Robust self-supervised learning for multimodal data

Context

Multimodal learning Multimodal learning [1] refers to the process of extracting information from multiple sources, such as text, audio, images, or tabular data. This approach is particularly significant in real-world applications where data inherently exists in diverse formats, such as in the medical field (clinical reports, medical imaging, dynamic health recordings), or autonomous driving (RGB cameras, LiDAR, and radar sensors). Utilizing multiple modalities is beneficial, as it not only improves predictive accuracy but also enhances the model's interpretability by providing additional sources of explanation for its predictions.

Multimodal aligment A wide variety of multimodal models are learned though contrastive learning which aims at extracting similar representations in a joint latent space [12]. These approaches are particularly successful to encode the information shared between modalities. Once learned, these models are powerful swiss army knifes able to retrieve information from a given modality [12, 6, 11]. However these approaches fail to encode precious modality specific information which ends up discarded. More complex dependencies between the modalities such as incompatibility or complementarity which give strong insights into the data are also not modeled. To address this oversimplification in the modality representation, recent works proposed new training strategies including complementary information and incompatibility [4, 13, 10].

Uncertainty quantification Machine learning models are often designed by assuming that the test data is similar to the training data. In practice, ambiguous, out-of-distribution (OOD), or outlier samples may occur, which compromise the reliability of the predictions. Being robust and quantify model and data uncertainty is therefore crucial in sensitive fields such as autonomous driving and healthcare. Many methods have been proposed in computer vision for uncertainty quantification [2, 8]. In multi-modal contexts, some works aim at quantifying uncertainties from the different sources [7, 14]. As discussed previously, when these approaches rely on contrastive learning, the sole focus of feature alignment can lead to a loss of information. For instance, modality-specific attributes that should be flagged as OOD could be lost when embedded in a common latent space that only focuses on the information shared by modalities.

Objectives & work plan

To tackle these challenges, this thesis aims to develop new methods for quantifying uncertainty in complex multimodal representations. The work plan is as follows:

- Evaluate the gain related to the multimodal strategies that preserve modality-specific information [4, 13] in typical robustness tasks such as OOD detection or failure prediction [3]. (1st year of PhD)

- Develop extension of well-known uncertainty quantification method [5, 9] to leverage modality specificity and interactions. The main challenge is that methods from [4, 13] representations that can entangle modality-specific information and well as cross-modal interactions (complementarity or incompatibilities). This requires thus careful adaptations to provide a latent space suitable to interpretable (modality-wise and interaction-wise) uncertainty quantification. (2nd and 3rd year of PhD).

Alignment with the initiative

This project is well aligned with the PostGenAI@Paris initiative, as it addresses both scientific and societal challenges by focusing on trustworthy AI. The supervision team is affiliated with research units at Sorbonne University and CNAM (both of which are part of the initiative). This thesis will thus reinforce the collaboration between the two teams.

Supervision team

Consortium: CEDRIC, Conservatoire National des Arts et Metiers & ISIR, Sorbonne Université

Supervision: Members of the supervisory team have contributed to recognized recent research in multimodal training and uncertainty quantification.

Arnaud Breloy (director)	CEDRIC CNAM	Robust statistics, Information geometry	[website]
Javiera Castillo Navarro	CEDRIC CNAM	Multimodal learning, Self-supervised learning	[website]
Clément Rambour	ISIR SU	Robustness in IA, Multimodal learning	[website]

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[2] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez. Addressing failure prediction by learning model confidence. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[3] H. Dong, Y. Zhao, E. Chatzi, and O. Fink. Multiood: Scaling out-of-distribution detection for multiple modalities. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

[5] D. Hendrycks and K. Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*, Nov. 2016.

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[7] M. Lafon, E. Ramzi, C. Rambour, N. Audebert, and N. Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282. Springer, 2024.

[8] M. Lafon, E. Ramzi, C. Rambour, and N. Thome. Hybrid energy based model in the feature space for out-of-distribution detection. In *International Conference on Machine Learning*, 2023.

[9] K. Lee, K. Lee, H. Lee, and J. Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[10] P. P. Liang, Z. Deng, M. Ma, J. Zou, L.-P. Morency, and R. Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[11] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[13] S. Swetha, M. N. Rizve, N. Shvetsova, H. Kuehne, and M. Shah. Preserving modality structure improves multi-modal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21993–22003, 2023.

[14] U. Upadhyay, S. Karthik, M. Mancini, and Z. Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1899–1910, 2023.