Investigation of modality-specific information in multimodal representation learning

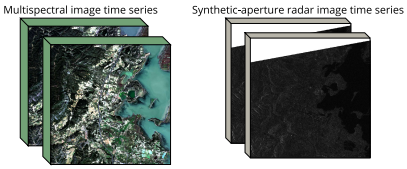
Laboratory MIA-Paris-Saclay - Mathématiques et Informatique Appliquées

Supervision Romain Thoreau, Vincent Guigue

Internship description

For many automatic tasks, **combining multiple data modalities has great potential** (Baltrušaitis et al., 2018), from movie genre classification (Fig. 1) to remote sensing (Fig. 1b). For instance, the combination of optical and radar satellite images has been proven to be valuable for **crop mapping** (Garnot et al., 2022) or **flood detection** (Rambour et al., 2020).

(a) A multimodal sample of MM-IMDb: a dataset for movie genre classification



(b) Optical and radar satellite image time series

Figure 1: Examples of multimodal data

In many real-world applications where labeled data are scarce, **self-supervised methods have become central in representation learning**; i.e. the process of extracting compressed, task-agnostic yet meaningful features from data. Large models are usually pretrained through a self-supervised task, and then finetuned on a supervised downstream task (Hu et al., 2022; Thoreau et al., 2025b). In particular, contrastive learning (CL) has recently gained considerable traction in order to learn representations from unlabeled multimodal data (Liang et al., 2024). The pretext task pertaining to CL methods consists in aligning (in terms of cosine similarity) the representations of the modalities in the latent

space. The idea of **alignment is based on the multi-view redundancy assumption, stating that task-relevant information is shared across modalities** (Tsai et al., 2020). Obviously, the multi-view redundancy assumption does not hold in many multimodal problems. For example, a medical image and a patient record, or altimetry and optical satellite data, provide complementary and non-redundant information about potential pathologies and forest properties, respectively. Xue et al. (2023) have formalized how **multimodal information divides into modality-generic and modality-specific information**. In this spirit, **recent works have challenged the capacity of CL algorithms to extract modality-specififc features** from multimodal data (Liang et al., 2023; Dufumier et al., 2025; Thoreau et al., 2025a).

The goal of this research internship is to deepen our understanding of machine learning models trained by contrastive learning when multimodal data provide rich and task-relevant modality-specific information.

— The first objective of the internship is to formalize when multimodal machine learning models trained by contrastive learning are expected to fail on downstream tasks (e.g. classification). In particular, the intern will introduce metrics in order to measure the gap between modality-specific and modality-generic information in multimodal samples.

— The second part of the internship will focus on visual and textual multimodal data sets and models. The intern will introduce specific data augmentation techniques in order to increase the gap between modality-generic and modality-specific information in multimodal samples. This data augmentation process will allow the intern to design an experimental protocol for studying CL methods in a controlled setting.

Candidate profile We are looking for a master student in machine learning, applied mathematics or computer science.

References

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Dufumier, B., Castillo-Navarro, J., Tuia, D., and Thiran, J.-P. (2025). What to align in multimodal contrastive learning? In *International Conference on Learning Representations*. ICLR.

Garnot, V. S. F., Landrieu, L., and Chehata, N. (2022). Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:294–305.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen,

W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Liang, P. P., Deng, Z., Ma, M. Q., Zou, J. Y., Morency, L.-P., and Salakhutdinov, R. (2023). Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998.

Liang, P. P., Zadeh, A., and Morency, L.-P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42.

Rambour, C., Audebert, N., Koeniguer, E., Le Saux, B., Crucianu, M., and Datcu, M. (2020). Flood detection in time series of optical and sar images. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:1343–1346.

Thoreau, R., Levillain, J., and Derksen, D. (2025a). Can multimodal representation learning by alignment preserve modality-specific information? *Accepted to MACLEAN workshop at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), 2025. arXiv preprint arXiv:2509.17943*.

Thoreau, R., Marsocci, V., and Derksen, D. (2025b). Parameter-efficient adaptation of geospatial foundation models through embedding deflection. *Accepted to IEEE/CVF International Conference on Computer Vision 2025. arXiv preprint arXiv:2503.09493*.

Tsai, Y.-H. H., Wu, Y., Salakhutdinov, R., and Morency, L.-P. (2020). Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*.

Xue, Z., Gao, Z., Ren, S., and Zhao, H. (2023). The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *The Eleventh International Conference on Learning Representations*.