Title: **Understanding Visual Processing in Vision Language Models**

Abstract:
Does language help make sense of the visual world? How important is it to actually see the world rather than having it described with words? These basic questions about the nature of intelligence have been difficult to answer because we only had one example of an intelligent system–humans–and limited access to cases that isolated language or vision.

However, the development of sophisticated Vision-Language Models (VLMs) by artificial intelligence researchers offers us new opportunities to explore the contributions that language and vision make to learning about the world. Given the unique potentials offered by language, we investigate how the VLM visual understanding process might utilize language. We hypothesize a computational process for image understanding in VLMs that involves semantic extraction of images and test this hypothesis by ablating components from the cognitive architecture of VLMs. In doing so, we identify the contributions of each component in the architecture to VLM visual understanding. We find that if VLMs process images via semantic extraction, they likely require three key components: prior knowledge, non-trivial reasoning abilities, and semantically meaningful examples to refine internal representations.

Textbooks:

1. Computer Vision: Algorithms and Applications (Szelinski) [second edition], 2022/2023
    - Chapter 5: Deep learning
    - Chapter 6: Recognition
2. Deep Learning by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
    - Chapter 2: Linear Algebra
    - Chapter 5: Machine Learning Basics
    - Chapter 9: Convolutional Networks
    - Chapter 15: Representation Learning
3. The Cambridge Handbook of Computational Psychology
    - Chapter 12: Mental Logic, Mental Models, and Simulations of Human Deductive Reasoning
    - Chapter 23: Computational Modeling of Visual Information Processing
4. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information (Marr)
    - Chapter 1: The Philosophy and the Approach
5. "The Language of Thought Hypothesis" by Michael Rescorla. *The Stanford Encyclopedia of Philosophy (Winter 2023 Edition)*, Edward N. Zalta & Uri Nodelman (eds.)
6. "Modularity of Mind" by Phillip Robbins, *The Stanford Encyclopedia of Philosophy (Winter 2017 Edition)*, Edward N. Zalta (ed.)

Research Papers:
Vision Language Models

1. Liu, Haotian, et al. "Visual Instruction Tuning." *Advances in neural information processing systems* 36 (2024).
2. Radford, Alec, et al. "Learning Transferable Visual Models from Natural Language supervision." *International conference on machine learning. PMLR*, 2021.

Language Model Analysis

1. Marjieh, Raja et al. “Words are all you need? Language as an Approximation for Human Similarity Judgments.” *International Conference on Learning Representations* (2022).
2. Menon, Sachit and Carl Vondrick. “Visual Classification via Description from Large Language Models.” *International Conference on Learning Representations* (2023)
3. Wilcox, Ethan Gotlieb , et al. “On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior.” *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2020.
4. Shiffrin, Richard, and Melanie Mitchell. "Probing the Psychology of AI Models." *Proceedings of the National Academy of Sciences* 120.10 (2023): e2300963120.
5. Webb, Taylor, Keith J. Holyoak, and Hongjing Lu. "Emergent Analogical Reasoning in Large Language Models." *Nature Human Behaviour 7.9* (2023): 1526-1541.
6. Mitchell, Melanie, Alessandro B. Palmarini, and Arseny Moskvichev. "Comparing Humans, GPT-4, and GPT-4V on Abstraction and Reasoning Tasks." *arXiv preprint arXiv:2311.09247* (2023).
7. Webb, Taylor, Shanka Subhra Mondal, and Jonathan D. Cohen. "Systematic Visual Reasoning Through Object-centric Relational Abstraction." *Advances in Neural Information Processing Systems 36* (2024).

Language in Human Visual Understanding:

1. Kim, Judy Sein, et al. "Shared Understanding of Color Among Sighted and Blind Adults." *Proceedings of the National Academy of Sciences* 118.33 (2021): e2020192118.
2. Saysani, Armin, Michael C. Corballis, and Paul M. Corballis. "Colour Envisioned: Concepts of Colour in the Blind and Sighted." *Visual cognition 26.5* (2018): 382-392.