This is the peer reviewd version of the followng article:

Towards Explainable Navigation and Recounting / Poppi, Samuele; Rawal, Niyati; Bigazzi, Roberto; Cornia, Marcella; Cascianelli, Silvia; Baraldi, Lorenzo; Cucchiara, Rita. - 14233:(2023), pp. 171-183. (22nd International Conference on Image Analysis and Processing (ICIAP 2023) Udine, Italy September 11-15, 2023) [10.1007/978-3-031-43148-7_15].

Springer Science and Business Media Deutschland GmbH
Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

05/03/2026 22:42

(Article begins on next page)

Towards Explainable Navigation and Recounting

Samuele Poppi[0000−0002−8428−501X], Roberto Bigazzi[0000−0002−6457−1860], Niyati Rawal[0000−0002−4142−0488], Marcella Cornia[0000−0001−9640−9385], Silvia Cascianelli[0000−0001−7885−6050], Lorenzo Baraldi[0000−0001−5125−4957], and Rita Cucchiara[0000−0002−2239−283X]

University of Modena and Reggio Emilia, Italy
{name.surname}@unimore.it

Abstract. Explainability and interpretability of deep neural networks have become of crucial importance over the years in Computer Vision, concurrently with the need to understand increasingly complex models. This necessity has fostered research on approaches that facilitate human comprehension of neural methods. In this work, we propose an explainable setting for visual navigation, in which an autonomous agent needs to explore an unseen indoor environment while portraying and explaining interesting scenes with natural language descriptions. We combine recent advances in ongoing research fields, employing an explainability method on images generated through agent-environment interaction. Our approach uses explainable maps to visualize model predictions and highlight the correlation between the observed entities and the generated words, to focus on prominent objects encountered during the environment exploration. The experimental section demonstrates that our approach can identify the regions of the images that the agent concentrates on to describe its point of view, improving explainability.

Keywords: Explainable AI · Visual Navigation · Image Captioning.

1 Introduction

Recent advances in the field of Embodied AI aim to foster the next generation of autonomous and intelligent mobile agents and robots. Research in this field includes visual navigation [10, 29], object-driven navigation [11], and the creation of new research platforms for simulation of embodied agents [33]. While this line has focused on providing mobile robots with perception and action capabilities, it is indubitable that future agents will need to interact seamlessly with human beings. In this regard, the research at the intersection of Computer Vision and NLP is of particular interest to the community, as it can provide robots with the ability to connect what they perceive with their linguistic abilities. Also, being able to describe what the robot sees can be a valuable means of explainability and bridge the gap between the black-box architecture and the user.

In this paper, we concentrate on a novel Embodied AI setting, which aims both to connect Vision and Language and provide explainability. In our setting, a robot navigates an unknown environment and has the ability to describe in

natural language what it sees. In particular, the agent needs to perceive the environment around itself, navigate it driven by an exploration goal, and describe salient objects and scenes in natural language. Beyond navigating the environment and translating visual cues in natural language, the agent also needs to identify appropriate moments to perform the explanation step, *i.e.*, it needs to employ an appropriate speaking policy. It is worthwhile to note that this setting poses challenges from different perspectives. Firstly, exploring an environment without any previous knowledge, nor a reference trajectory is, by itself, a significant challenge for Embodied AI. Secondly, while describing visual content in Natural Language has been previously addressed by the image captioning community [39], existing approaches have been applied to either natural or web-scale images – and applications of image captioning to images taken from the perspective of a mobile robot are still very limited in literature [12]. Moreover, an appropriate speaking policy needs to be defined to interconnect the navigation and captioning components of the approach and select appropriate moments for describing the visual perception. Lastly, for the generated descriptions to be a valuable explainable mean, their quality should be properly assessed, and their content properly aligned with the robot's actual perception.

In the following, we jointly address all the previously mentioned points. We devise an exploration strategy based on a surprisal reward [24], in which the agent is trained in a self-supervised manner to explore previously unknown environments. Further, we endow our agent with a Transformer-based captioner that can generate natural language descriptions. Lastly, we devise a component that can measure the explanation capability of the caption by aligning its content with what the visual encoder is attending from the input image. Experimentally, we assess the performance of the proposed approach on the Matterport3D dataset [9] by employing the Habitat simulation platform [33]. In particular, we analyze the navigation, captioning, and explainability capabilities of the model and demonstrate the appropriateness of the proposed solutions. Overall, our proposal makes a step forward in the direction of mobile agents that can be explainable by design and interact with human beings in natural language.

