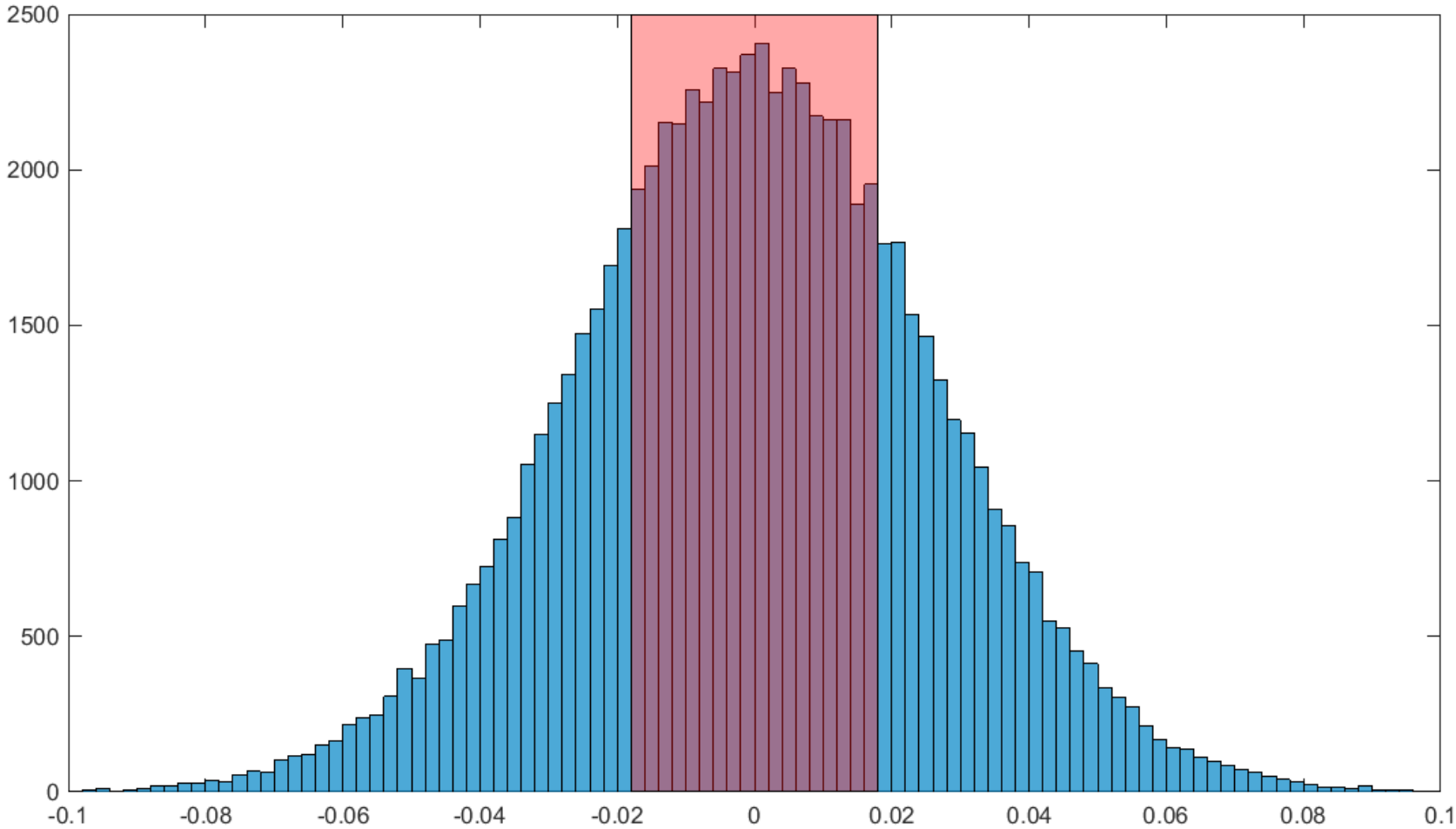


$$q_{\mathrm{fxp}} = \mathcal{Q}(x_{\mathrm{flp}}) = \mu + \sigma \cdot \mathrm{round}(\sigma^{-1} \cdot (x - \mu))$$

Scale factor: $\sigma = 2^{-\mathbf{frac\_bits}}$ $\quad \mathbf{int\_bits} = \mathrm{ceil}(\log_2(\max_{x \in \mathcal{S}} |x|))$

