A Proper Orthogonal Decomposition approach for parameters reduction of Single Shot Detector networks

Laura Meneghetti, Nicola Demo, and Gianluigi Rozza   *Mathematics Area, mathLab, SISSA, Trieste, Italy*

The Research Hub By ELECTROLUX PROFESSIONAL

Electrolux PROFESSIONAL

Paper ID: 2448

Introduction

Convolutional Neural Networks (CNNs) have achieved an impressive success in solving many problems in several fields including computer vision and image processing, attracting a huge interest from the industrial sector. However, running these deep neural networks in embedded systems with limited hardware represent a challenging task for several engineering applications.

Objectives

Goals of the project:

1. Reduction of the memory storage required for an object detector;
2. Application in an embedded system;
3. Accurate performances;
4. Real-time predictions.

Reduced SSD-type Object Detector

Assumption: SSD-type architecture, composed of:

- a base net (a CNN);
- some additional convolutional layers;
- two siblings predictors, one for localization prediction and one for class prediction.

1. Network Splitting

Let $\mathit{Obj\_Det} : \mathbb{R}^{\text{in}} \to \mathbb{R}^{n_{\text{class}} \times 4}$ be an object detector. It can be described as the composition of L functions f_j, representing the different layers of the net:

$$\mathit{Obj\_Det} = f_{L+1} \circ f_L \circ \cdots \circ f_1 \qquad (1)$$

Denoting with ℓ the cut-off index, we can define the pre-model and the post-model as:

$$\begin{aligned} \mathit{basenet}^l_{\text{pre}} &= f_l \circ f_{l-1} \circ \cdots \circ f_1, \\ \mathit{basenet}^l_{\text{post}} &= f_L \circ f_{L-1} \circ \cdots \circ f_{l+1}. \end{aligned} \qquad (2)$$

NOTE: ℓ has to be chosen carefully!

The output of the pre-model $\mathbf{x}^{(\ell)}$ lies in a high-dimensional space $\to$ project into a low-dimensional one.

2. Dimensionality Reduction

Let $\mathbf{S} = [\mathbf{x}^{(l),1}, \dots, \mathbf{x}^{(l),N_{\text{train}}}]$ be the snapshot matrix.We can compute the SVD of it:

$$\mathbf{S} = \mathbf{\Psi}\mathbf{\Sigma}\mathbf{\Theta}^T, \qquad (3)$$

- the columns of $\mathbf{\Psi}$ are the left-singular vectors, also called POD modes;
- $\mathbf{\Sigma}$ contains the corresponding eigenvalues.

Fixed r to be the reduced, we define the projection matrix $\mathbf{\Psi}_r$ by keeping the first r modes.

$$\mathbf{z}^i = \mathbf{\Psi}_r^T \mathbf{x}^{(l),i}, \quad \text{for } i = 1, \dots, N_{\text{train}}. \qquad (4)$$

