Learning to Combine Vision and Language: Towards Novel Multimodal Solutions for Media Technologies

Lorenzo Baraldi[1,*], Marcella Cornia[1] and Rita Cucchiara[1,2]

[1]University of Modena and Reggio Emilia, Modena, Italy
[2]IIT-CNR, Pisa, Italy

Abstract

Research on Vision-and-Language integration is a fundamental ingredient towards the development of novel multimodal solutions for media technologies and is nowadays one of the most productive and flourishing research areas in AI. This paper provides an overview of the research activities carried out at the AImageLab laboratory of the Department of Engineering "Enzo Ferrari" of the University of Modena and Reggio Emilia, which has been actively working in this area for more than five years. The task being addressed include image captioning, cross-modal retrieval, the adaptation of Large Language Models and Multimodal Foundational Models, as well as the integration of Vision, Language and Action in the context of the development of navigation algorithms for personalized robotics.

Keywords

Vision-and-Language, Multimodal Learning, Image Captioning, Embodied AI

1. Introduction

The integration of Vision and Language is a core research field in Multimedia which stands at the intersection of Computer Vision and Natural Language Processing. AImageLab, the research laboratory at the "Enzo Ferrari" Department of Engineering of the University of Modena and Reggio Emilia, has been actively tackling this research area for more than five years from now, addressing core Vision-and-Language (V&L) tasks such as *image captioning*, *cross-modal retrieval*, *visual question answering* and, more recently, the integration and customization of *foundational models* like Large Language Models or Multimodal Large-Scale models. Beyond the pure integration of the visual and the textual modality, the group has also developed a significant expertise in multimodal models which combine Vision, Language and Action, such as *navigation models for embodied agents*, and human-robot interaction algorithms.

This research is conducted within different European and national projects, such as the PERSEO MSCA project for personalized robotics, the CREATIVE PRIN project for generative and multimodal AI, the ELSA European Project ("European Lighthouse on Secure and Safe AI"), the FIT4MEDROB PNRR project for medical robotics, and the more recent FAIR (Future AI Research) PNRR project and its transversal project on Vision and Language. A close collaboration is also active with CINECA and NVIDIA, in the context of the NVIDIA AI Technology Centre of Modena.

This paper presents an overview of some of the most recent research activities the group has carried out, ranging from enhancing V&L models with long-tail capabilities [1] to building learnable metrics with high correlation with the human judgment [2] Finally, we will also provide a progress report on the development of navigation algorithms which integrate Vision, Language, and Explainability [3, 4, 5].

2. Describing Images in Natural Language

Image captioning aims at generating textual descriptions from visual inputs. As such, it entails modeling the connections between the visual and textual modalities and can be seen as a fundamental step toward machine intelligence [6]. In the following, we describe our recent research activities in this domain, focusing on both architectural advances and the evaluation of generated captions.

2.1. Separating Semantics and Style for Image Captioning

Recent works automatically collected large-scale datasets with noisy image-text pairs from the web, partially solving the semantic scale issue of popularly-used human-annotated datasets like COCO [8], but at the cost of reducing the quality of the annotations (see upper part of

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
lorenzo.baraldi@unimore.it (L. Baraldi); marcella.cornia@unimore.it (M. Cornia); rita.cucchiara@unimore.it (R. Cucchiara)
0000-0001-5125-4957 (L. Baraldi); 0000-0001-9640-9385 (M. Cornia); 0000-0002-2239-283X (R. Cucchiara)

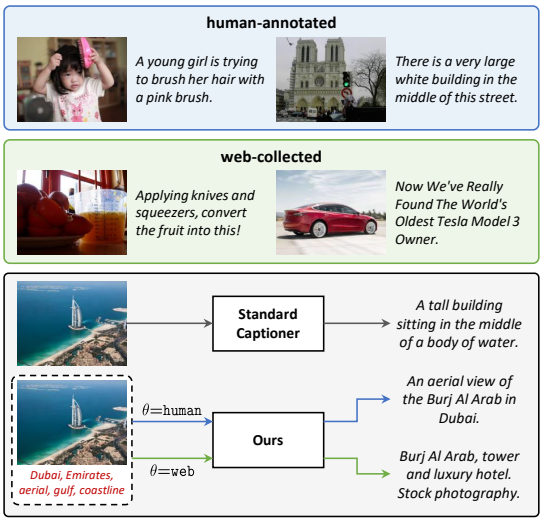

Figure 1: Samples of human-annotated and web-collected (image, caption) pairs and overview of our approach with multi-source training [7].

