A Reduced Order Approach for Artificial Neural Networks Applied to Object Recognition

AUTHORS

Laura Meneghetti,
Nicola Demo, Gianluigi Rozza

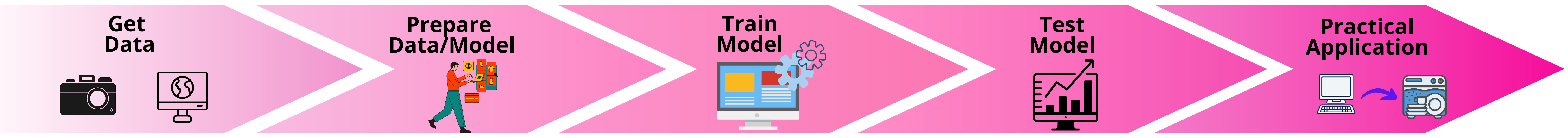
AFFILIATIONS

SISSA MathLab, Trieste, Italy
Fast Computing Srl, Trieste, Italy

Finanziato dall'Unione europea NextGenerationEU

INTRODUCTION

Development pipeline of an Artificial Neural Network for the problem of Object Recognition to be later deployed in vision embedded systems.



- Benchmark Dataset
- Custom Dataset

- Convolutional Neural Networks
- Object Detectors

- high number of parameters (space required)
- need GPU for training
- demanding hardware resources

Embedded Systems

- limited hardware resources
- strict memory constraints
- low CPU performance

DIMENSIONALITY PROBLEMS

A REDUCED ORDER APPROACH

Choose a dataset for the problem under consideration and determine the train dataset as the 80% of the complete dataset. Its components can be described as input-(expected) output pairs: $\mathcal{D}_{\text{train}} = \{\mathbf{x}^{(0),j}, \mathbf{y}^j\}_{j=1}^{N_{\text{train}}}$.

1 Original Neural Network

Take an Artificial Neural Network. It can be described as the compositions of its layers: $\mathcal{ANN} \equiv f_L \circ f_{L-1} \circ \cdots \circ f_1$.

2 Network Splitting

Choose a cut-off layer l (empirically). Define the pre- and post-models as follows:

$$\mathcal{ANN}^l_{\text{pre}} = f_l \circ f_{l-1} \circ \cdots \circ f_1, \qquad \mathcal{ANN}^l_{\text{post}} = f_L \circ f_{L-1} \circ \cdots \circ f_{l+1} .$$

Discard the post-model.

3 Dimensionality Reduction

- Compute the pre-model outputs for each input in the train dataset.
- Project them into a low-dimensional space of size R, using the projection matrix obtained through the chosen reduction method.
- Goal of the reduction methods: determine the most important parameters (eingenvalue analysis).

Tested Reduction Methods:

