Active Visual Exploration Based on Attention-Map Entropy

Adam Pardyl [1,2] Grzegorz Rypeść [1,3] Grzegorz Kurzejamski [1] Bartosz Zieliński [1,2,5] Tomasz Trzciński [1,2,3,4]

[1]IDEAS NCBR [2]Jagiellonian University [3]Warsaw University of Technology [4]Tooploox [5]Ardigen

#2446

Abstract

Active visual exploration addresses the issue of limited sensor capabilities in real-world scenarios, where successive observations are actively chosen based on the environment. To tackle this problem, we introduce a new technique called Attention-Map Entropy (AME). It leverages the internal uncertainty of the transformer-based model to determine the most informative observations. In contrast to existing solutions, it does not require additional loss components, which simplifies the training. Through experiments, which also mimic retina-like sensors, we show that such simplified training significantly improves the performance of reconstruction, segmentation and classification on publicly available datasets.

Active visual exploration

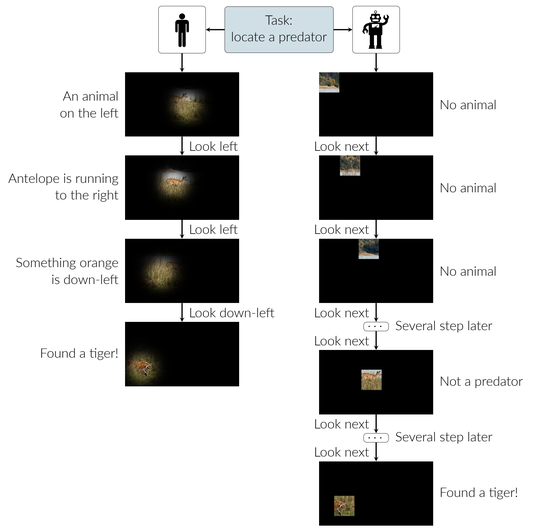

Figure 1. **Visual exploration – human versus AI**: Humans naturally visually explore surrounding environment, using already observed areas as clues to where the wanted object can be located [1]. At the same time, common state-of-the-art artificial intelligence solutions analyze all available data, which is inefficient and waste time and computational resources. In this project, we introduce a novel Active Visual Exploration method, enabling AI agents to efficiently explore their environment.

Method

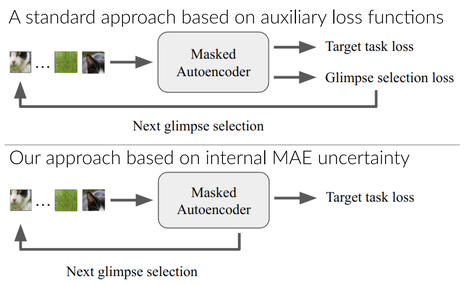

Figure 2. **Attention-Map Entropy (AME):** Our approach chooses the most informative observations by reusing the internal uncertainty coded in the attention maps. In contrast to existing methods, it does not require any auxiliary loss functions dedicated to active exploration. Therefore, the training concentrates on the target task loss, not on an auxiliary loss, which improves overall performance.

Architecture

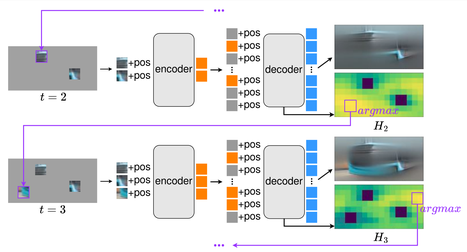

Figure 3. **Architecture for reconstruction:** The agent observed two patches of the image, which are processed by the encoder to produce their feature representations (orange rectangles). These outputs are combined with the masked patches (shown as gray rectangles) and passed through the decoder. The decoder reconstructs the missing image patches. Additionally, our method generates the entropy map for one of the decoder's multi-head self-attention layers and uses it to select the location of the third glimpse. The process repeats till we reach the assumed number of glimpses.

Attention-map entropy

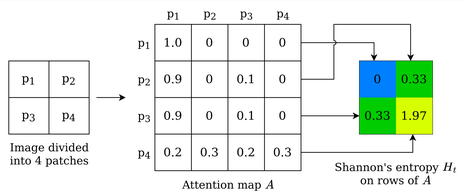

Figure 4. **Entropy map:** To explain the idea of the entropy map based on attention in the transformer layer, let us consider an image divided into four patches (2×2) on the left. Its attention map will be a 4×4 matrix, where each row represents the attention weights used to calculate the output in the next transformer layer for a corresponding patch. Calculating Shannon's entropy for each row will result in a 2×2 entropy map. The patch with the highest entropy value is selected as the next glimpse.

Active visual exploration example

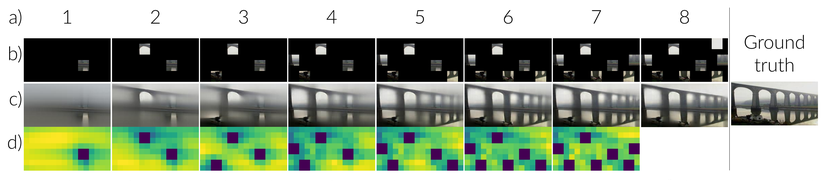

Figure 5. **Glimpse-based reconstruction step-by-step:** The figure shows a glimpse selection process based on AME for 8×32^2 glimpses for a sample 256×128 image. The rows correspond to a) step number, b) model input (glimpses), c) model prediction given, d) decoder attention entropy (known areas are explicitly set to zero). The algorithm explores the image in places where the reconstruction result is blurry.