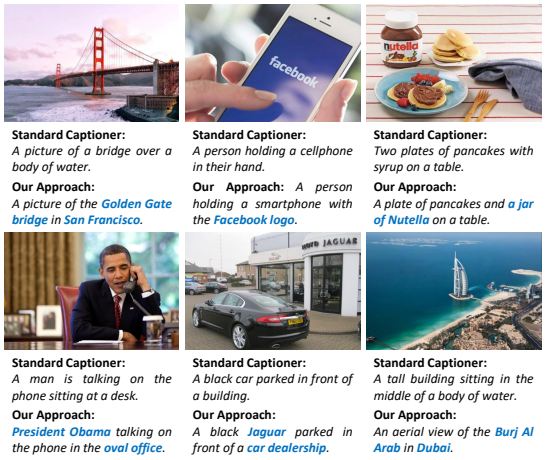

Figure 2: Sample descriptions generated by our model in comparison with a Transformer-based captioner trained on COCO. Our approach generates high-quality captions by separating content from style.

Figure 1). Some attempts that directly train on noisy datasets [9, 10] have been proposed, but generally generate captions with low quality, while others solutions that perform a self-supervised pre-training [11, 12] do not take into account the difference in descriptive style between sources. To overcome these issues, we focused on generating pertinent captions which can be richer in terms of semantics and include proper names and long-tail concepts, by jointly leveraging web-collected and human-annotated sources and maintaining the style and fluency of human-annotated captions [7].

The core idea behind our approach is that of separating semantics and descriptive style while training on non-homogeneous data sources. This is achieved through the introduction of a *style token* that can condition the network both at training and generation time. During training, the token is employed to distinguish between human-annotated and web-crawled sources. At generation time, the style token can be used to generate human-like descriptions enriched by the semantics learned on web-collected datasets.

Further, to better represent semantics, we extracted textual keywords through a novel retrieval-based approach [13], which avoids the need of using tags or descriptions from object detectors [14, 11]. This also allowed us to scale beyond a limited set of categories and fully represent the semantics of the image regardless of its source. The addition of the style token and of textual keywords foster the transfer of descriptive style and semantic concepts between data sources. An overview of our architecture is shown in Figure 1.

From an experimental point of view, our model outperforms existing proposals in terms of caption quality, sometimes also surpassing models trained on significantly larger datasets [12], and shows an improved capability of generating named entities to improve the description pertinence (Figure 2). Overall, this research activity demonstrates that heterogeneous data sources can be properly exploited, together with a selective architecture, to increase the performance of image captioning systems.

2.2. Evaluating Captions with Positive-Augmented Constrastive Learning

The task of image captioning has not only witnessed methodological and architectural innovations but also the proposal of different evaluation metrics, shifting from early translation metrics [15, 16] to more effective text-based [17, 18] and multimodal solutions [19]. In particular, cross-modal models that match visual and textual data have been emerging as a viable strategy to build high-quality metrics [20, 21]. The CLIP model [13], which is pre-trained on web-collected data, has also been tested for image captioning evaluation, resulting in the CLIP-Score [21], which has a significant correlation with human judgment. However, using large-scale models pre-trained on web-collected data has limitations, as captions collected from alt-tags lack style and distribution, which is not aligned with those on which image captioning systems are evaluated. At the same time, recent advances in both image [22] and text generation [11, 23] have made it possible to synthetically generate data in both modalities, with controlled style and quality.

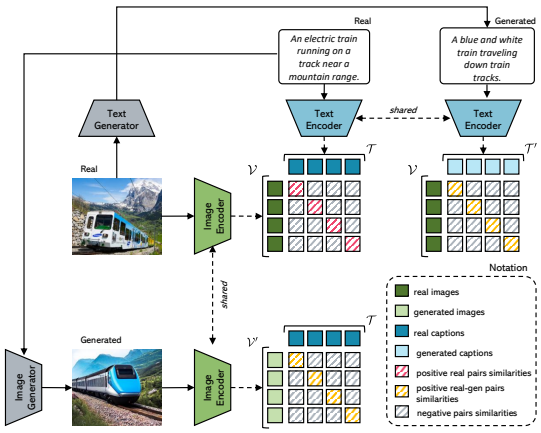

Figure 3: Overview of positive-augmented constrastive learning for captioning evaluation [2].