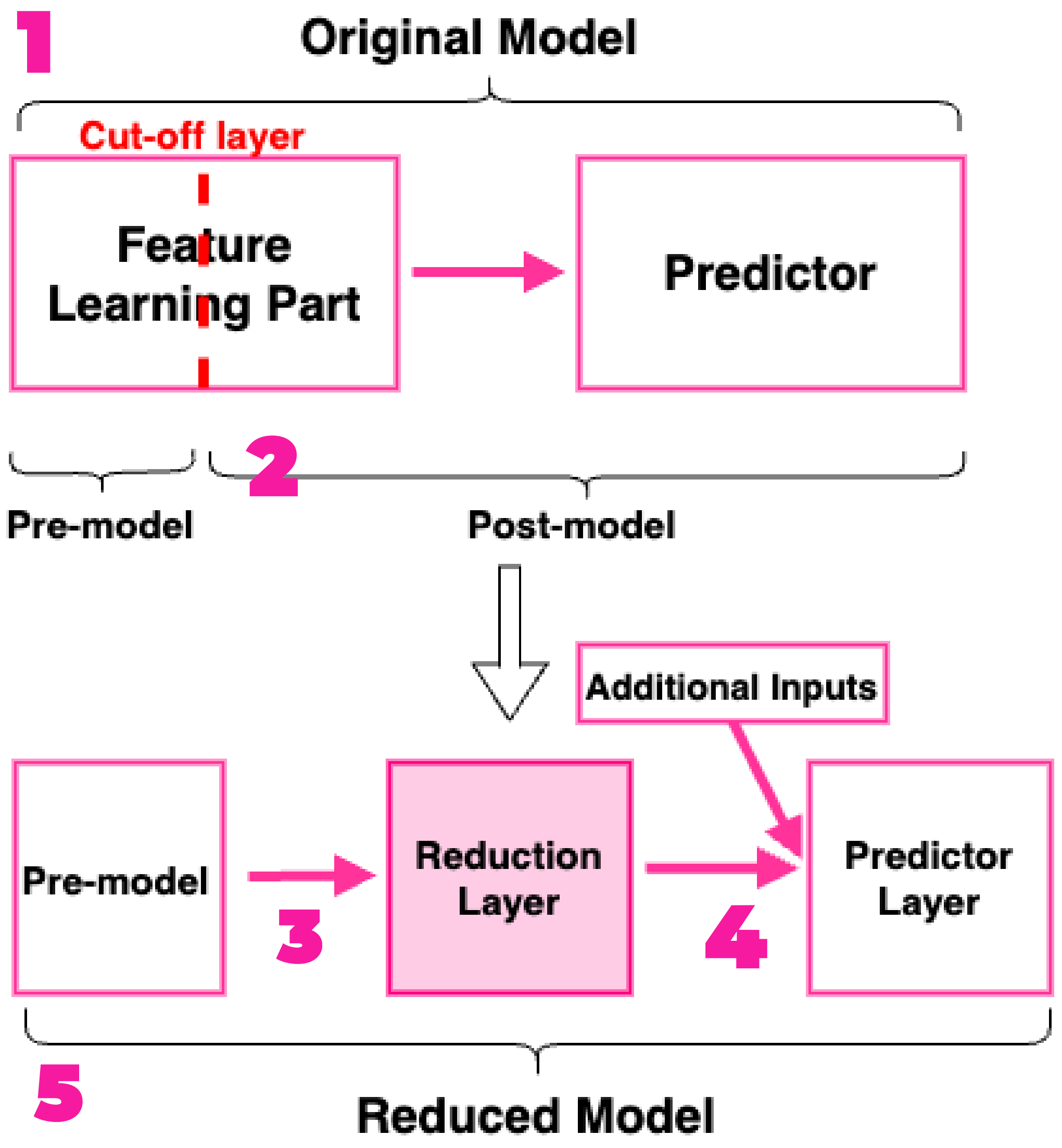
- **POD**: Compute the trunctated Singular Value Decomposition (SVD) of the matrix containing the linearized pre-model outputs. R represents the reduced output dimension.
- **HOSVD**(generalization in higher dimensions of SVD): Take into account the tensorial structure of pre-model outputs