2 Related Work

As the goal of this work is to enable embodied agents to generate user-explainable descriptions of the perceived environment, this section describes state-of-the-art research in explainable AI, image captioning, and embodied exploration.

Explainable AI. Explainability properties are becoming an increasingly desired feature for deep learning models, especially when these models are employed by final users and not by experts. When dealing with models for Computer Vision applications, several explainability approaches have been developed, depending on the specific task to perform and the type of features used [16, 23, 26, 20, 13, 27]. In this work, we borrow ideas from the image classification task, where the predominant approach in literature is that of building a saliency map out of the visual encoder, achieving some level of explainability by visualizing the most

salient regions in the input image. Subsequently, an explanation map can be computed as the pixel-wise multiplication of the input image and the saliency map. Many different approaches have been proposed to obtain explanation maps, like visualization tools [37, 45], gradient-based approaches [38, 40], and Class Activation Mapping (CAM)-based methods [46, 35, 42]. In this work, we resort to a CAM-based approach [46] for providing explanation maps.

Image Captioning. Being a task at the intersection between vision and language, image captioning has benefited from the technical advancements in both these fields. The goal of the task is to generate a natural language description of a given image. To this end, the image must be properly represented. In most works, this has been done by employing convolutional neural networks to extract global or grid features [19, 43], or image region features containing visual entities [1]. More recent approaches employ fully-attentive Transformer-like architectures [41] as visual encoders, which can also be applied directly to image patches [13]. The image representation is used to condition a language model that generates the caption. The language model can be implemented as a recurrent neural network [19, 7, 18] or Transformer-based fully-attentive models [15, 14, 32]. In this work, we develop a fully-attentive image captioning approach as our captioning module, as it is potentially suitable to be employed in a wide variety of real-world settings such as the embodied exploration one.

Embodied Exploration. Research in Embodied AI has witnessed increasing interest from the community, thanks to the development of photo-realistic 3D simulation environments [9, 33] that bring exploration and navigation agents one step closer to real-world deployment. Those agents are commonly trained with reinforcement learning adopting a modular, hierarchical approach [10] where the agent learns to explore the environment by optimizing a self-supervision signal in the form of a reward function [10, 29, 30, 4]. Once trained on the aforementioned simulators, the developed agents can be easily deployed on physical robotic platforms [6]. However, many state-of-the-art architectures are still considered black boxes, as their behavior lacks explainability [2]. In this respect, some attempts have been made to make the robot navigation and decision-making processes more interpretable for the end-user by letting it produce natural language descriptions of what it observes [5, 12, 3]. In this work, we consider a curiosity-driven exploration agent [30] and equip it with the ability to produce natural language descriptions of what it observes while navigating the environment, also exploiting explainable maps to enhance the interpretability of the descriptions.

3 Proposed Method

In our proposed approach, a deep reinforcement learning exploration agent navigates an unknown environment and collects interesting views according to a heuristic speaker policy. These images are then passed to an encoder-decoder captioning model, combined with an explainability technique to provide a user-

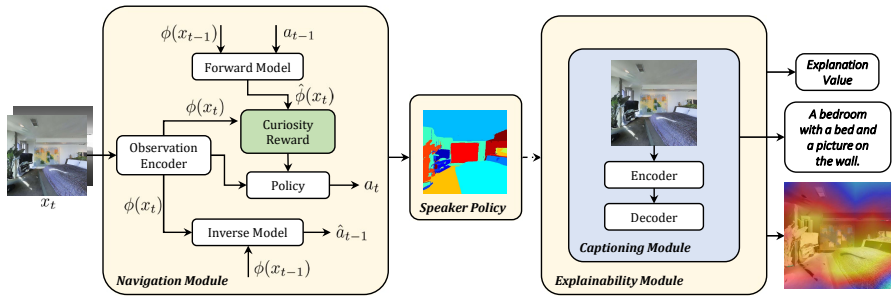

Fig. 1. Schema of the proposed "navigation and recounting" framework.

understandable interpretation of the environment internal representation of the agent. An overview of the proposed approach is depicted in Fig. 1.