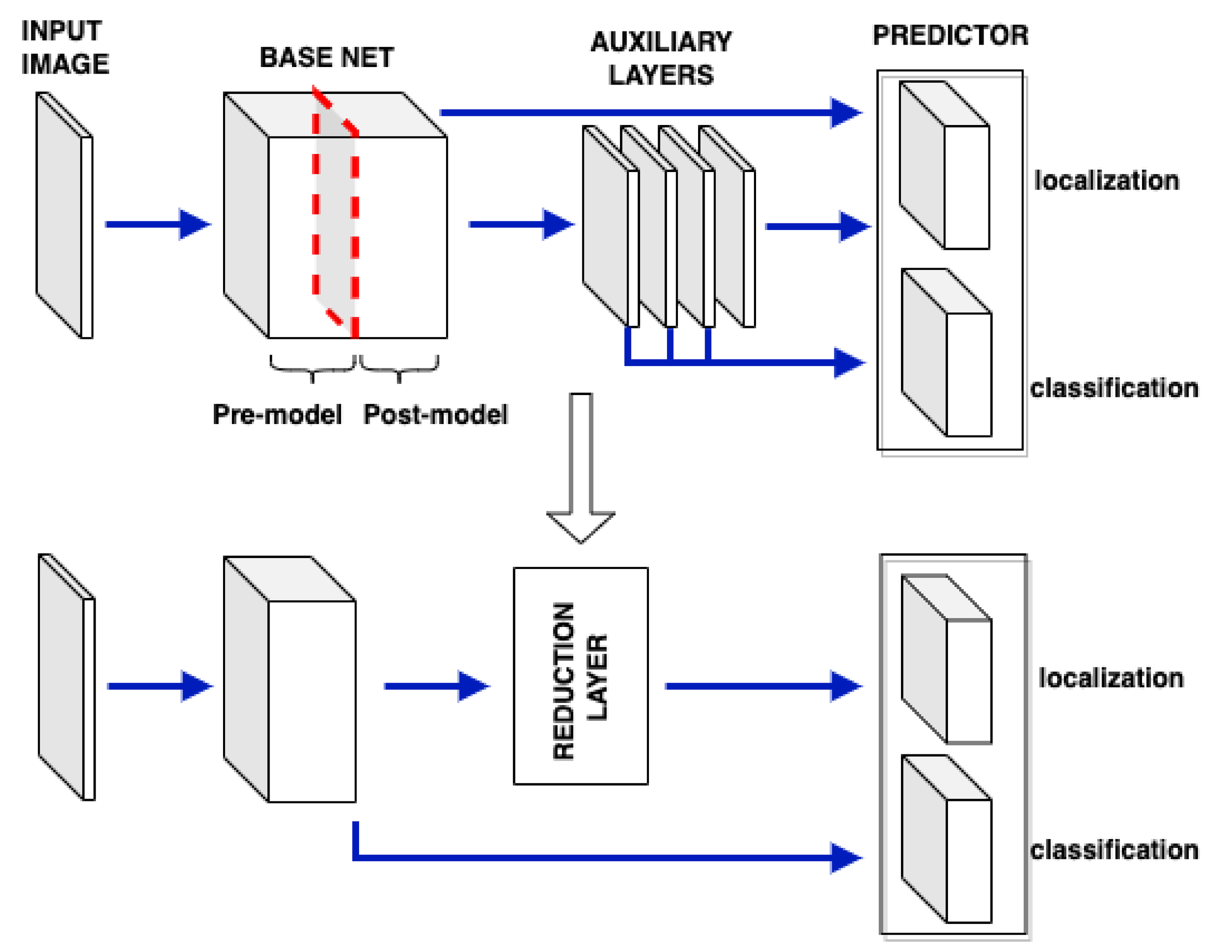
3. Predictor

Same predictor as before:

$$\hat{\mathbf{y}}_{\text{loc}}, \hat{\mathbf{y}}_{\text{cls}} = \mathit{predictor}(\mathbf{x}^{(l)}, \mathbf{z}) \qquad (5)$$

BUT, since the number of inputs is changed, the scale parameter and the number of anchor boxes has to be adjusted.

Reduced Approach- Main Idea



Algorithm 1 Pseudo-code for the construction of the reduced object detector.

Inputs:

- train dataset $\mathcal{D}_{\text{train}} = \{\mathbf{x}^{(0),j}, \mathbf{y}^j_{\text{loc}}, \mathbf{y}^j_{\text{cls}}\}_{j=1}^{N_{\text{train}}}$,
- $\mathit{Obj\_Det} = [\mathit{basenet}, \mathit{auxlayers}, \mathit{predictor}]$,
- reduced dimension r,
- index of the cut-off layer l,
- a test dataset $\mathcal{D}_{\text{test}} = \{\mathbf{x}^i, \mathbf{y}^i_{\text{loc}}, \mathbf{y}^i_{\text{cls}}\}_{i=1}^{N_{\text{test}}}$.

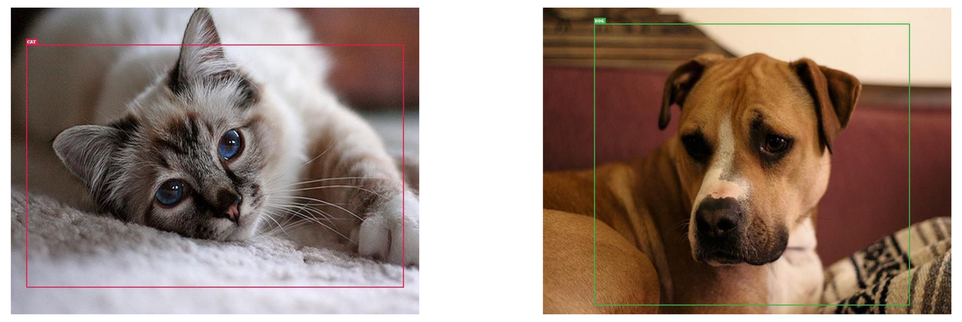
Output: Reduced Object Detector $\mathit{Obj\_Det}^{\text{red}}$

1: $\mathit{basenet}^l_{\text{pre}}, \mathit{basenet}^l_{\text{post}} = \text{splitting\_net}(\mathit{basenet}, l)$
2: $\mathbf{x}^{(l)} = \mathit{basenet}^l_{\text{pre}}(\mathbf{x}^{(0)})$
3: $\mathbf{z} = \text{reduce}(\mathbf{x}^{(l)}, r)$
4: $\hat{\mathbf{y}}_{\text{loc}}, \hat{\mathbf{y}}_{\text{cls}} = \mathit{predictor}(\mathbf{x}^{(l)}, \mathbf{z})$
5: Training of the constructed reduced net using $\mathcal{D}_{\text{test}}$.

