Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

Links to Further Works I

- FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture [Li et al., 2024]
- Cola: Large Language Models are Visual Reasoning Coordinators [Chen et al., 2023b, Chen et al., 2023a]
- Improving language-supervised object detection with linguistic structure analysis [Rai and Kovashka, 2023]
- Linguistic Structures as Weak Supervision for Visual Scene Graph Generation [Ye and Kovashka, 2021]
- Clue: Cross-modal Coherence Modeling for Caption Generation [Alikhani et al., 2020] [Alikhani et al., 2022]
- Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color [Abdou et al., 2021]
- Benchmarking Vision Language Models for Cultural Understanding [Nayak et al., 2024]
- Situated Instruction Following [Min et al., 2024]

References I

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. (2021).
Can language models encode perceptual structure without grounding? a case study in color.
In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132.

Alikhani, M., Han, F., Ravi, H., Kapadia, M., Pavlovic, V., and Stone, M. (2022).
Cross-modal coherence for text-to-image retrieval.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10427–10435.

Alikhani, M., Sharma, P., Li, S., Soricut, R., and Stone, M. (2020).
Cross-modal coherence modeling for caption generation.
In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535.

Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., and Liu, Z. (2023a).
Language models are visual reasoning coordinators.
In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., and Liu, Z. (2023b).
Large language models are visual reasoning coordinators.
Advances in Neural Information Processing Systems.

Li, W., Zhang, X., Li, J., Peng, Q., Tang, R., Zhou, L., Zhang, W., Hu, G., Yuan, Y., Søgaard, A., Hershcovich, D., and Elliott, D. (2024).
Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture.
CoRR, abs/2406.11030.

Min, S. Y., Puig, X., Chaplot, D. S., Yang, T.-Y., Rai, A., Parashar, P., Salakhutdinov, R., Bisk, Y., and Mottaghi, R. (2024).
Situated instruction following.
In *ECCV*.

References II

Nayak, S., Jain, K., Awal, R., Reddy, S., van Steenkiste, S., Hendricks, L. A., Stańczak, K., and Agrawal, A. (2024).
Benchmarking vision language models for cultural understanding.
arXiv preprint arXiv:2407.10920.

Rai, A. and Kovashka, A. (2023).
Improving language-supervised object detection with linguistic structure analysis.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5569.

Ye, K. and Kovashka, A. (2021).
Linguistic structures as weak supervision for visual scene graph generation.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8289–8299.