Mean value $\mu = 0$

Pruning



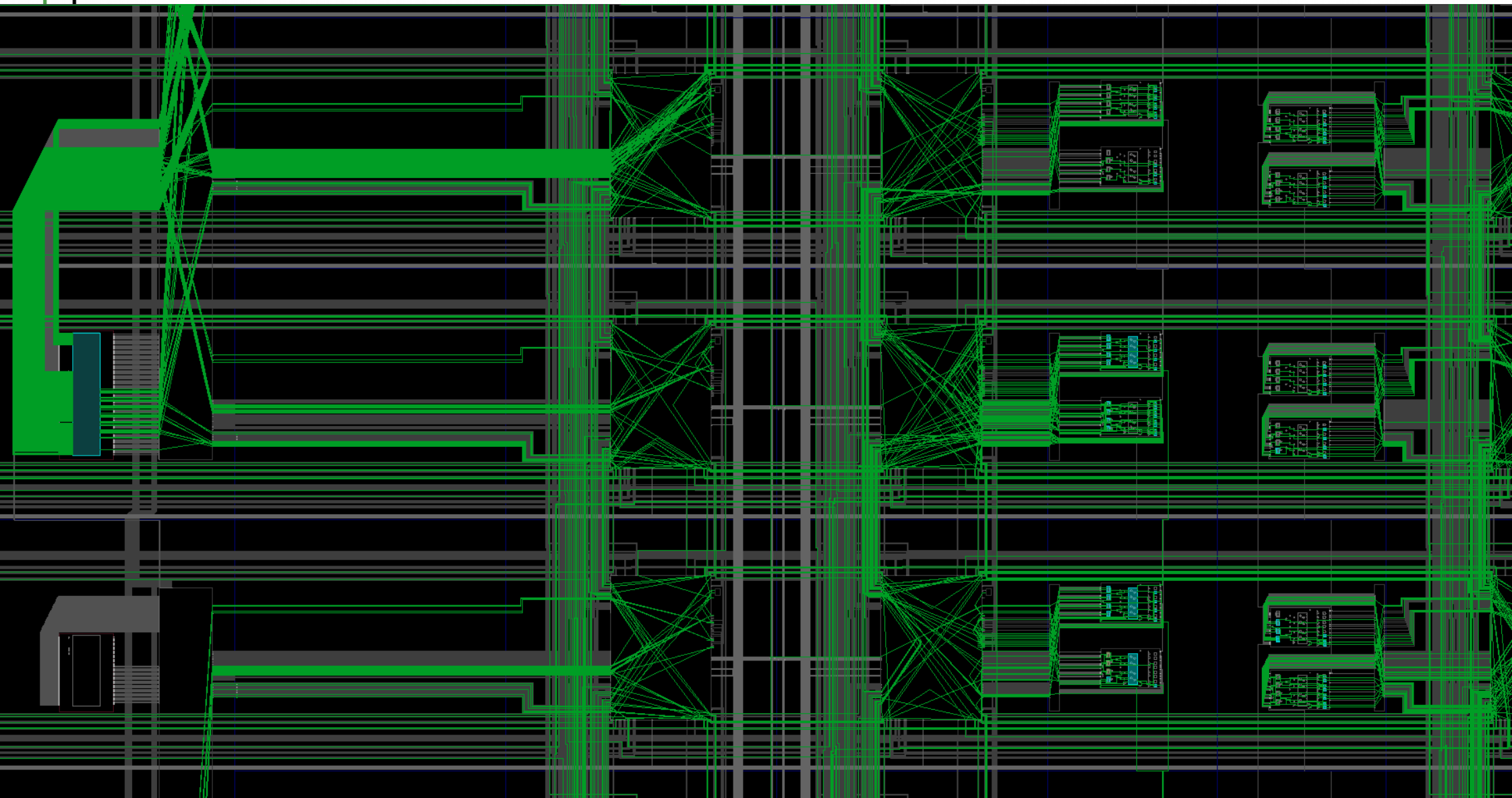
Pruning

Algorithm 1 1D convolution layer with pruning

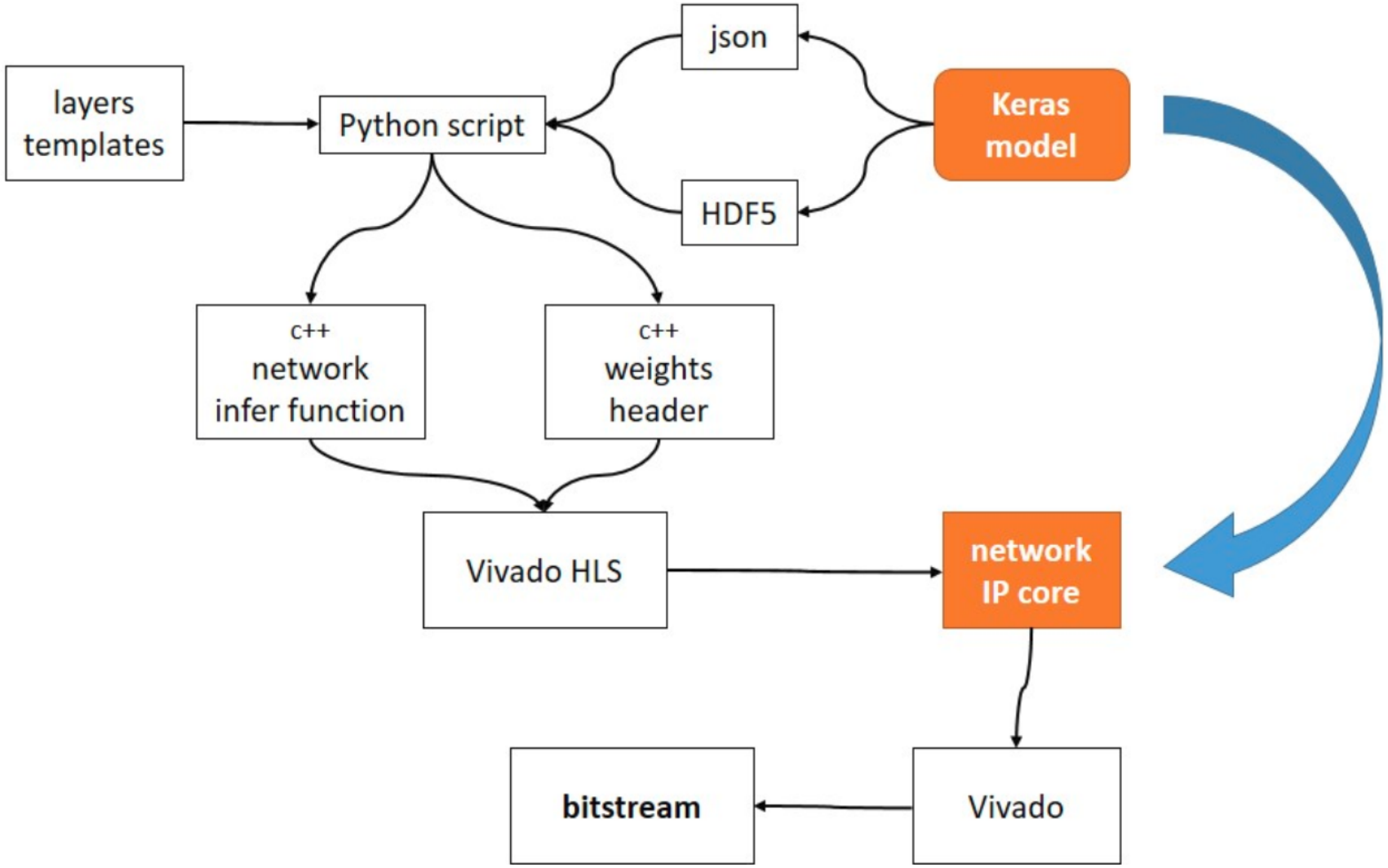
```
1:  for outS in output_size do
2:      for outFM in output_feature_maps do
3:          for inFM in input_feature_maps do
4:              for kerS in kernel_size do
5:                  if abs(weight[inFM][outFM][kerS]) >= pruning_threshold then
6:                      output(outS, outFM) += input(outS + kerS, inFM) * weight[inFM][outFM][kerS]
7:                  end if
8:              end for
9:          end for
10:         output(outS, outFM) += bias[outFM]
11:     end for
12: end for
```