Following these insights, we recently proposed a learnable metric that fuses the advantages of both these scenarios, by leveraging the quality of the pre-training on web-collected data and that of cleaned data, and also regularizing the training by considering additional positive samples hailing from visual and textual generators [2]. Specifically, we started from the dual-encoder architecture popularized by CLIP [13], which comprises an image encoder [24, 25] and a text encoder [26]. In this architecture, the multimodal interaction is performed in a late fusion fashion, by projecting the output of both encoders to a common dimensionality and then on the ℓ_2 hypersphere via normalization. The visual and the textual inputs can then be compared via cosine similarity. Starting from a trained embedding space, an evaluation metric for image captioning can be defined by simply scaling, and eventually thresholding, the similarity computed inside of the embedding itself. To overcome the fact that the textual annotations in alt-tags and the distribution of web-scale images may not align well with the images used to evaluate captioning systems, we advocated the usage of synthetic generators of both visual and textual data, which showcase sufficiently high-quality levels when generating both images and text and are controllable in terms of visual distribution.

Formally, given a batch of N real images $\mathcal{V} = [v_1, v_2, ..., v_N]$ and their corresponding captions $\mathcal{T} = [t_1, t_2, ..., t_N]$, we augmented it by generating images $\mathcal{V}' = [v'_1, v'_2, ..., v'_N]$ (using the Stable Diffusion model [22]) and texts $\mathcal{T}' = [t'_1, t'_2, ..., t'_N]$ (employing the BLIP captioner [23]), and defined multiple $N \times N$ matrices containing pairwise cosine similarities between the different inputs. We then adopted a symmetric InfoNCE loss [27] which aims at maximizing the cosine similarity between the N matching pairs and minimize those of the $N^2 - N$ non-matching pairs (Figure 3). The loss which compares real images $\mathcal{V}$ with respect to real texts $\mathcal{T}$, for instance, can be defined as

$$L_{\mathcal{V},\mathcal{T}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(v_i,t_i)/\tau}{\sum_{j=1}^{N}\exp(\cos(v_i,t_j)/\tau} + \quad (1)$$

$$-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(v_i,t_i)/\tau}{\sum_{j=1}^{N}\exp(\cos(v_j,t_i)/\tau}, \quad (2)$$

where τ is a temperature parameter. In addition to a loss term between real images and real texts, we also add symmetrical loss terms between cross-modal generated and real pairs, *i.e.* between generated images and human-annotated texts ($L_{\mathcal{V}',\mathcal{T}}$), and between original images and generated texts ($L_{\mathcal{V},\mathcal{T}'}$). In this way, generated items act as additional positive samples for the real matching pairs, thus adding a supervisory signal without paying the cost of the noisy data on which contrastive-based features extractors like CLIP are learned. The final loss function is a weighted combination of the three loss terms.

We then used the positive-augmented CLIP model to evaluate captions for both image and video settings. Specifically, we followed the CLIP-S paradigm [21] to evaluate image-text pairs and the EMScore framework [28] to evaluate video-text ones, by employing our image-text model finetuned with additional positive samples. Experimentally, the proposed Positive-Augmented Contrastive learning Score (PAC-S) outperforms existing metrics on multiple datasets and achieves the highest correlation with human judgments on both images and videos. The source code and trained models are publicly available at https://github.com/aimagelab/pacscore.

3. Multimodal Embodied AI

3.1. Exploring and Explaining Embodied Environments

The development of embodied agents that can communicate with humans in natural language has gained increasing interest over the last years, as it facilitates the diffusion of robotic platforms in human-populated environments. As a step towards this objective, we tackled a setting for visual navigation in which an autonomous agent needs to explore and map an unseen indoor environment while portraying interesting scenes with natural language descriptions [29, 5]. Our approach can generate smart scene descriptions that maximize semantic knowledge of the environment and avoid repetitions. Further, such descriptions offer user-understandable insights into the robot's representation of the environment by highlighting the prominent objects and the correlation between them as encountered during the exploration.

Our architecture is composed of three main components: a *navigator*, in charge of the exploration, a *cap-*

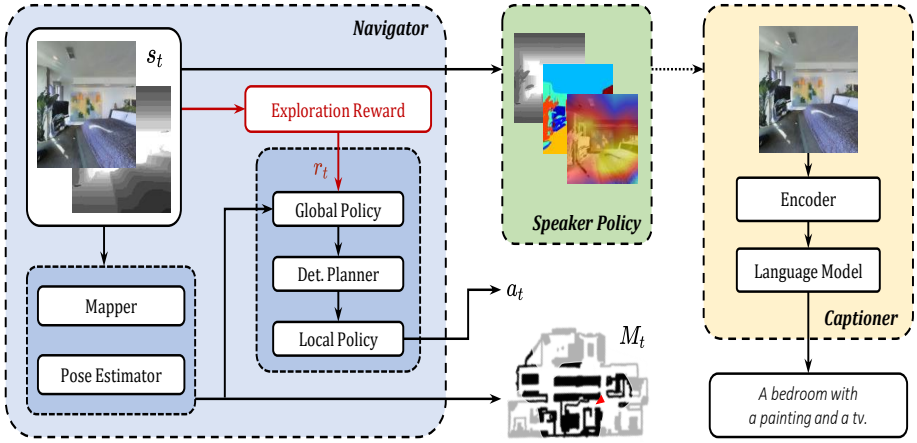

Figure 4: Overview of the approach for smart scene description, comprising a navigator, a speaker policy, and a captioner.

