In the Depths of Hyponymy: A Step Towards Lifelong Learning

Tommaso Boccato*, Timothy Patten†,
Markus Vincze† and Stefano Ghidoni*

*Department of Information Engineering, Università degli Studi di Padova, Padova, Italy
Email: tommaso.boccato@studenti.unipd.it, stefano.ghidoni@unipd.it
†Automation and Control Institute, TU Wien, Vienna, Austria
Email: {patten, vincze}@acin.tuwien.ac.at

Authors

Tommaso Boccato received his B.Sc. in Information Engineering from the University of Padova, Italy. He is completing his master's thesis project at the Vision for Robotics (V4R) laboratory, Technical University of Vienna (TU Wien), Austria. His research interests include machine learning, computer vision and network science.

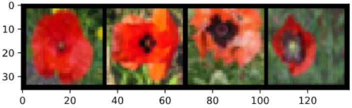
Dr. **Timothy Patten** received his PhD from the Australian Centre for Field Robotics at the University of Sydney, Australia. He is now a postdoctoral researcher with the V4R laboratory at TU Wien. He has been involved in a number of research and industry sponsored projects in which he worked on object segmentation, recognition, grasping and task planning. Currently, he is the principle investigator at TU Wien in the CHIST-ERA project InDex, for which he is developing methods for object tracking and semantic grasping.

Prof. **Markus Vincze** received his Diploma in Mechanical Engineering from TU Wien in 1988 and a Master of Science from Rensselaer Polytechnic Institute, USA, 1990. He finished his PhD at TU Wien in 1993. With a grant from the Austrian Academy of Sciences he worked at HelpMate Robotics Inc. and at the Vision Laboratory of Gregory Hager at Yale University. In 2004 he obtained his habilitation in robotics. He started the V4R laboratory at TU Wien in 1996.

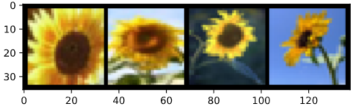
Stefano Ghidoni is Assistant Professor at the University of Padova. His research interests are in semantic segmentation, people re-identification systems, and deep learning for pattern recognition and scene understanding, mainly targeting robotics application. His research interests also include deep learning for human-robot cooperation, focused on robot task and motion planning in cooperative environments. He has also been involved in a number of projects for bringing intelligence into industrial processes, including airplane and boat production.

Introduction

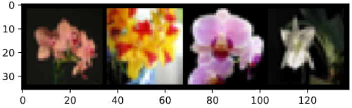
- The applications in which a robot should be able to understand what it sees are countless: human-robot interaction, healthcare, service robotics, industrial robotics, logistics, connected and autonomous vehicles.
- The last decade of advancements in deep learning have led to astonishing results in the applications that respond to the so called closed-world assumption [KSH12].
- Robots, however, operate in dynamic and uncontrolled environments.



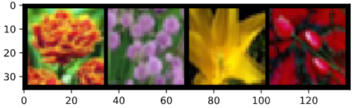
(a) `poppies`



(b) `sunflowers`



(c) `orchids`



(d) `tulips`

Figure 1: Randomly sampled batches extracted from 4 different CIFAR-100 [Kri09] classes. Images in the top row show homogeneous visual properties while images in the bottom row are characterized by very different visual properties. Yet, all the batches belong to specific categories. A question arises: *How does the intra-class variability impact a classifier, and how can an agent (e.g., a robot) recognize and exploit this phenomenon?*

- In the presented work, with reference to the classification task, a step is taken towards relaxing the aforementioned assumption by introducing a novel framework capable of allowing the refinement of the classes encoded into a classifier during its operational life.

Lifelong Learning Framework

- The objective pursued by the definition of the framework is to theoretically describe the operational life of a classifier trained on a set of semantic categories or classes labeled by the positive integers $\mathcal{K}_1 = \{1, \ldots, N_1\}$.
- It is therefore natural to define $\mathcal{K}_t \subseteq \mathbb{N}^+$ as the set of classes encoded into the classifier at time t.
- Let $\boldsymbol{x} \in \mathbb{R}^d$ be the features associated to a new sample seen by the classifier.
- Let $\mathcal{T}_t \subseteq \mathbb{R}^d \times \bigcup_{j=1}^{t} \mathcal{K}_j$ be the set containing all the samples, with the respective labels, seen by the classifier up to time t.

A model, to function within the defined framework, must be characterized by the following main ingredients.

1. A *multi-class recognition function* $F_t : \mathbb{R}^d \longrightarrow \mathcal{K}_t$.
2. The *additional state information* $\mathcal{S}_t = \{s_t^i\},\ \forall i \in \mathcal{K}_t$. For each semantic category, the corresponding element of the set should contain all the necessary information to compute its intra-class variability after the classification performed in the previous time step.
3. A formalization of the intra-class variability computation $V : \mathcal{S}_{t+1} \longrightarrow \mathbb{R}$.
4. A *trigger* $T : \mathbb{R} \longrightarrow \{0, 1\}$ defined in accordance with a criterion selected by the designer in order to establish whether class i needs to be split or not.