3.1 Navigation Module

While moving inside the environment, our agent captures an RGB-D image from its current position x_t, which is encoded via a convolutional neural network ϕ. Note that, during training, the encoding network is maintained fixed for guaranteeing the stability of the resulting features during training, thus resulting in a more efficient training of our agent, as demonstrated in [8].

Our agent can move inside the environment by taking atomic actions at each timestep, $a_t \in \{\textit{turn left 15°},\ \textit{turn right 15°},\ \textit{move forward 0.25m}\}$. After the execution of each action, the agent captures a new observation x_{t+1}. From the observation x_t and action a_t, we can define the forward dynamics problem of predicting the next observation as $\hat{\phi}(x_{t+1}) = f(\phi(x_t), a_t; \theta_F)$, where $\hat{\phi}(x_{t+1})$ is the predicted visual embedding for x_{t+1} and f is the forward dynamics model, whose parameters are θ_F. Moreover, given two consecutive observations (x_t, x_{t+1}), we can define the inverse dynamics problem of predicting the action a_t performed between the two observations as $\hat{a}_t = g(\phi(x_t), \phi(x_{t+1}); \theta_I)$, where $\hat{a}_t$ is the predicted estimate for the action a_t and g is the inverse dynamics model, whose parameters are θ_I. The parameters θ_F and θ_I of the dynamics models are determined by minimizing respectively $\mathcal{L}_F$ and $\mathcal{L}_I$ losses:

$$\mathcal{L}_F = \frac{1}{2}\left\|\hat{\phi}(x_{t+1}) - \phi(x_{t+1})\right\|_2^2, \quad \text{and} \quad \mathcal{L}_I = y_t \log \hat{a}_t, \tag{1}$$

where y_t is the one-hot representation of a_t.

Note that the actions the agent performs at each timestep are selected by a policy $\pi(\phi(x_t); \theta_\pi)$, that is trained to maximize the expected sum of a reward expressed as the discrepancy of the predictions of dynamics models and the actual observation, *i.e.*,

$$\max_{\theta_\pi} \mathbb{E}_{\pi(\phi(x_t);\theta_\pi)}\left[\sum_t r_t\right] \quad \text{s.t.} \quad r_t = \frac{\eta}{2}\left\|f\big(\phi(x_t), a_t\big) - \phi(x_{t+1})\right\|_2^2 - p_t, \tag{2}$$

where η is a scaling factor and p_t is a penalty factor to prevent the agent from repeating the same action multiple consecutive times. In this work, p_t is set to 0 and then becomes equal to a constant value $\tilde{p}$ if the same action is repeated $\tilde{t}$ times. By combining the loss functions in Eq. 1 and 2, we obtain the overall optimization problem:

$$\min_{\theta_\pi,\theta_F,\theta_I} \Big[-\lambda \mathbb{E}_{\pi(\phi(x_t);\theta_\pi)} \Big[\sum_t r_t \Big] + \beta \mathcal{L}_F + (1-\beta)\mathcal{L}_I \Big] \tag{3}$$

where λ and β are regularization factors.

3.2 Object-driven Speaker Policy

As the navigation proceeds, the relevant objects in the scene can be recognized from the observation x_t. Based on the analysis presented in [5], in this work, we adopt an object-driven speaker policy that elicits the description of the observation if a minimum number of salient objects are present in the scene. By adopting this policy, the captioning module produces descriptions when a sufficient number of characteristic objects are observed, which allow connoting, and thus, distinguishing, that view of the environment. In this work, we set to five the minimum number of objects in the scene for it to be described.

3.3 Captioning Module

The goal of the captioning module is that of modeling an autoregressive distribution probability $p(\boldsymbol{w}_t|\boldsymbol{w}_{\tau<t}, \mathbf{V})$, where $\mathbf{V}$ is an image captured from the agent and $\{\boldsymbol{w}_t\}_t$ is the sequence of words comprising the generated caption. This is usually achieved by training a language model conditioned on visual features to mimic ground-truth descriptions.