by computing the truncated SVD in each tensor direction.

4 Predictor

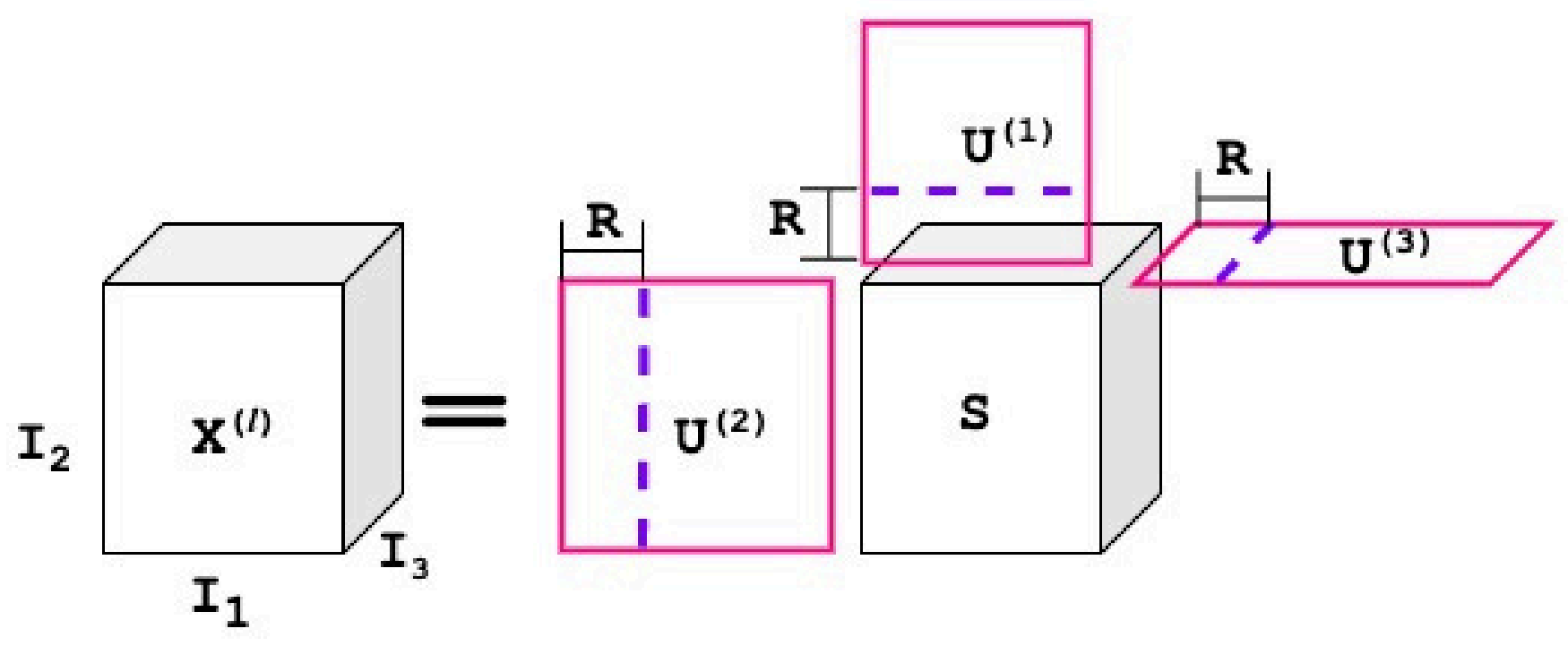
- The output of Step 3 becomes the input for the predictor, together with some additional inputs.
- Maintain the predictor's architecture of the original model, adjusting it for the variation in size of its inputs.