5. A *labeling process function* $L_t : \mathcal{P}(\mathcal{T}_t^i) \longrightarrow \mathcal{P}(\mathbb{N}^+ \setminus \bigcup_{j=1}^{t} \mathcal{K}_j)$, where $\mathcal{P}(\bullet)$ denotes the power set and $\mathcal{T}_t^i = \{(\boldsymbol{x}, k) \in \mathcal{T}_t \text{ s.t. } k = i\}$. The function aims to retrieve the sub-class labels of class i when its split is triggered. Once the new categories are collected, the classifier class structure has to be updated.
6. A *data retrieval function* $R : \mathcal{P}(\mathcal{K}_{t+1}) \longrightarrow \mathcal{P}(\mathbb{R}^d \times \mathcal{K}_{t+1})$. The function is responsible for retrieving the new data for the incremental training of the model.
7. An *incremental learning function* whose objective is to incrementally update the model by replacing the obsolete per-class recognition function with the ones related to the new semantic categories.

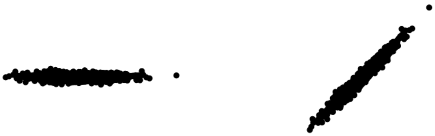
Metric for Intra-class Variability

- Let $\boldsymbol{X}$ be the matrix whose columns are the samples, belonging to or classified as belonging to class $i \in \mathcal{K}_t$, seen by the considered model up to time t.
- If the used classifier belongs to the category of deep models, $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^n$ can be defined as the function responsible for extracting deep representations from the generic sample features $\boldsymbol{x} \in \mathbb{R}^d$.
- Let $\phi(\boldsymbol{X})$ be the matrix obtained applying function ϕ to $\boldsymbol{X}$ columnwise.
- The matrix can be thought of as the repeated sampling of a probability distribution over $\mathbb{R}^n$.

- The intuition is to link the abstract concept of intra-class variability to the shape of the $\phi(\boldsymbol{X})$ sampling in the space of the deep representations.
- The formulated hypothesis follows: *The lower the intra-class variability of class i, the better the sampling $\phi(\boldsymbol{X})$ approximates a hyperball.*
- The concept of approximation introduced in the formulated hypothesis needs to be formalized.

- A first proposal consists of analyzing the per-component variances of the random vector $\phi(\boldsymbol{x})$.
- Let $\boldsymbol{C}_{\phi(\boldsymbol{X})}$ be the (sample) covariance matrix associated to the $\phi(\boldsymbol{X})$ data.
- Let $\boldsymbol{\sigma} = [\sigma_1^2, \ldots, \sigma_n^2]$ be the vector containing the diagonal terms of $\boldsymbol{C}_{\phi(\boldsymbol{X})}$ and $\tilde{\boldsymbol{\sigma}} = [\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2]$ be its normalized counterpart.
- Let $H(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \log_2 p_i$ be the entropy of the generic distribution $\boldsymbol{p} = [p_1, \ldots, p_n]$.
- The proposed metric is defined to be

$$V(\boldsymbol{C}_{\phi(\boldsymbol{X})}) = H(\tilde{\boldsymbol{\sigma}}) \tag{1}$$



(a) $\boldsymbol{C}_{\phi(\boldsymbol{x})_{0^\circ}} = \begin{bmatrix} 1 & 0 \\ 0 & 0.01 \end{bmatrix}$ (b) $\boldsymbol{C}_{\phi(\boldsymbol{x})_{45^\circ}} \simeq \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$

Figure 2: Rotated versions of the same set of samples. As reported by the captions, ${\sigma_{0^\circ}}^2_x \gg {\sigma_{0^\circ}}^2_y$ while ${\sigma_{45^\circ}}^2_x = {\sigma_{45^\circ}}^2_y$. The two cases lead to different aggregated scores.

- A subtle problem arises: rotated versions of the same sampling could lead to different aggregated scores.

- A possible solution is inspired by Principal Component Analysis (PCA) [ShI14].
- Let $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]$ be the eigenvalues of $\boldsymbol{C}_{\phi(\boldsymbol{X})}$ and $\tilde{\boldsymbol{\lambda}} = [\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n]$ be the distribution extracted from $\boldsymbol{\lambda}$.
- The final proposal consists of modifying (1) into

$$V(\boldsymbol{C}_{\phi(\boldsymbol{X})}) = H(\tilde{\boldsymbol{\lambda}}) \tag{2}$$

Qualitative Hypothesis Verification

1. The DeepNCM classifier [GCM18] is trained on 20 modified CIFAR-100 [Kri09] super-classes made of only one randomly selected sub-class.
2. 5000 unseen samples belonging to the same sub-classes exploited during the model training are supplied to the classifier.
3. After each classification, the model state is updated and the score produced by the metric computation is stored.
4. 5000 unseen samples, from randomly chosen sub-classes, different from the ones of the training phase, are supplied to the classifier and the corresponding metric scores are computed and stored.