Results

Method	SUN360	ADE20k	MSCOCO	Image res.	Glimpse regime	Pixel %	Area %
AttSeg	37.6	36.6	41.8	128×256	8×48^2 (retinal)	18.75	56.25
GlAtEx	33.8	41.9	40.3	128×256	8×48^2 (retinal)	18.75	56.25
Ours (retinal)	**23.6**	**23.8**	**25.2**	128×256	8×48^2 (retinal)	18.75	56.25
Ours (non-retinal)	37.9	40.7	43.2	128×256	8×16^2 (non ret.)	6.25	6.25
Ours (non-retinal)	29.8	30.8	32.5	128×256	8×32^2 (non ret.)	25.00	25.00
Ours (non-retinal)	**20,1**	**20.6**	**22.1**	128×256	8×48^2 (non ret.)	56.25	56.25
SimGlim (detached)	26.2	27.2	29.8	224×224	37×16^2 (non ret.)	18.75	18.75
SimGlim (end-to-end)	28.0	28.8	31.3	224×224	37×16^2 (non ret.)	18.75	18.75
Ours (non-retinal)	**23.4**	**26.2**	**28.6**	224×224	37×16^2 (non ret.)	18.75	18.75

Table 1. **Reconstruction results:** Comparison of our model in reconstruction task against AttSeg [5], GlAtEx [4] and SimGlim [2] on SUN360 [6], ADE20K [7] and MS COCO [3] datasets. The metric used is a root mean square error (RMSE; lower is better). For each experiment, we provide a training and evaluation regime defined by a number of glimpses of a specific resolution. Pixel % and area % denote respectively: the percentage of image pixels known to the model and the percentage of image area seen by the model. Differences in both measures occur when dealing with retina-like glimpses, which have lower pixel counts by design. Our method outperforms competitive solutions in all configurations.

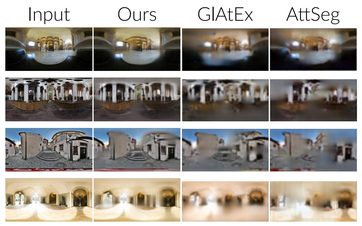

Figure 6. **Reconstruction quality for SUN360:** Reconstruction results of our method compared with AttSeg [5] and GlAtEx [4] on the SUN360 dataset. Reconstructions done with our method are visibly better than the competition.

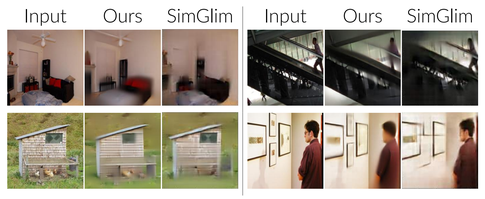

Figure 7. **Reconstruction quality for ADE20K:** Figure shows difference in reconstruction quality against SimGlim [2] on the ADE20K dataset. SimGlim reconstructs a single object slightly better, but our method recovers more objects in the scene.

Method	mPA(%)	PA(%)	IoU(%)
AttSeg	-	47.9	-
GlAtEx	-	52.4	-
Ours (MAE-weights)	32.2	**70.27**	24.4
Ours (SETR-weights)	**35.6**	69.5	**27.6**

Table 2. **Segmentation results:** Comparison of our model against AttSeg [5] and GlAtEx [4]. The metric used is mean Pixel-wise Accuracy (mPA, higher is better), Pixel-Accuracy (PA, higher is better), and Intersection over Union (IoU, higher is better). Our solution outperforms competitive methods on all metrics.

Method	Accuracy (%)
AttSeg	52.6
GlAtEx (full)	56.4
GlAtEx (no decoder)	67.2
Ours (head-only)	70.1
Ours (train-all)	**75.7**

Table 3. **Classification results:** Comparison of our model's classification performance against AttSeg [5] and GlAtEx [4] on the SUN360 dataset. The metric used is accuracy (higher is better). Our method outperforms competitive methods in both Train-all and Head-only training options.

Results cont.

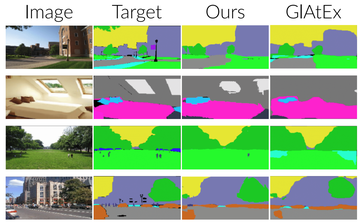

Figure 8. **Segmentation quality for ADE20K:** Semantic segmentation results of our method compared with GlAtEx [4] on the ADE20k dataset. Qualitatively, segmentation maps produced by our method are at least as good as those of the competition.

Acknowledgments

This research was funded by National Science Centre, Poland (grants no 2020/39/B/ST6/01511, 2022/45/B/ST6/02817, and 2021/41/B/ST6/01370), Foundation for Polish Science (grant no POIR.04.04.00-00-14DE/18-00 carried out within the Team-Net program co-financed by the European Union under the European Regional Development Fund). We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2022/015753. The research was supported by a grant from the Faculty of Mathematics and Computer Science under the Strategic Programme Excellence Initiative at Jagiellonian University.

References

[1] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

[2] Abhishek Jha, Soroush Seifi, and Tinne Tuytelaars. Simglim: Simplifying glimpse based active visual reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 269–278, January 2023.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] Soroush Seifi, Abhishek Jha, and Tinne Tuytelaars. Glimpse-attend-and-explore: Self-attention for active visual exploration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16137–16146, 2021.

[5] Soroush Seifi and Tinne Tuytelaars. Attend and segment: Attention guided active semantic segmentation. *CoRR*, abs/2007.11548, 2020.

[6] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[7] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

Read the paper at

IJCAI/2023 MACAO