We represent each training image-caption pair as a pair of image and text $(\mathbf{V}, \mathbf{W})$, where $\mathbf{V}$ is encoded with a set of fixed-length visual descriptors. The text input is tokenized with lower-cased Byte Pair Encoding [36] with a vocabulary of 49,152 tokens. For multimodal fusion, we employ an encoder-decoder Transformer [41] architecture. Each layer of the encoder employs multi-head self-attention (MSA) and feed-forward layers, while each layer of the decoder employs multi-head self- and cross-attention (MSCA) and feed-forward layers. For enabling text generation, sequence-to-sequence attention masks are employed in each self-attention layer of the decoder. The visual descriptors $\mathbf{V} = \{\boldsymbol{v}_i\}_{i=1}^N$ are encoded via bi-directional attention in the encoder, while the token embeddings of the caption $\mathbf{W} = \{\boldsymbol{w}_i\}_{i=1}^L$ are inputs of the decoder, where N and L indicate the number of visual embeddings and caption tokens, respectively. The overall network operates according to the following schema:

$$\begin{aligned} &\text{encoder} \quad && \tilde{\boldsymbol{v}}_i = \text{MSA}(\boldsymbol{v}_i, \mathbf{V}) \\ &\text{decoder} \quad && \mathbf{O}_{\boldsymbol{w}_i} = \text{MSCA}(\boldsymbol{w}_i, \tilde{\mathbf{V}}, \{\boldsymbol{w}_t\}_{t=1}^i), \end{aligned} \tag{4}$$

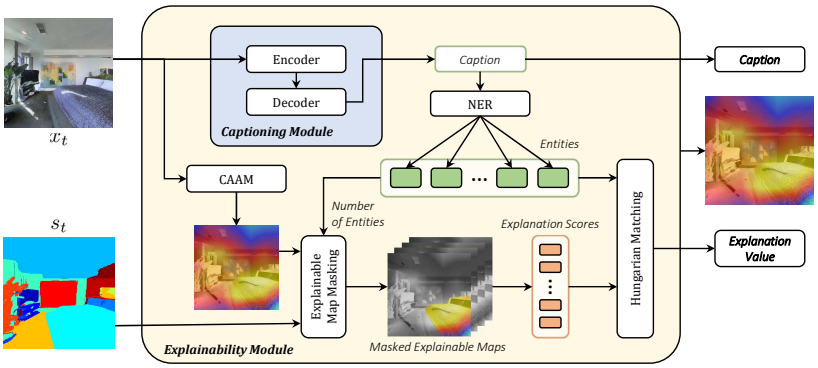

Fig. 2. Schema of the explainability module of our approach.

where $\mathbf{O}$ is the network output, $\text{MSA}(\boldsymbol{x}, \mathbf{Y})$ a self-attention with $\boldsymbol{x}$ mapped to query and $\mathbf{Y}$ mapped to key-values, and $\text{MSCA}(\boldsymbol{x}, \mathbf{Y}, \mathbf{Z})$ a self-attention with $\boldsymbol{x}$ as query and $\mathbf{Z}$ as key-values, followed by cross-attention with $\boldsymbol{x}$ as query and $\mathbf{Y}$ as key-values. We omit feed-forward layers and the dependency between consecutive layers for ease of notation. The captioning network is trained with a unidirectional language modeling loss based on cross-entropy.

Inference. Once the model is trained, at each time step t, it samples a token $\hat{\boldsymbol{w}}_t$ from the output probability distribution. This is then concatenated to previously predicted tokens to form a sequence $\{\hat{\boldsymbol{w}}_\tau\}_{\tau=1}^{t}$, which is employed as the input for the next iteration. Since the representation of output tokens does not depend on subsequent tokens, the past intermediate representations are kept in memory to avoid repeated computation and increase efficiency at prediction time.

Visual features. To obtain the set of visual features $\mathbf{V}$ for an image, we employ a visual encoder pre-trained to match vision and language [28]. Compared to using features extracted from object detectors [1], our strategy is beneficial in terms of both computational efficiency feature quality. Specifically, we use one of the encoders proposed in CLIP [28], which can be either based on CNNs or Vision Transformers. In both cases, we employ the entire grid of features coming from the last visual encoder layer.

3.4 Captioning Assessment via Explanation Maps

As a measure of the explainability degree of the caption, given the output of the captioner, we align its content with portions of the image on which the visual encoder focuses. Given an input image x, we firstly compute a Class-Agnostic Activation Map (CAAM) inspired by the Class Activation Mapping [46] literature. This is defined as a linear combination of the activation maps of the last layer of the encoder, weighted by their mean score, as follows:

$$\text{CAAM}(x) = \text{ReLU}\left(\sum\nolimits_{k=1}^{N_l} \alpha_k A_k\right), \tag{5}$$