tioner, that describes interesting scenes, and the *speaker policy* that decides when the captioner should be activated. An overview of our complete architecture is shown in Figure 4.

Navigator. The exploration capabilities of the agent are strictly dependant on the performance of the navigation module, therefore relying on a proper navigation approach is of fundamental importance. Following recent literature on embodied visual navigation [30, 31], we devised a hierarchical policy coupled with a learned neural occupancy mapper and a pose estimator. The hierarchical policy sets long and short-term navigation goals, while the neural mapper builds an occupancy grid map representation of the environment and the pose estimator locates the agent on such map.

The output of the mapper is a global map of the environment that keeps track of the non-traversable space in its first channel and the area explored by the agent in the second one. At each time step, the mapper processes the RGB-D observation coming from the agent and predicts an egocentric local map representing the state in front of the agent. At every timestep, the local map is transformed using the estimated pose of the agent and registered to a global map with a moving average. The navigation policy adopts a hierarchical structure as used in [30, 31]. Specifically, the navigation policy comprehends three modules: a high-level global policy, a deterministic planner, and an atomic local policy. The hierarchical policy is adopted to decouple high-level and low-level concepts like moving across rooms and avoiding obstacles. It samples a goal coordinate on the map, while the deterministic planner uses the global goal to compute a local goal in close proximity of the agent. The local policy then predicts actions to reach the local goal.

As global exploration reward, we compared various approaches such as curiosity [32], coverage [33], anticipation [31], and impact [4]. All the considered methods obtain the reward by exploiting visual input sensors only. Exemplar exploration trajectories resulting from the different rewards are reported in Figure 5.

Captioner. 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. For multimodal fusion, we employed an encoder-decoder Transformer [26] architecture. Each layer of the encoder employs multi-head self-attention and feed-forward layers, while each layer of the decoder employs multi-head self- and cross-attention and feed-forward layers. For enabling text generation, sequence-to-sequence attention masks are employed in each self-attention layer of the decoder.

To obtain the set of visual features $\mathbf{V}$ for an image, our model employs a visual encoder that is pre-trained to match vision and language (*i.e.* CLIP [13]). Compared to using features extracted from object detectors [14, 11], our strategy is beneficial in terms of both computational efficiency and feature quality.

Speaker Policy. While exploring the environment, the agent sees various RGB observations. Even if the agent was navigating efficiently, the majority of the observations would be overlapped with each other, and the same objects would be observed at multiple consecutive timesteps. Since the agent should describe only relevant scenes during exploration and avoid uninformative captions or unnecessary repetitions, a component that controls caption generation becomes necessary. We thus introduce a speaker policy which is responsible for triggering the captioner depending on the current view. We compared three approaches that exploit different modali-

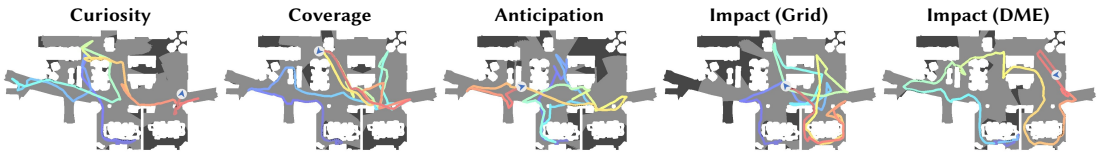

Figure 5: Qualitative exploration trajectories of different navigation agents on the same episode.

ties: a *depth-based* policy that employs mean depth value of the current observation, an *object-based* policy that uses the number of relevant objects in the RGB observation, and a *visual activation-based* policy that instead employs the activation maps of the visual encoder used by the captioner.

3.2. Generating Synthetic Instructions for VLN

In line with the trend observed in the past section, we have also focused on the Vision-and-Language Navigation (VLN) task for personalized embodied navigation. When performing VLN, an agent or a robot can perceive the 360° view of the environment and is given human instructions such as "*Walk forward, make a right turn around the kitchen island. Continue past the living area, enter the bedroom, then make a left turn and enter the bathroom*". The agent has to follow the instructions and navigate a previously unknown environment to reach the specified goal and stop there.

