Where to look next – guiding active visual exploration with internal model uncertainty

Computer vision. In recent years, Deep Learning has revolutionized the field of Computer Vision, propelling it to unprecedented heights. Current deep neural network models based on the vision transformer (ViT) architecture have been proven not only to achieve state-of-the-art performance, but also to generalise into multiple domains without the need for additional training. However, as artificial neural networks have grown in size and complexity, so have their computational demands and energy consumption. Furthermore, traditional computer vision solutions often assume complete access to input data, overlooking the challenges posed by real-world applications. Active visual exploration (AVE) aims to be a remedy for those problems.

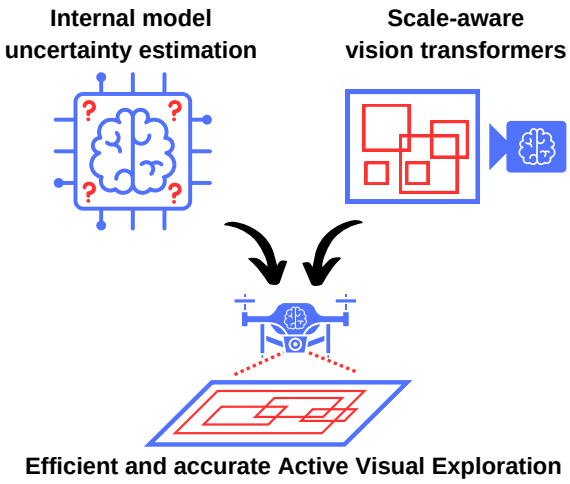

Figure 1: This project combines internal model uncertainty estimation and vision transformer architecture to create efficient accurate active visual exploration methods. To achieve this goal, we introduce exploration strategies based on model uncertainty and adapt the vision transformer architecture to support multi-scale input.

Active visual exploration. AVE introduces a new paradigm, allowing machines to explore their environments actively, much like humans do. Instead of processing the entire visual field at maximum resolution, AVE empowers the agent to selectively sample and focus computational resources on crucial areas. By intelligently determining where and at what resolution to gather visual information, AVE promises more efficient and accurate computer vision. However, current solutions for AVE have serious restrictions. Firstly, they rely on complex training, leading to computationally and memory intensive algorithms. Secondly, they use fixed-size visual samples, disregarding most basic abilities of robotic platforms such as camera zoom or freely rotating cameras. This project seeks to address these limitations by creating a new branch in active visual exploration research.

Neural network uncertainty. The first aspect of this project focuses on leveraging the internal uncertainty estimation capabilities of vision transformers. By incorporating this uncertainty into the AVE process, agents can make informed decisions about where to explore to maximize situation awareness and actively seek out regions where the model predictions are ambiguous, leading to a more accurate understanding of the environment. This novel approach avoids the need for additional trainable modules and reduces complexity.

Embracing Scale and Flexibility. Another key objective is to make the vision transformer model scale-aware and suitable for AVE tasks. Currently, ViTs use fixed grid sampling, hindering their ability to capture crucial details in varying situations. This project proposes modifying the input layers of ViTs to accept arbitrary sampled patches, enabling the robotic camera ability to zoom in and out of scenes.

Optimizing Glimpse Selection. The final stage of this project aims to refine the glimpses selection process. By integrating the knowledge gained from uncertainty estimation and scale-awareness, the selection process will determine which areas require multiple high-resolution patches and which can be covered with low-resolution glimpses. This optimized approach will significantly reduce the number of glimpses needed while maintaining high accuracy.

Implications and Future Prospects. The success of this project holds vast implications for various domains, including robotics, autonomous vehicles, and image analysis. Embodied agents, such as robots or unmanned aerial vehicles, will benefit from more efficient and accurate visual exploration, enabling them to navigate complex and dynamic environments effectively. Furthermore, the advancements made in ViT architectures and active exploration strategies can lead to more sustainable AI systems, reducing computational demands and energy consumption.