where N_l denotes the number of activation maps, A_k is the k-th channel of the activation, and α_k are weight coefficients indicating the importance of each activation map, each of them defined as the average of A_k over the two spatial axes. A ReLU activation is employed to consider only the features that have a positive influence. Then, given a semantic segmentation map of the input image, *i.e.*, a set of binary masks, each associated to a semantic class, we average the values of CAAM(x) over each mask to obtain an "explanation score" for each semantic class. The explanation score of a semantic class indicates how much the network has focused on that semantic class. Finally, we align explanation scores with respect to concepts present in the caption. To this end, we firstly extract nouns from the generated caption by using Named Entity Recognition[1]. We then build a matching between the set of nouns in the caption and the set of semantic classes found in the segmentation map by using the Hungarian algorithm and GloVe embeddings [25] as similarity measure. With the obtained association between nouns mentioned in the caption and semantic concepts, we take a weighted average of explanation scores, according to the weights defined by the Hungarian matching for the optimal association. The final explanation value is a measure of how much the caption aligns with what the visual encoder is actually observing, and therefore it is an indirect score for the explainability power of the caption. In the following, we refer to this explanation value as EXPL-S.

4 Experiments

4.1 Exploration Experiments

The exploration capability of the agent influences the variability and information content of the observations extracted from the environment. We qualitatively validate the navigation module presenting the trajectories of the agent on some sample test episodes.

Implementation details. In our experiments, we use Habitat simulator [33] with scenes from the Matterport3D dataset [9], for which dense semantic annotations are available with 40 different object categories. The observations for our navigation agent and the subsequent modules have a 640×480 resolution. Before being fed to the navigation module, the observations are resized to 84×84 grayscale images. Furthermore, instead of using a single observation for the current timestep, we use the last four observations stacked together for modeling temporal dependencies. Stacked grayscale observations are fed to the observation encoder to compute the features used by the dynamics models and the policy. The navigation module has been trained for 10K updates on the training set of Matterport3D dataset for a total of $\approx$ 1.3M frames, then, experiments are done on unseen environments of the test split. Episodes have a maximum length T of 500 and 1000 steps for the training and the testing phases, respectively.

[1] `https://spacy.io/`

The policy and the dynamics models are trained using PPO [34] The parameters in Eq. 3 are set to $\lambda = 0.1$ and $\beta = 0.9$. The penalty introduced in Eq. 2 is triggered if the agent repeats the same action for $\tilde{t} = 5$ and assumes a value $p_t = \tilde{p} = 0.01$.

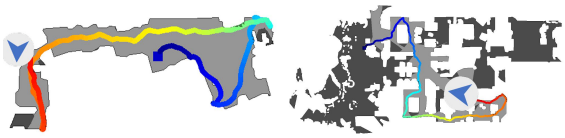

Fig. 3. Qualitative results of the agents' trajectories in sample exploration episodes.

Exploration results. In Fig. 3 we show the agent's trajectory on the top-down environment map in some sample episodes. These results show that the agent is able to explore small environments completely while navigating efficiently for the duration of the time budget to optimize the area seen in larger environments.

4.2 Captioning and Explainability Experiments

In the considered setting, the explainability properties of the navigation agent depend on the quality of the captions produced by the captioning module and to their explainability power. In this section, we evaluate these two aspects.

Datasets. To assess the performance of the captioning and explainability modules, we first consider the COCO image captioning dataset [22]. Specifically, we follow the splits defined in [19], which consists of 113,287 images for training, 5,000 images for validation, and 5,000 for test. The images depict people and common-use objects and come with five ground-truth captions each. Moreover, for evaluating these modules on the Matterport3D dataset, we use the images gathered from the robot that are considered to be worth describing according to the speaker policy. In particular, these are the images in which the number of objects is greater than five, resulting in 16,828 images. Note that no ground-truth caption is available for these images, which are therefore used only in inference.

Implementation details. We consider four variants of the captioning module, each of them based on a different CLIP-based visual encoder [28]. In particular, we consider three encoders based on CNNs (*i.e.*, CLIP-RN50, CLIP-RN50×4, and CLIP-RN50×16), and an encoder based on a Vision Transformer (*i.e.*, CLIP-ViT-B16). In all the variants, the decoder takes as inputs d-dimensional vectors with $d = 384$, and has $l = 3$ layers and $H = 6$ attention heads. To represent the position of the input words, we exploit the sinusoidal positional encoding as in [41]. We train the four variants of the captioning module by optimizing a standard cross-entropy loss with the LAMB optimizer [44]. We employ the learning rate scheduling strategy proposed in [41], which entails a warmup of 6,000 iterations whose resulting learning rate is multiplied by 5, and minibatch size equal to 1,080. We additionally fine-tune the models with the SCST strategy, by employing the Adam optimizer [21] and fixed learning rate 5×10^{-6}.