As collecting manual annotations has a significant cost, we also proposed a novel computational model that can generate synthetic instructions starting from unlabeled navigation paths in an environment. In the model that we propose, we combine a multimodal Generative Pre-Trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT) in an adversarial manner to generate better quality instructions. In particular, the model consists of a Transformer decoder (GPT-2) that generates sentences for a sequence of images from the environment describing the agent's path. The BERT-like encoder, instead, serves as a discriminator and is trained to distinguish between real and fake instructions. Both the generator (GPT-2) and discriminator (BERT) are trained simultaneously.

Acknowledgments

The research activities described in this work are partially supported by the Marie Sklodowska-Curie Action Horizon 2020 project "Personalized Robotics as Service Oriented Applications (PERSEO)", co-funded by the European Union, and by the PNRR and PRIN projects "Future Artificial Intelligence Research (FAIR)", FIT4MEDROB and "CREATIVE: CRoss-modal understanding and gEnerATIon of Visual and tExtual content", all co-funded by the Italian Ministry of University and Research.

References

[1] M. Cornia, L. Baraldi, G. Fiameni, R. Cucchiara, Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training, arXiv preprint arXiv:2111.12727 (2022).

[2] S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara, Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation, in: CVPR, 2023.

[3] M. Cornia, L. Baraldi, R. Cucchiara, SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability, in: ICRA, 2020.

[4] R. Bigazzi, F. Landi, S. Cascianelli, L. Baraldi, M. Cornia, R. Cucchiara, Focus on Impact: Indoor Exploration with Intrinsic Motivation 7 (2022) 2985–2992.

[5] R. Bigazzi, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara, Embodied Agents for Efficient Exploration and Smart Scene Description, in: ICRA, 2023.

[6] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, From Show to Tell: A Survey on Deep Learning-based Image Captioning, IEEE Trans. PAMI (2022).

[7] M. Cornia, L. Baraldi, G. Fiameni, R. Cucchiara, Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training, arXiv preprint arXiv:2111.12727 (2021).

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common Objects in Context, in: ECCV, 2014.

[9] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, in: ACL, 2018.

[10] S. Changpinyo, P. Sharma, N. Ding, R. Soricut, Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts, in: CVPR, 2021.

[11] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, VinVL: Revisiting visual representations in vision-language models, in: CVPR, 2021.

[12] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, Y. Cao, SimVLM: Simple Visual Language Model Pretraining with Weak Supervision, in: ICLR, 2022.
[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: ICML, 2021.
[14] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: CVPR, 2018.
[15] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: ACL, 2002.
[16] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: ACL Workshops, 2005.
[17] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-based Image Description Evaluation, in: CVPR, 2015.
[18] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic Propositional Image Caption Evaluation, in: ECCV, 2016.
[19] M. Jiang, Q. Huang, L. Zhang, X. Wang, P. Zhang, Z. Gan, J. Diesner, J. Gao, TIGEr: Text-to-Image Grounding for Image Caption Evaluation, in: EMNLP, 2019.
[20] H. Lee, S. Yoon, F. Dernoncourt, D. S. Kim, T. Bui, K. Jung, ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT, in: EMNLP Workshops, 2020.
[21] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, Y. Choi, CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in: EMNLP, 2021.
[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, 2022.
[23] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, in: ICML, 2022.
[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016.
[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: ICLR, 2021.
[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NeurIPS, 2017.
[27] A. v. d. Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, arXiv preprint arXiv:1807.03748 (2018).
[28] Y. Shi, X. Yang, H. Xu, C. Yuan, B. Li, W. Hu, Z.-J. Zha, EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching, in: CVPR, 2022.
[29] R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara, Explore and Explain: Self-supervised Navigation and Recounting, in: ICPR, 2020.
[30] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, R. Salakhutdinov, Learning To Explore Using Active Neural SLAM, in: ICLR, 2019.
[31] S. K. Ramakrishnan, Z. Al-Halah, K. Grauman, Occupancy Anticipation for Efficient Exploration and Navigation, in: ECCV, 2020.
[32] D. Pathak, P. Agrawal, A. A. Efros, T. Darrell, Curiosity-driven exploration by self-supervised prediction, in: ICML, 2017.
[33] D. S. Chaplot, D. P. Gandhi, A. Gupta, R. R. Salakhutdinov, Object Goal Navigation using Goal-Oriented Semantic Exploration, in: NeurIPS, 2020.