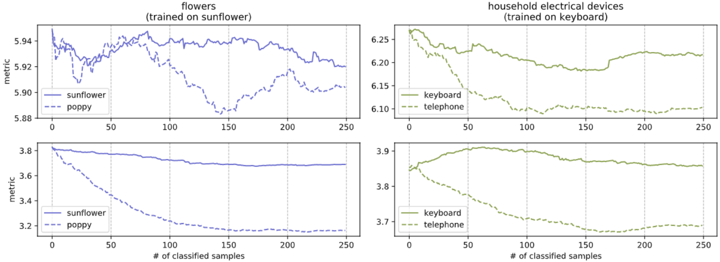


Figure 3: Metric scores for 2 randomly chosen example classes. Top row reports computations with the $V(\boldsymbol{C}_{\phi(\boldsymbol{x})}) = H(\tilde{\boldsymbol{\sigma}})$ definition while the bottom row reports computations with the $V(\boldsymbol{C}_{\phi(\boldsymbol{x})}) = H(\tilde{\boldsymbol{\lambda}})$ definition. Solid lines show the "constant" scenario and dashed lines show the "drift" scenario.

- Considering each super-class separately, most cases present lower metric values, under the same number of classified samples, for the "drift" scenario confirming the correctness of the formulated hypothesis with respect to the considered dataset/classifier pair.

Quantitative Metric Evaluation

- The separability of the scores associated to the "constant" and "drift" scenarios is investigated.
- The experiment analyzes the family of threshold triggers acting on the metric scores after the 10000 sample classifications.
- The investigation is performed by computing the Receiver Operating Characteristic (ROC) curves for both the $V(\boldsymbol{C}_{\phi(\boldsymbol{x})}) = H(\tilde{\boldsymbol{\sigma}})$ and $V(\boldsymbol{C}_{\phi(\boldsymbol{x})}) = H(\tilde{\boldsymbol{\lambda}})$ definitions.

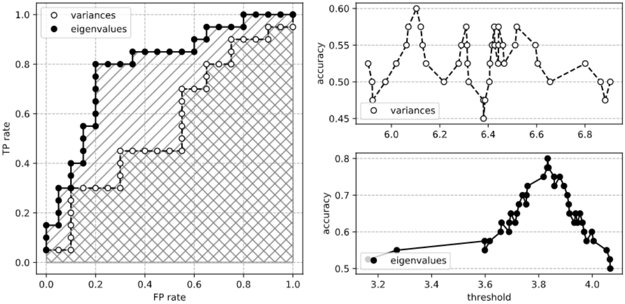


Figure 4: Quantitative evaluation of the considered scores/trigger pairs. The plot on the left reports the produced ROC curves while the plots on the right report the computed accuracies. White dotted lines refer to the $V(\boldsymbol{C}_{\phi(\boldsymbol{x})}) = H(\tilde{\boldsymbol{\sigma}})$ definition while black dotted lines refer to the $V(\boldsymbol{C}_{\phi(\boldsymbol{x})}) = H(\tilde{\boldsymbol{\lambda}})$ definition.

- The computation of the eigenvalues reveals to be necessary with a final AUC of 0.79, a net improvement over the direct use of the per-component variances, characterized by an AUC of 0.56.

Conclusion

- This paper presented a novel lifelong learning framework and metric in order to manage and quantify the intra-class variability of a trained classifier.
- The proposed work is an important step to extend the life of robots, thus enabling them to operate longer in real uncontrolled environments.
- For future work, we intend to fully implement the introduced framework and test the full framework's real-world performance on a robot platform.

Acknowledgment

The research leading to these results has received funding from the Austrian Science Fund (FWF) under grant agreement No. I3969-N30 (InDex). This work was partly supported by MIUR (Italian Minister for Education) under the initiative Departments of Excellence (Law 232/2016).

Samantha Guerriero, Barbara Caputo, and Thomas Mensink, *DeepNCM: Deep nearest class mean classifiers*, Proc. of International Conference on Learning Representations - Workshop, 2018.

Alex Krizhevsky, *Learning multiple layers of features from tiny images*, Tech. report, Department of Computer Science University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in Neural Information Processing Systems 25, 2012, pp. 1097–1105.

Jonathon Shlens, *A tutorial on principal component analysis*, ArXiv **abs/1404.1100** (2014).