CLIP-Guided Vision-Language Pre-training for Question Answering in 3D Scenes

M. Parelli*, A. Delitzas*, N. Hars, G. Vlassis, S. Anagnostidis, G. Bachmann, T. Hofmann

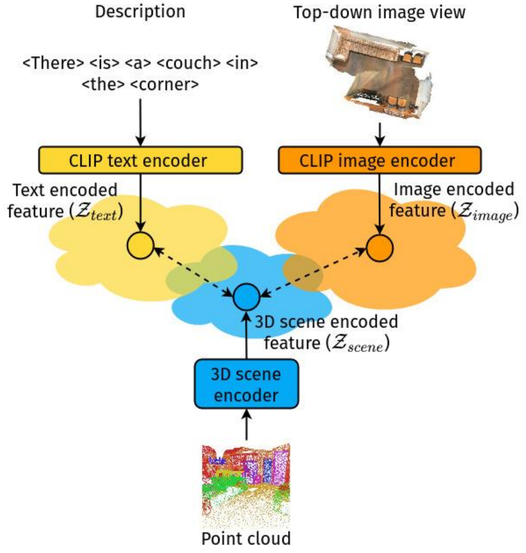
* indicates equal contribution

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Overview

We propose a novel *3D Vision-Language pre-training method* that helps a model learn *language-grounded and semantically meaningful* 3D scene point cloud representations for *question answering tasks* in 3D scenes.

Pre-training method



We align the scene embedding to the corresponding text and image representations extracted by the CLIP model via a cosine similarity loss:

$$\mathcal{L} = \mathcal{L}_{det} + \alpha\mathcal{L}_{text} + \beta\mathcal{L}_{image}$$

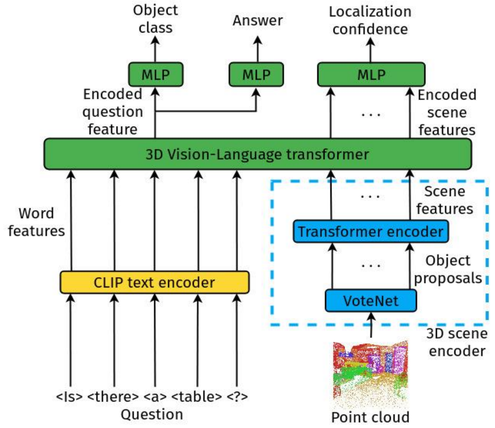
where

$$\mathcal{L}_{text} = 1 - \cos\left(Z_{text}, Z_{scene}\right)$$

$$\mathcal{L}_{image} = 1 - \cos\left(Z_{image}, Z_{scene}\right)$$

3D Visual Question Answering

We transfer the weights pre-trained by our method to the downstream task of 3D-VQA.



Quantitative results

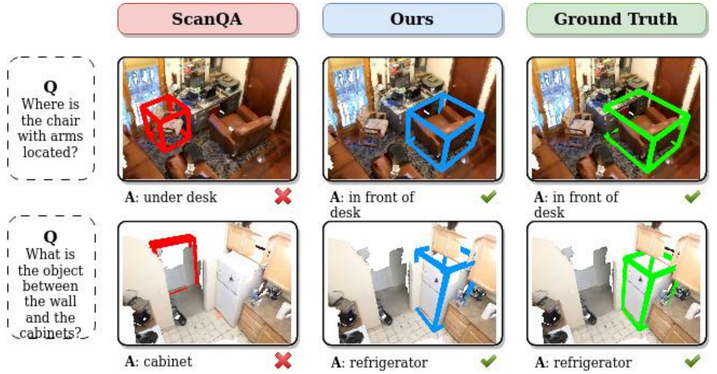
Question Answering on ScanQA test sets

Method	EM@1	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Test set w/ objects						
Scanrefer + MCAN	20.56	27.85	7.46	30.68	11.97	57.36
ScanQA w/o multiview	22.49	30.82	9.66	33.37	13.17	64.55
ScanQA	23.45	31.56	12.04	34.34	13.55	67.29
Ours w/o pre-training	22.76	31.08	13.31	33.84	13.28	65.81
Ours	**23.92**	**32.72**	**14.64**	**35.15**	**13.94**	**69.53**
Test set w/o objects						
Scanrefer + MCAN	19.04	26.98	7.82	28.61	11.38	53.41
ScanQA w/o multiview	20.05	30.84	12.80	30.60	12.66	59.95
ScanQA	20.90	30.68	10.75	31.09	12.59	60.24
Ours w/o pre-training	20.71	31.22	11.49	31.35	12.80	60.75
Ours	**21.37**	**32.70**	**11.73**	**32.41**	**13.28**	**62.83**

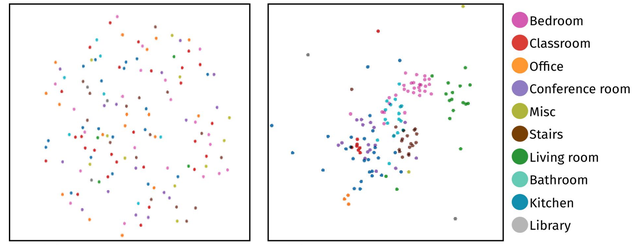
Referred Object Localization on ScanQA valid set

Method	Acc@0.25	Acc@0.5
Scanrefer + MCAN	23.53	11.76
ScanQA w/o multiview	25.17	16.21
ScanQA	24.96	15.42
Ours w/o pre-training	26.57	18.58
Ours	**29.61**	**21.22**

Qualitative results



Qualitative results on ScanQA valid set



(a) Random initialization (b) Our pre-training method

T-SNE visualizations of scene-level features in ScanNet