5 Compressed Neural Network

Re-training of the constructed Reduced Artificial Neural Network to achieve a good level of accuracy, comparable to the original model.



HOSVD decomposition

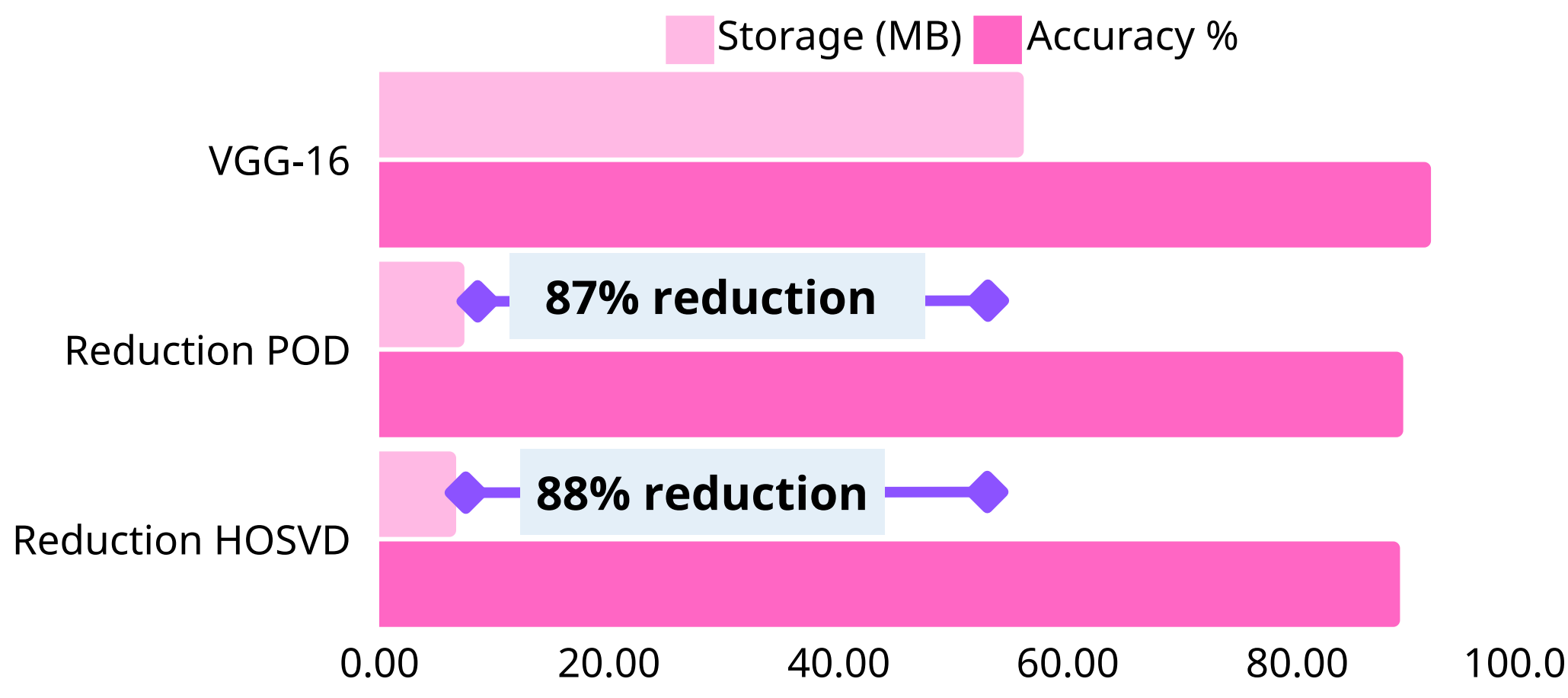


RESULTS

Some experimental results obtained with our reduction technique.

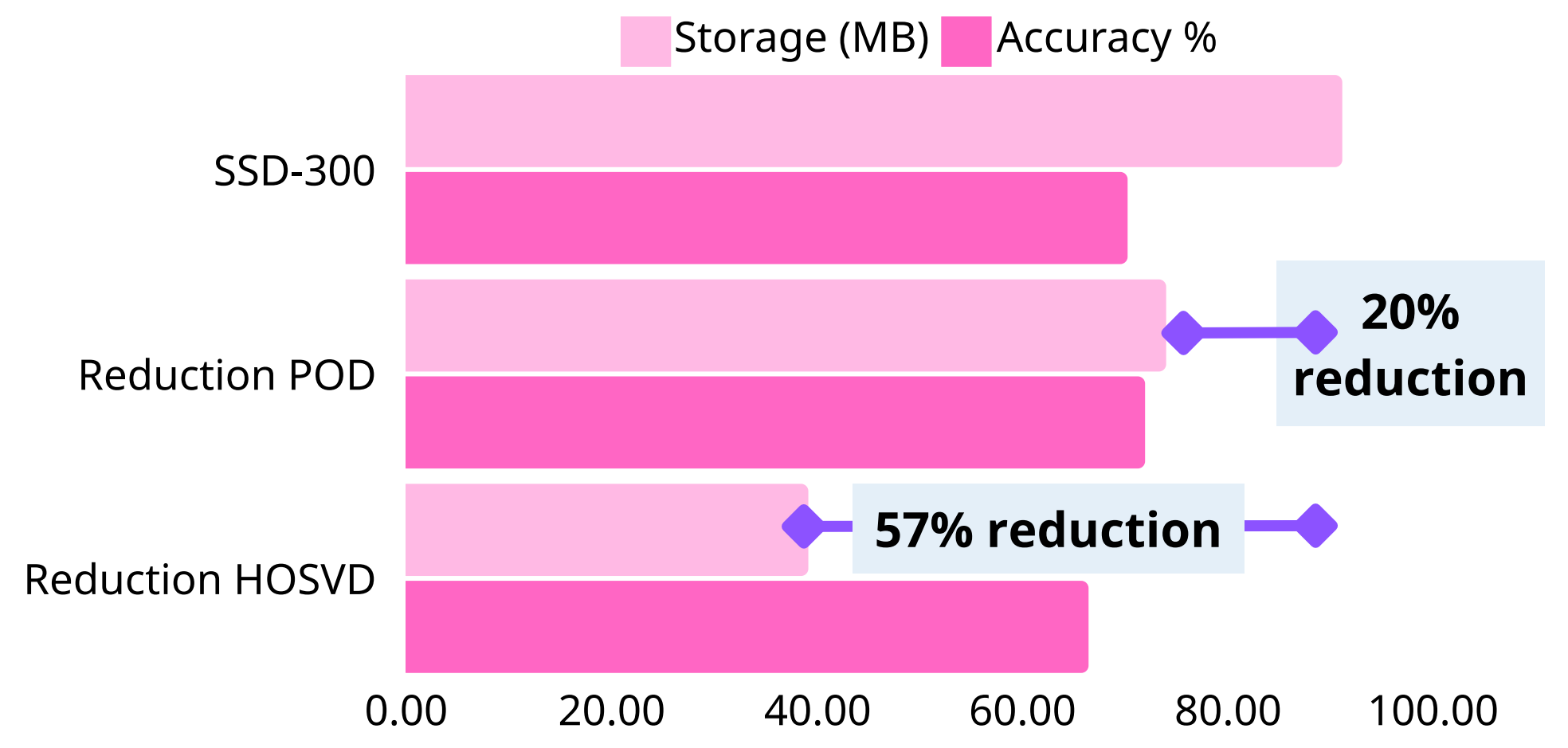
IMAGE RECOGNITION

- Original Model: VGG-16
- Dataset: CIFAR10
- Cut-off layer: 7
- Reduced dimension R: 50 for POD, 3x3x35 for HOSVD

