Learning discriminative representations to interpret image recognition models

Environment: QARMA (machine learning) Team at Laboratoire d'Informatique et Systèmes (LIS)
Location: Ecole Centrale Marseille (ECM), Technopôle de Château-Gombert, Marseille
Supervision: **LIS:** R. Sicre, T. Artieres, S. Ayache; **INRIA Rennes:** Yannis Avrithis
Keywords: computer vision, deep learning, image recognition, interpretability, unsupervised learning

As for machine learning, computer vision has witnessed a fundamental change with the re-popularization of Deep Neural Networks (DNN) since 2012. Within a few years, DNN have been applied to various problems, such as image retrieval, object detection, instance segmentation, etc. These tasks benefit from the pre-training of networks on a large annotated corpus to obtain a superior visual representation that allows generalization and adaptation.

A large number of methods are based on intermediate representation learning to better encapsulate variation in parts of the data. In fine-grained recognition for instance, part-like representations are largely used [9, 2] and are clearly outperforming other methods. Chen *et al.* [3] introduce Prototype-agnostic Scene Layout to model scenes. Several detection methods propose to learn deformable models based on parts [4]. Detection of object proposals can also be used to improve image representation for retrieval [5]. While many methods use additional annotation such as bounding boxes, semantic part location, etc, several methods use only image-level labels or no supervision at all [10, 8, 11].

An additional benefit of these part-based methods is that the learned representations can often be visualized to gain understanding about inner workings of complex models. Recently, Chen *et al.* [2] improved interpretability by showing the contribution of latent parts to the final classification prediction.

This PhD aims at studying novel approaches to learn intermediate image representations. The objectives are to improve both recognition capabilities and interpretability of model predictions. While existing methods improve recognition and interpretability, several limitations remain: the computational cost, the inability to handle large datasets, complex optimization procedures, and the requirement for large amounts of annotation.

A first objective of this PhD is to address these limitations. Specifically, simple and efficient end-to-end methods will be investigated. Unsupervised representation learning will be investigated, adapting methods from self-supervised learning [1]. These methods will then be investigated to address a number of tasks and supervision settings, such as semi-supervised learning, metric learning, open-set recognition [7], few-shot learning, instance retrieval, and object detection.

A second objective is to produce better interpretation of model predictions. Several works enable interpretation, by using CAM [12] or grouping regions [6] for instance. Common part-based models already produce part-level information that can be linked to the classification prediction [2]. Following these works, improved methods will be investigated to learn intermediate representations that favour interpretability. Constraints will be considered on learned representations, so that these would be more discriminative, generative, binary, disentangled, etc. Such interpretation can help address other tasks with few or no labels and also help the user gain knowledge from data.

Contact: ronan.sicre@lis-lab.fr

References

[1] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2959–2968, 2019.

[2] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939, 2019.

[3] Gongwei Chen, Xinhang Song, Haitao Zeng, and Shuqiang Jiang. Scene recognition with prototype-agnostic scene layout. *arXiv preprint arXiv:1909.03234*, 2019.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[5] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257. Springer, 2016.

[6] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2020.

[7] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11814–11823, 2020.

[8] Ronan Sicre, Yannis Avrithis, Ewa Kijak, and Frédéric Jurie. Unsupervised part learning for visual recognition. In *Computer Vision and Pattern Recognition*, 2017.

[9] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. Technical report, 2015.

[10] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *Proceedings of the European Conference on Computer Vision*, pages 73–86. Springer, 2012.

[11] Jian Zhang, Runsheng Zhang, Yaping Huang, and Qi Zou. Unsupervised part mining for fine-grained image classification. *arXiv preprint arXiv:1902.09941*, 2019.

[12] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.