Thesis topic: Leveraging multimodal unlabeled data for visual scene understanding

Description

Multimodal learning has proven to be an effective way of leveraging multiple information for better decision-making [1]. In particular, in the context of computer vision, multimodal visual data provide complementary information about the same scene, thus enhancing the accuracy and robustness of complex scenes analysis. Several approaches have been proposed in literature, which mainly focus on the best fusion strategy of the different modalities [2].

However, most, if not all, machine learning methods used in practice are supervised approaches that rely on annotated training data. While these supervised techniques have shown excellent results, it has also been shown in literature that they often perform poorly under domains and distributions shifts [3, 4, 5]. From a theoretical point of view, the key assumption of classic supervised methods is the i.i.d. assumption, postulating that the training and test data are from the same distribution. When this assumption is violated, we face an out-of-distribution (O.O.D) generalization problem, which has to be addressed for better performances of learning methods. In real-world scenarios, the environment on which the vision system is deployed may drastically diverge from the training environment, leading to a domain and distribution shift. Moreover, the test distribution is typically unknown during training, and the system may encounter new unseen categories at test time, for example when operating in unconstrained or unseen environments.

Moreover, collecting and labeling large amount of data is a difficult and costly task. Thus, fully supervised methods cannot be used in many applications. Recently, self- supervised methods, and in particular contrastive learning approaches have been successfully developed in different application areas, including visual scene analysis and understanding [6, 7].

On the other hand, in training a deep learning model, all training examples are randomly presented to the model, ignoring the various complexities of data samples and the current learning status of the model. Whereas, it has been show that carefully selecting the order in which to present training data for learning improves the generalization capacity and convergence rate of the model [8]. This learning strategy known as curriculum learning consist in training a model from easier to harder examples. The basic idea is to train a model with easier data subsets (or easier subtasks) and gradually increase the difficulty level of the data (or subtasks) until the complete training dataset is used [9].

Thus, the goal of this project is threefold : i) avoid the cost and burden of dense annotations by employing self-supervised learning approaches, ii) benefit from multimodal information to tackle difficult recognition cases, and iii) explore the use of curriculum learning for better training.

In particular, we aim to leverage multimodal information to answer the following questions:

- How to use data from different modalities for self-supervised learning?
- How to measure difficulty? That is how to decide the relative “easiness” of a training example?
- How to select examples? That is how to decide the order in which data are used in the training process, because a data sample considered “easy” in one modality can be “difficult” in another and vice-versa?

Therefore, this research topic opens news questions never addressed before in the literature and can provide new ideas and methods for different tasks such as segmentation, detection or recognition.

Applications

Applicants with a strong background in machine learning, computer vision, and related topics are invited to apply.

A master degree in Computer Sciences, Mathematics or related fields is required.

Please send your resume, a motivation letter and transcripts of your Bachelor and Master programs to:

- Prof. Désiré Sidibé, drodesire.sidibe@univ-evry.fr
- Associate Prof. Dominique Fourer, dominique.fourer@univ-evry.fr

Application deadline: May 10th, 2024.

References

[1] Baltruaitis et al. "Multilmodal machine learning: a survey and taxonomy", IEEE PAMI, 2017
[2] Zhang et al. "Deep multimodal data fusion for semantic image segmentation: a survey", Image and Vision Computing, 2021
[3] M. Arjovsky et al. "Invariant risk minimization", arXiv:1907.02893, 2019
[4] E. Creager et al. "Environment inference for invariant learning", ICML, 2021
[5] Z. Shan et al. "Stable learning via sample reweighting", AAAI, 2020
[6] T. Chen, et al., "A simple framework for contrastive learning of visual representations", ICML 2020
[7] R. Balestriero et al., "A coockbook of self-supervised learning", arXiv:2304.12210, 2023
[8] Bengio et al. "Curriculum learning", ICML, 2009
[9] Wang et al. "A survey on curriculum learning", IEEE PAMI, 2021