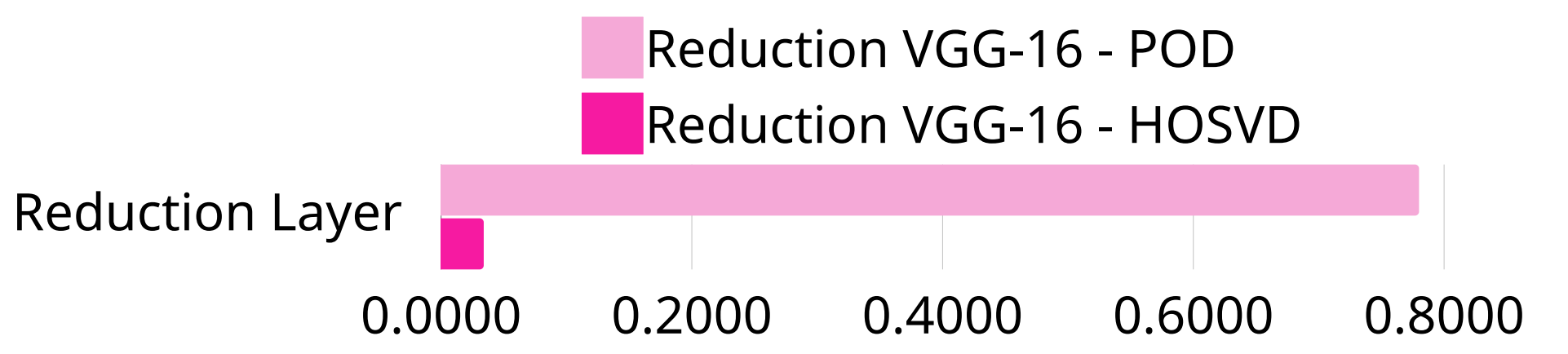


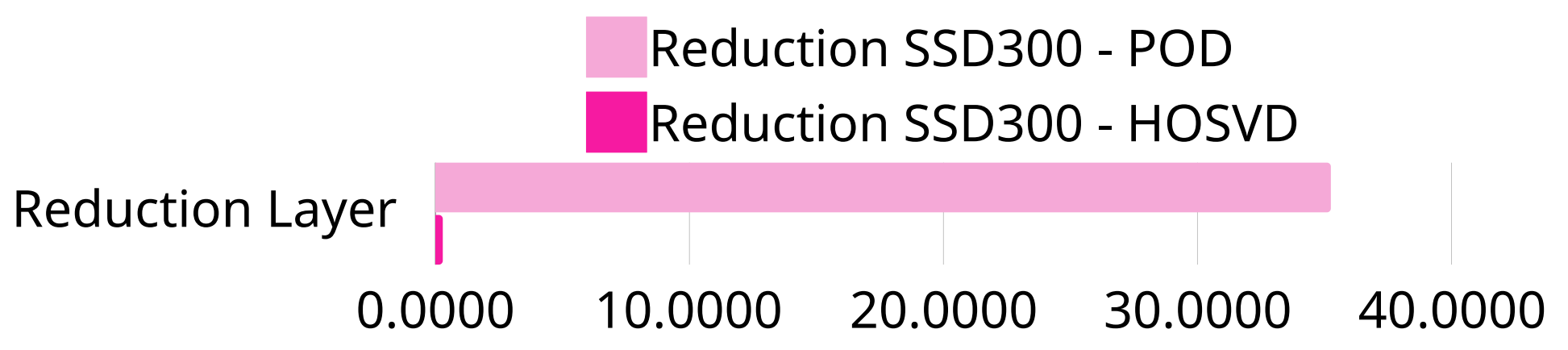
OBJECT DETECTION

- Original Model: SSD300
- Dataset: smaller dataset (300 images and two categories) extracted from PASCALVOC
- Cut-off layer: 11
- Reduced dimension R: 50 for POD, 3x3x150 for HOSVD