AGH

FPGA



Keras model to IP core



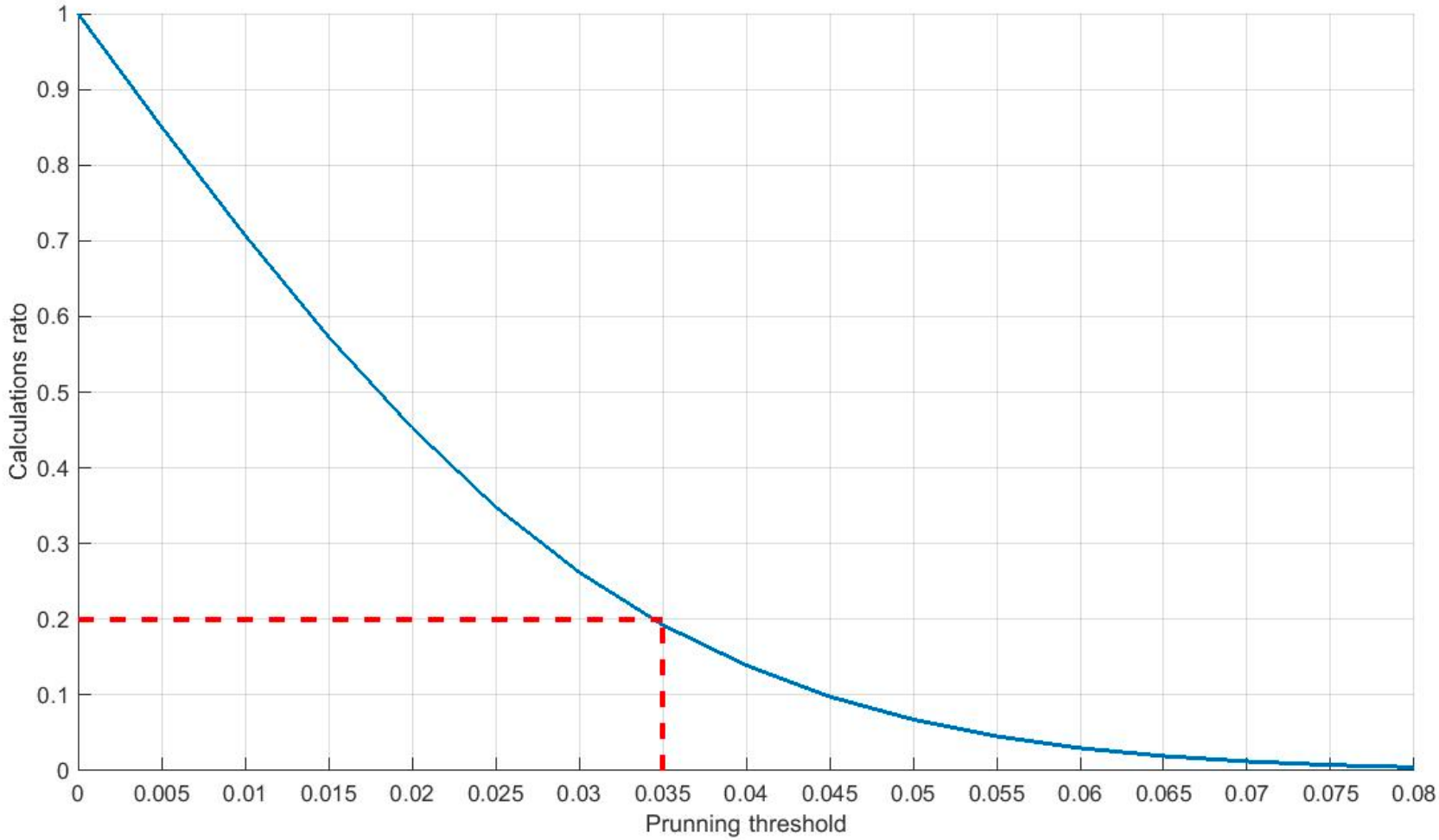
Movie Review dataset

- 1000 positive reviews
- 1000 negative reviews
- Validation split: 80% -20%

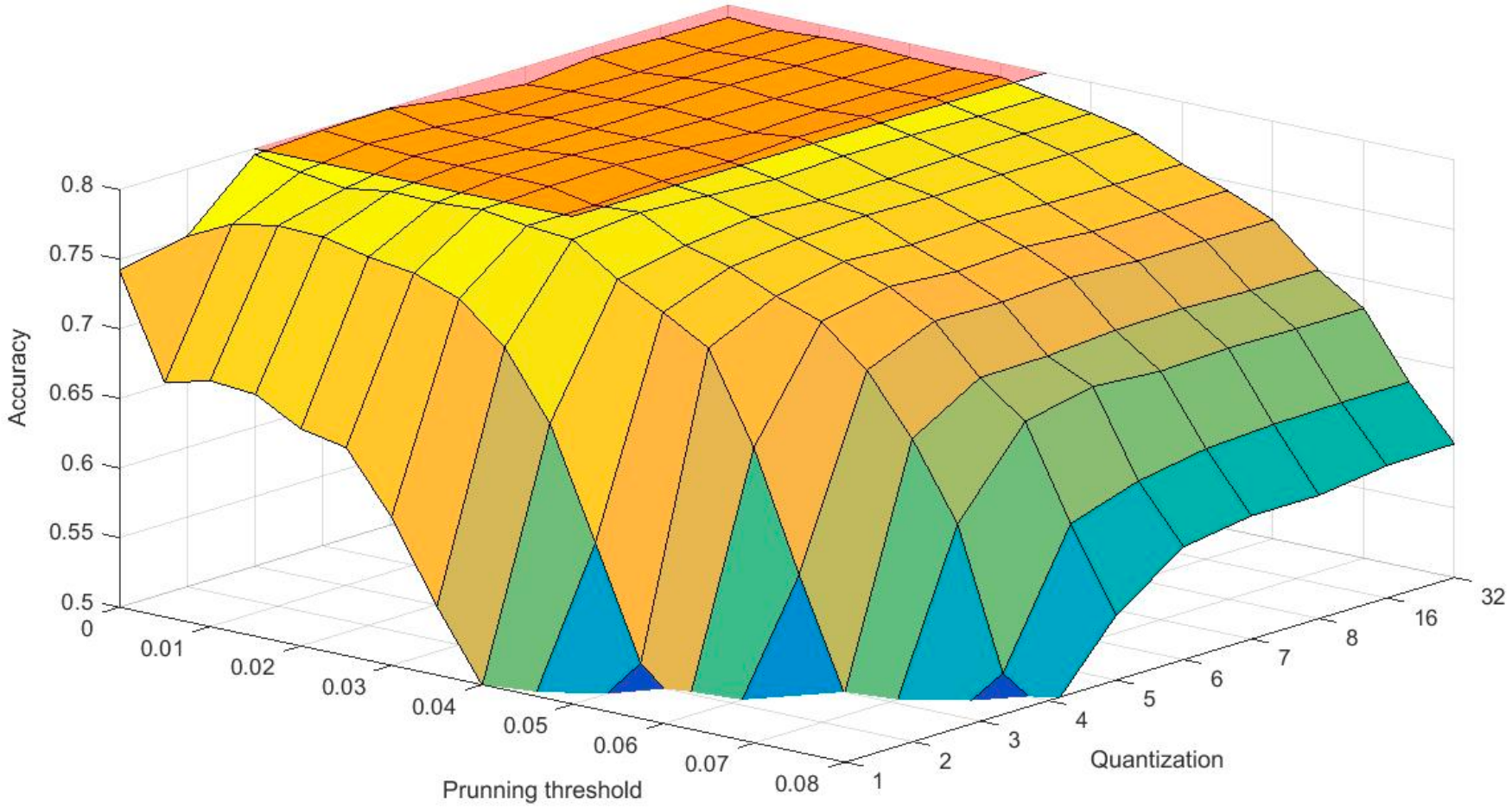
POSITIVE: the fact there's a dead body in the corner goes to enhance the feeling of paranoia and a mysterious , hurried call telling him to leave immediately is also very chilling

NEGATIVE: you could easily sleep through whole sections of the film (as some fellow critics did) and wake up in a scene exactly like the one you nodded off in , not having missed anything worthwhile