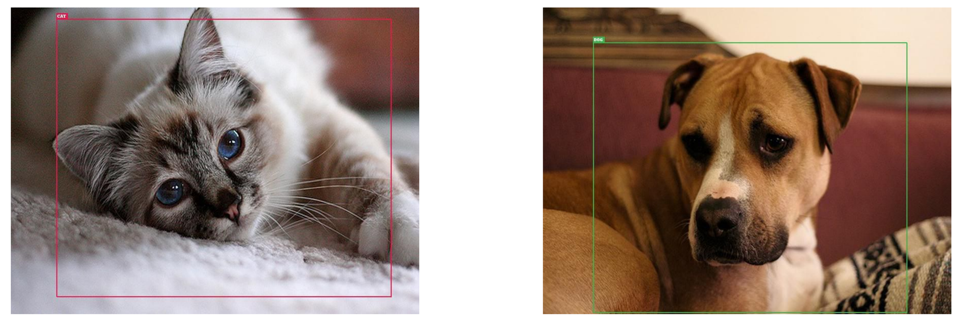
Results1- Cats & Dogs

Network	mAP	Storage (Mb)	Training Time
SSD300	70.2%	91.09	43.5 h
SSD300_red	59%	77.45	26 h

SSD300



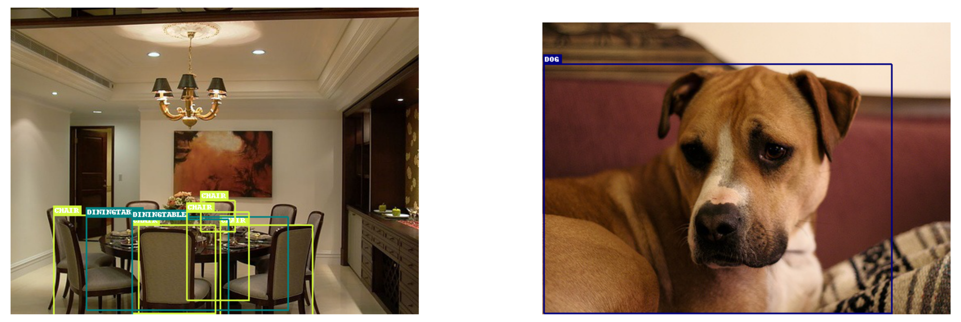
Reduced SSD300



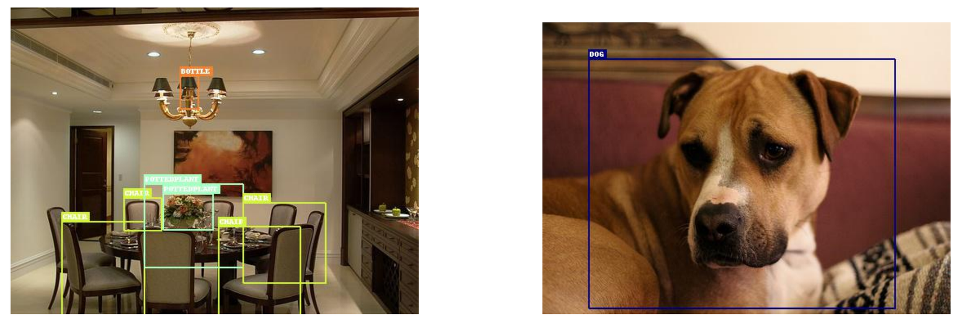
Results-2 PASCAL VOC

Network	mAP	Storage (Mb)	Training Time
SSD300	77.8%	100.23	48 h
SSD300_red	39%	76.23	18 h

SSD300



Reduced SSD300



References

[1] Laura Meneghetti, Nicola Demo, and Gianluigi Rozza. A Dimensionality Reduction Approach for Convolutional Neural Networks. *arXiv preprint arXiv:2110.09163*, 2021.

[2] Chunfeng Cui, Kaiqi Zhang, Talgat Daulbaev, Julia Gusak, Ivan Oseledets, and Zheng Zhang. Active subspace of neural networks: Structural analysis and universal attacks. *SIAM Journal on Mathematics of Data Science*, 2(4):1096–1122, 2020.

[3] Jan S. Hesthaven, Gianluigi Rozza, and Benjamin Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. Springer Briefs in Mathematics. Springer, Switzerland, 1 edition, 2015.

Future Research

- Knowledge Distillation for Object Detection;
- Employment of hyperreduction techniques or POD variants.
- Iterative procedure for the choice of ℓ.

Contact Information

Code https://github.com/mathLab/Smithers

Email laura.meneghetti@sissa.it
nicola.demo@sissa.it
gianluigi.rozza@sissa.it