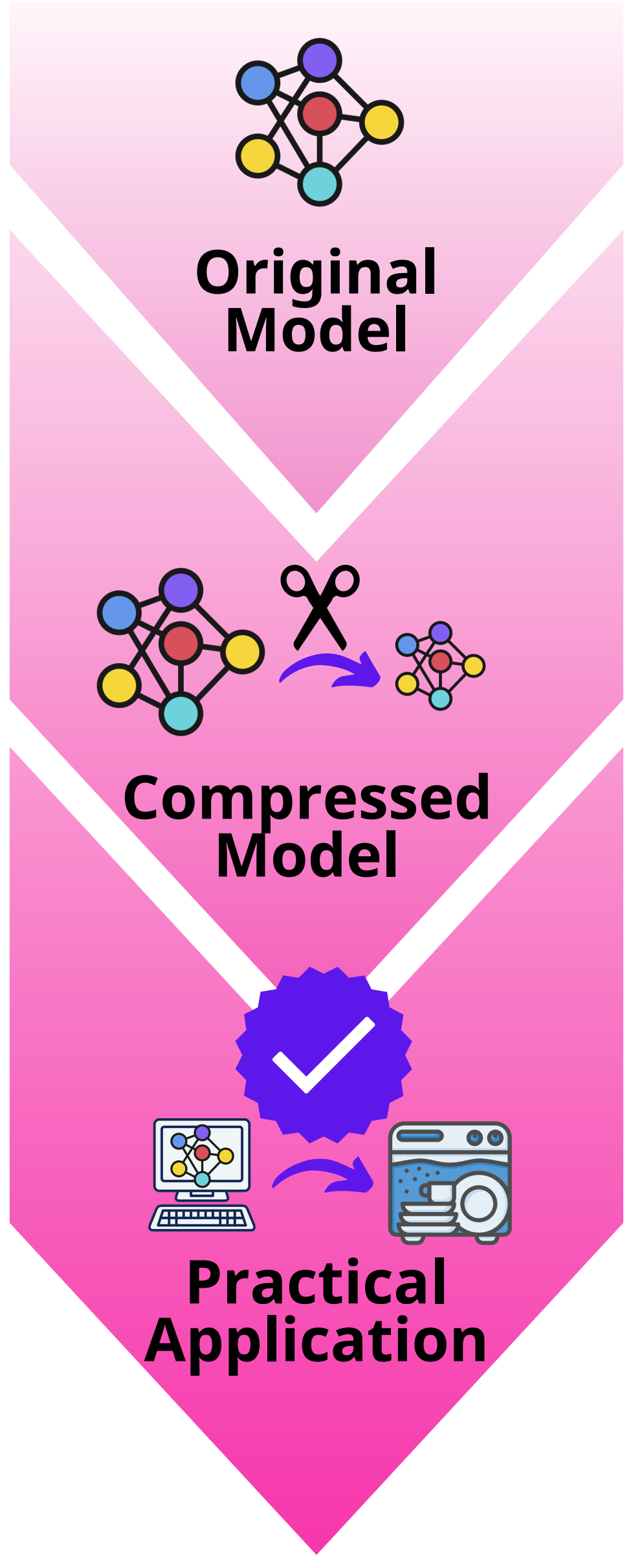


Storage comparison (MB) for the reduction layers using POD or HOSVD.





CONCLUSIONS



SOME FUTURE DEVELOPMENTS

Generalizability of the approach
more tasks, more architectures, more datasets

Criteria for cut-off index
Information theory notions (e.g. entropy) to understand the most important neurons/layers

More reduction tecniques
e.g. non linear ones

Comparison and integration of other compression methods
e.g. pruning, quantization,..

CONTACT INFORMATION

laura.meneghetti@sissa.it
nicola.demo@fastcomputing.net
gianluigi.rozza@sissa.it

Check our GitHub page for the code!

REFERENCES

- Meneghetti, L., Demo, N., Rozza, G.: A dimensionality reduction approach for convolutional neural networks. Applied Intelligence 53(19), 22818–22833 (2023). https://doi.org/10.1007/s10489-023-04730-1
- Meneghetti, L., Demo, N., Rozza, G.: A Proper Orthogonal Decomposition Approach for Parameters Reduction of Single Shot Detector Networks. In: 2022 IEEE ICIP. pp. 2206–2210 (2022). https://doi.org/10.1109/ICIP46576.2022.9897513
- Meneghetti, L., Zanin, S., Demo, N., Rozza,: Deep Neural Network Compression via Tensor Decomposition, (2024) submitted