Evaluation setup. For evaluating the performance of the captioning module on the COCO dataset, we consider the standard image captioning metrics (*i.e.*, BLEU-4, METEOR, ROUGE, CIDEr, SPICE) [39] and, following recent

Table 1. Captioning results on the COCO Karpathy-test split.

	BLEU-4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S
CLIP-RN50	36.9	28.1	57.5	125.2	21.5	0.738
CLIP-ViT-B16	38.6	29.2	58.8	132.5	23.0	0.749
CLIP-RN50×4	39.2	29.2	58.8	132.9	22.7	**0.753**
CLIP-RN50×16	**40.2**	**29.5**	**59.4**	**137.0**	**23.1**	0.750

Table 2. Captioning and explainability results on the images gathered from the agent in the Matterport3D dataset, and described according to the speaker policy.

	CLIP-S	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	EXPL-S
CLIP-RN50	0.697	0.492	0.569	0.625	0.721	**0.670**
CLIP-ViT-B16	**0.722**	0.521	0.591	0.651	0.735	0.694
CLIP-RN50×4	0.721	0.520	0.591	0.649	0.732	0.623
CLIP-RN50×16	0.719	**0.528**	**0.605**	**0.654**	**0.738**	0.554

advancements in the field [17, 31], the CLIP-S [17] in its reference-free definition. As for the evaluation on Matterport3D, in which no ground-truth captions are available, the standard metrics mentioned above cannot be computed. For this reason, we consider a variant of the soft coverage score, computed between the set of nouns mentioned in the caption and the set of semantic classes in the image, as defined in [5]. In the considered variant, only objects whose area is greater than a threshold are included in the compared sets. Moreover, we compute the CLIP-S, as done for the images from COCO, and the EXPL-S proposed.

Captioning and explainability results. First, we evaluate the performance of the captioning module alone on the COCO dataset. The results of this analysis are reported in Table 1. From the table, it can be observed that the CLIP-RN50×16 variant is the best performing in terms of the classical paired metrics and the second-best in terms of the unpaired CLIP-S metric. In terms of this latter metric, CLIP-RN50×4 performs best. The performance of the CLIP-ViT-B16 is close to that of the mentioned variants. These results suggest that the considered captioners have the ability to generate meaningful and grammatically correct captions, and that larger visual encoders can generally improve caption quality and quantitative performance.

We also evaluate the considered variants for the captioning module on the images from the Matterport3D dataset. The results of this analysis are reported in Table 2. It emerges that, when applied to these images, CLIP-ViT-B16 is the best performing variant in terms of CLIP-S, while CLIP-RN50×4 is the second-best. The obtained scores are in line with those obtained on the COCO dataset, indicating the generalization capabilities of the considered captioners. In terms of the proposed EXPL-S, the variant with the highest explanation power is CLIP-ViT-B16. In terms of Coverage, the variant with CLIP-RN50×16 achieves the best result, while CLIP-ViT-B16 is the second best. This underlines that CLIP-RN50×16 can name more coherent semantic concepts, while CLIP-ViT-B16 is better terms of both caption quality and explainability power.

The high absolute values of EXPL-S achieved by most of the visual encoders outline that there is a significant agreement between nouns mentioned in the

Fig. 4. Qualitative captioning and explainability results. For each image, we report the generated caption, the saliency map produced by CAAM, and the EXPL-S score.

caption and regions attended by the visual encoders. This can also be observed in Fig. 4, where we show qualitative results on Matterport3D environments. For each sample, we report the RGB view of the agent, the generated caption, and the saliency map produced by CAAM, which is the basis for the computation of the EXPL-S score. As it can be noticed, the EXPL-S score quantifies the degree of alignment between the generated caption and the saliency map.

