Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

SUBJECT:

Multi-Modal Continual Learning

SUPERVISOR:

dr hab. inż. Szymon Łukasik

NASK-PIB

dr inż. Bartłomiej Twardowski

IDEAS NCBR / Computer Vision Center Universitat Autònoma de Barcelona

DESCRIPTION:

The traditional approach to deep learning is to train once on a large amount of data. However, in many systems, the distribution of data changes over time, or new tasks, classes or objects appear. A simple attempt to adapt the model results in a strong forgetting phenomenon, i.e. a degraded ability of the model to solve previous tasks correctly.

A simple solution would be to retrain the model on the original data with the addition of new data. However, this leads to excessive resource consumption. A more optimal solution is to use continual learning methods that retrain the model in a way that limits catastrophic forgetting.

Current methods have been developed primarily in domains such as image processing. The area of continual learning in a multimodal setting, such as vision-text models, has not yet been widely explored. However, learning from an everchanging stream of multimodal data pose some unique challenges, i.e., *misalignment between modalities* when only one modality is changed and finetuned with a new data. Then, the catastrophic forgetting is expected in finetuned modality, however, the influence on the other modalities is not investigated.

The recent advancements in Large Language Models (LLMs) and foundation models demonstrate impressive performance even in multimodal settings, such as information retrieval. These models are extensively large and trained on vast amounts of data collected from the Internet over many years. When evaluated using common benchmarks for single modalities, they show similar performance. However, recent models from OpenAI exhibit less robustness in retrieval tasks compared to the latest models from the OpenCLIP repository (more up-to-date models), highlighting their susceptibility to evolving data distributions over time and a need for update [10].

Another critical aspect of this research is sustainability and energy efficiency. Continuing the current trend of training ever-larger LLMs and foundation models from scratch demands increasingly significant financial investment in computational resources. This approach is neither sustainable nor efficient. Multimodal continual learning directly addresses this issue by enhancing knowledge retention and enabling constant knowledge accumulation. Instead of starting training from scratch, we can continually evolve the model, thus preventing performance degradation over time.

The purpose of this study is to explore a more holistic approach to the multimodal continual learning, considering both the most promising methods for counteracting forgetting, as well as ways to counteract misalignment, and the possibility of using interpretive methods to reliably investigate the source of the problem and propose appropriate solutions.

The candidate should propose and test new methods for detecting the phenomenon of misalignment for text-vision models in a continual learning setup. The candidate is expected to explore ways to increase the interpretability of the mechanisms that take place during the learning process in order to better identify the reasons for misalignment and finally propose an appropriate method to mitigate it.

BIBLIOGRAPHY:

1. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., & You, Y. (2023). Preventing zero-shot transfer degradation in continual learning of vision-language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 19125-19136).
2. Yu, J., Zhuge, Y., Zhang, L., Wang, D., Lu, H., & He, Y. (2024). Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters. arXiv preprint arXiv:2403.11549.
3. Cai, Y., Thomason, J., & Rostami, M. (2023). Task-attentive transformer architecture for continual learning of vision-and-language tasks using knowledge distillation. arXiv preprint arXiv:2303.14423.
4. Wang, K., Herranz, L., & van de Weijer, J. (2021). Continual learning in cross-modal retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3628-3638).
5. Srinivasan, T., Chang, T. Y., Pinto Alva, L., Chochlakis, G., Rostami, M., & Thomason, J. (2022). Climb: A continual learning benchmark for vision-and-language tasks. Advances in Neural Information Processing Systems, 35, 29440-29453.
6. Ni, Z., Wei, L., Tang, S., Zhuang, Y., & Tian, Q. (2023, July). Continual vision-language representation learning with off-diagonal information. In International Conference on Machine Learning (pp. 26129-26149). PMLR.

7. Skantze, G., & Willemsen, B. (2022). Collie: Continual learning of language grounding from language-image embeddings. Journal of Artificial Intelligence Research, 74, 1201-1223.
8. Wang, L., Zhang, X., Su, H., & Zhu, J. (2024). A comprehensive survey of continual learning: Theory, method and application. IEEE Transactions on Pattern Analysis and Machine Intelligence.
9. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural networks, 113, 54-71.
10. Garg, S., Farajtabar, M., Pouransari, H., Vemulapalli, R., Mehta, S., Tuzel, O., Shankar, V. and Faghri, F., 2024. TiC-CLIP: Continual Training of CLIP Models. ICLR.
11. Del Chiaro, R., Twardowski, B., Bagdanov, A., & Van de Weijer, J. (2020). Ratt: Recurrent attention to transient tasks for continual image captioning. Advances in Neural Information Processing Systems, 33, 16736–16748.
12. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A. D., & Van De Weijer, J. (2022). Class-incremental learning: survey and performance evaluation on image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), 5513–5533.
13. Yu, L., Twardowski, B., Liu, X., Herranz, L., Wang, K., Cheng, Y., Jui, S., & Weijer, J. van de. (2020). Semantic drift compensation for class-incremental learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6982–6991.