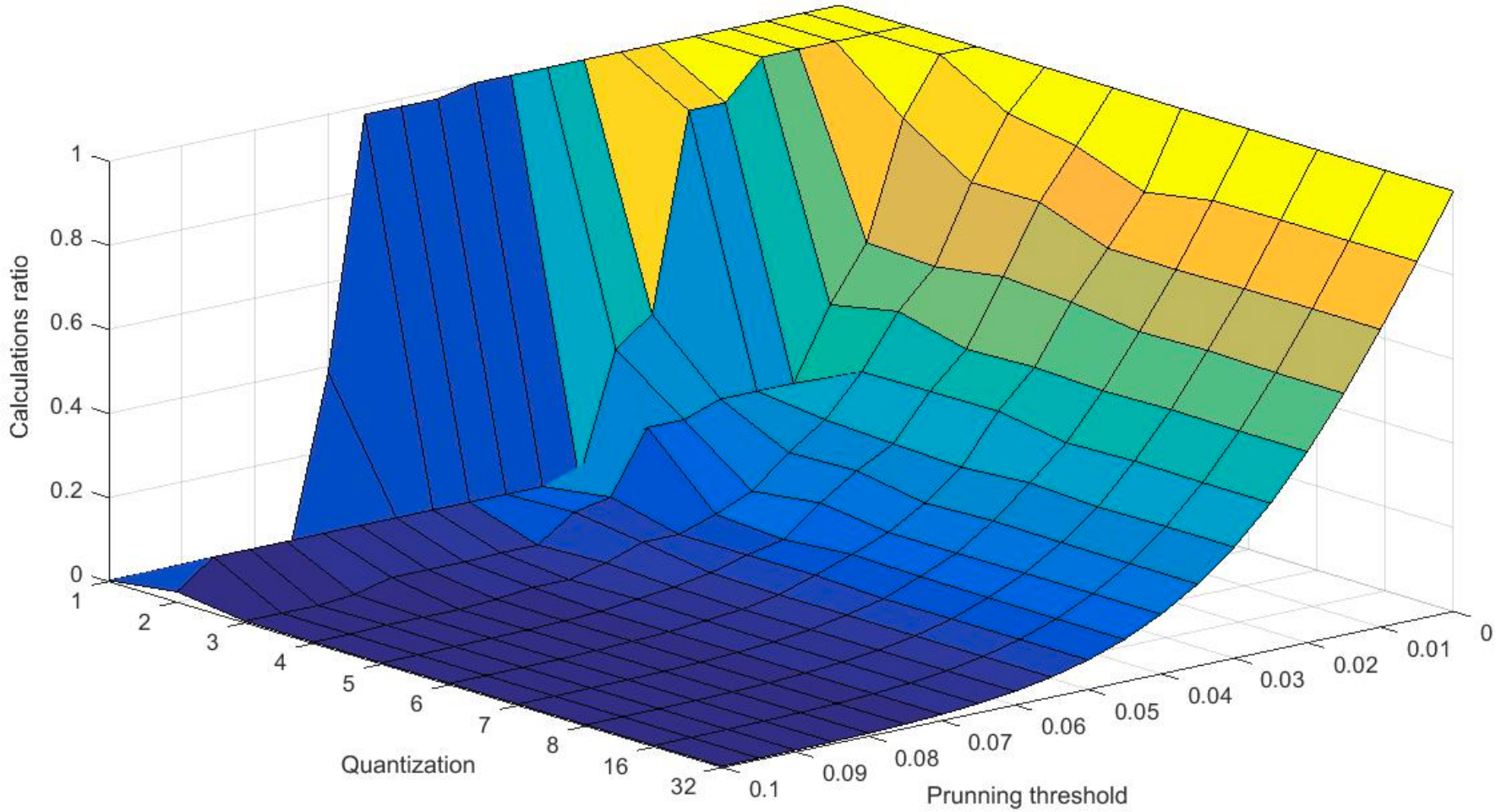
Extensiveness of pruning



Accuracy drop



The order matters



FPGA implementation

Precision	Pruning threshold	FF	LUT
32	0	8118	60750
	0.035	4449	14906
3	0	1952	14553
	0.035	1899	14047

Conclusions

- No more than 1% drop in accuracy after compression
- Around 4 times less FPGA resources needed

Future work

- Non linear quantization
- Retraining
- Architectural changes

Dwa serwery z dwoma kartami firmy Nallatech zawierającymi układy FPGA Altera Stratix V

Parametry karty typ 1:

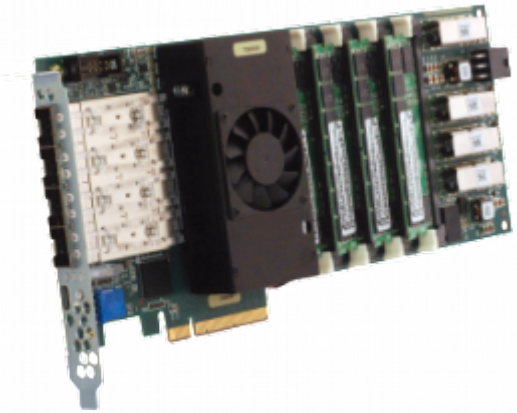
Nallatech 395-AB: Stratix V AB

32GB of DDR3

Parametry karty typ 2:

Nallatech 395: Stratix V D8

32GB of DDR3



Programowanie:

OpenCL

Altera Quartus

Narzędzia:

Altera OpenCL SDK

Altera Quartus

OpenCL

ALTERA PREFERRED BOARD FOR OpenCL™