5 Conclusion

In this work, we have presented an embodied agent for exploration, whose internal representation of the environment can be interpreted even by non-expert users. This has been achieved by equipping the agent with the ability to produce a natural language description of the observed scene when the scene is deemed interesting according to a speaking policy. In addition, we have defined an explanation map-based score to measure the explainability power of the description produced, which matches the explanation map, the produced caption, and the observed scene. The experimental results have shown that the proposed approach is a viable solution to gain insights into the perception and navigation capabilities of embodied agents.

Acknowledgements This work has been supported by the "Fit for Medical Robotics" (Fit4MedRob) project, funded by the Italian Ministry of University and Research, and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 955778 for project "Personalized Robotics as Service Oriented Applications" (PERSEO).

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
2. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: AAMAS (2019)

3. Bigazzi, R., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: Embodied Agents for Efficient Exploration and Smart Scene Description. In: ICRA (2023)
4. Bigazzi, R., Landi, F., Cascianelli, S., Baraldi, L., Cornia, M., Cucchiara, R.: Focus on Impact: Indoor Exploration with Intrinsic Motivation. RA-L **7**(2), 2985–2992 (2022)
5. Bigazzi, R., Landi, F., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: Explore and Explain: Self-supervised Navigation and Recounting. In: ICPR (2020)
6. Bigazzi, R., Landi, F., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: Out of the Box: Embodied Navigation in the Real World. In: CAIP (2021)
7. Bolelli, F., Baraldi, L., Pollastri, F., Grana, C.: A Hierarchical Quasi-Recurrent approach to Video Captioning. In: IPAS (2018)
8. Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., Efros, A.A.: Large-scale study of curiosity-driven learning. arXiv preprint arXiv:1808.04355 (2018)
9. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D Data in Indoor Environments. In: 3DV (2017)
10. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning To Explore Using Active Neural SLAM. In: ICLR (2019)
11. Chaplot, D.S., Gandhi, D.P., Gupta, A., Salakhutdinov, R.R.: Object Goal Navigation using Goal-Oriented Semantic Exploration. In: NeurIPS (2020)
12. Cornia, M., Baraldi, L., Cucchiara, R.: SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability. In: ICRA (2020)
13. Cornia, M., Baraldi, L., Cucchiara, R.: Explaining Transformer-based Image Captioning Models: An Empirical Analysis. AI Communications **35**(2), 111–129 (2022)
14. Cornia, M., Baraldi, L., Fiameni, G., Cucchiara, R.: Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training. arXiv preprint arXiv:2111.12727 (2022)
15. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-Memory Transformer for Image Captioning. In: CVPR (2020)
16. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: ECCV (2016)
17. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2021)
18. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on Attention for Image Captioning. In: ICCV (2019)
19. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
20. Kim, S.S., Meister, N., Ramaswamy, V.V., Fong, R., Russakovsky, O.: HIVE: Evaluating the Human Interpretability of Visual Explanations. In: ECCV (2022)
21. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV (2014)
23. Lovino, M., Bontempo, G., Cirrincione, G., Ficarra, E.: Multi-omics classification on kidney samples exploiting uncertainty-aware models. In: ICIC (2020)
24. Pathak, D., Agrawal, P., Efros, A.A., Darrell, T.: Curiosity-driven exploration by self-supervised prediction. In: ICML (2017)
25. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. In: EMNLP (2014)
26. Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Revisiting The Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis. In: CVPR Workshops (2021)

27. Poppi, S., Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: Multi-Class Explainable Unlearning for Image Classification via Weight Filtering. arXiv preprint arXiv:2304.02049 (2023)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021)
29. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: Occupancy Anticipation for Efficient Exploration and Navigation. In: ECCV (2020)
30. Ramakrishnan, S.K., Jayaraman, D., Grauman, K.: An Exploration of Embodied Visual Exploration. IJCV **129**, 1616–1649 (2021)
31. Sarto, S., Barraco, M., Cornia, M., Baraldi, L., Cucchiara, R.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In: CVPR (2023)
32. Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: Retrieval-Augmented Transformer for Image Captioning. In: CBMI (2022)
33. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019)
34. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347 (2017)
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: ICCV (2017)
36. Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. In: ACL (2016)
37. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
38. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
39. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From Show to Tell: A Survey on Deep Learning-based Image Captioning. IEEE Trans. PAMI **45**(1), 539–559 (2022)
40. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML (2017)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
42. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: CVPR Workshops (2020)
43. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
44. You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., Hsieh, C.J.: Large batch optimization for deep learning: Training bert in 76 minutes. In: ICLR (2019)
45. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
46. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)