Exposé for Bachelor's Thesis

GazeControl

Florentin Doll
Supervisor: Tomáš Daniš
Advisor: Prof. Dr. Martin V. Butz
Date: May 21, 2025

1. Introduction and Motivation

Humans solve complex visual reasoning tasks every day with remarkable speed and accuracy. In contrast, computer vision systems typically process all pixels of an image at a uniform resolution, overlooking the natural attentional focus provided by foveated vision. By emulating human-like vision—allocating high resolution to task-relevant regions and lower resolution elsewhere—artificial agents can become more efficient and their decision-making more interpretable by revealing which areas drew their attention.

2. Research Question and Objectives

Research Question: Can we build artificial agents that, like humans, sequentially "look" at parts of an image using a limited foveal window and integrate observations over time to solve visual reasoning tasks more efficiently?
My main objective is to train a model that mirrors human visual attention and decision processes. To this end, I will:

1. Model the human fovea and implement a variable-focus sampling mechanism.
2. Integrate a memory module (, GRU or LSTM) to accumulate evidence across fixations.
3. Learn a reinforcement learning policy (, Actor-Critic or PPO) that decides at each step whether to shift gaze or to issue a final decision.
4. Add a decoder that visualizes the internal memory state at each time step, providing insights into the agent's reasoning.

3. Related Literature and Resources

Datasets

- **Compositional Visual Relations (CVR)** (Zerroug *et al.*, 2022): A benchmark of 103 compositional odd-one-out tasks evaluating sample efficiency and generalization.

- **Pathfinder Challenge** (Linsley *et al.*, 2018): Tasks requiring the agent to determine whether two marked points are connected by a continuous path, testing long-range spatial integration.

Key Related Work

- **Recurrent Attention Model (RAM)** (Mnih *et al.*, 2014): A pioneering hard-attention network using reinforcement learning to select sequential high-resolution glimpses for image classification.
- **Active Vision RL under Limited Observability** (Shang and Ryoo, 2023): An RL framework that simultaneously learns motor and gaze policies, rewarding sensorimotor alignment to improve task performance.
- **Horizontal GRU (hGRU)** (Linsley *et al.*, 2018): A convolutional RNN inspired by cortical association fields that excels on the Pathfinder dataset, demonstrating efficient long-range grouping.

4. Methodology

Our architecture will combine:

- A **foveated sampler** to extract high-resolution features at the current fixation (via blurring + CNN or a differentiable foveation module).
- A **memory module** (GRU/LSTM) that integrates new glimpses with past information.
- An **actor** that, based on the memory state, chooses between a gaze shift or issuing the final classification.
- A **decoder** that upsamples the memory state into an image, visualizing the agent's internal representation.

To train the model, we will rely primarily on a reinforcement learning loss: correct decisions are rewarded, incorrect ones are penalized, and a small penalty is applied for each additional step taken to encourage efficiency. However, relying solely on this sparse reward signal may result in slow or unstable learning.

To support and guide training, we introduce an auxiliary decoder loss. The decoder reconstructs the agent's current memory state into an image, and we compare this reconstruction either to the original image (autoencoder-style) or to a feature-based representation. Instead of using a full image loss—which could force the model into premature and inaccurate reconstructions—we will mask the loss to only include regions that the model has already seen. This can be implemented using a heatmap that tracks visited locations. Alternatively, we may extract features from the reconstructed and original images using the same encoder and compare them at the feature level to prioritize meaningful information over exact pixel values. This allows the decoder to guide learning without interfering with the sequential nature of attention.