Entity-aware cross-modal pretraining for Knowledge-Based Visual Question Answering

Omar Adjali, Olivier Ferret, Sahar Ghannay, Hervé Le Borgne

Entity-Aware Cross-Modal Pretraining for Knowledge-based Visual Question Answering

Omar Adjali[1 [0002−6021−7776]], Olivier Ferret[1 [0003−0755−2361]], Sahar Ghannay[2 [0002−7531−2522]], and Hervé Le Borgne[1 [0003−0520−8436]]

[1] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[omar.adjali,olivier.ferret,herve.le-borgne]@cea.fr
[2] Université Paris-Saclay, CNRS, LISN, Orsay, France
sahar.ghannay@lisn.upsaclay.fr

Abstract. Knowledge-Aware Visual Question Answering about Entities (KVQAE) is a recent multimodal retrieval task aiming to answer visual questions about named entities from a multimodal knowledge base. In this context, we focus more particularly on cross-modal retrieval and propose to inject information about entities in the representations of both texts and images during their building through two pretraining auxiliary tasks, namely entity-level masked language modeling and entity type prediction objectives. We show competitive performance over existing approaches on 3 KVQAE standard benchmarks, revealing the interest of raising entity awareness during cross-modal pretraining and specifically for the KVQAE task[3].

1 Introduction

Contrastive Learning (CL) has emerged as a powerful paradigm for improving the representation learning capabilities of Vision-Language Pretrained Models (VLPMs) by aligning Vision-Language representations in a common latent vector space, which often spills over into the performances of numerous multimodal applications, including image captioning, visual question answering, and cross-modal retrieval. CLIP [22] is a notable example that demonstrated great zero-shot performance in several cross-modal tasks. However, several works pointed out the limitation of CLIP when adapting to downstream tasks, in particular, due to its unique instance-level contrastive pretraining [26,18]. In this context, we investigate in this paper how pretrained CLIP can handle entity-level information and its potential benefits on the multimodal KVQAE retrieval task. Indeed, the KVQAE task involves unstructured text rich with general domain entity-level information (see Figure 1). We argue that leveraging such knowledge can help capturing additional fine-grained cross-modal interactions and retrieve more semantically relevant text-image pairs that boost the performance of multimodal retrieval. Our contributions are the following. (1) We formulate a contrastive objective for disentangling passage representations of distinct entities. (2) We extend the standard

[3] This manuscript is a preprint submitted to a conference which allow the release of the Accepted Manuscript (i) on the authors personal websites *e.g* here (ii) on the employer and/or funder repositories twelve (12) months after first publication

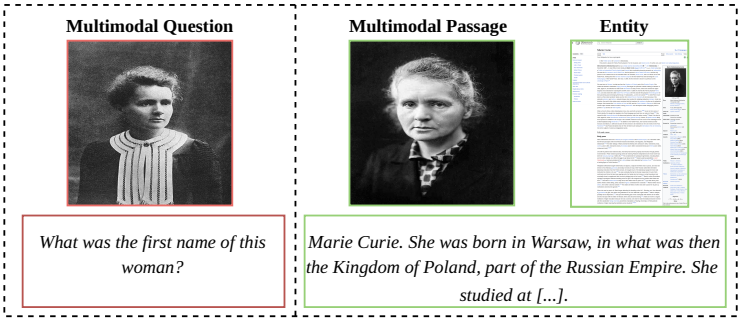

Fig. 1: KVQAE illustrative example.

CLIP pretraining framework with two mask-based self-supervised learning objectives, designed to implicitly incorporate entity information into CLIP representations. (3) Our experiments on 3 datasets demonstrate that our CLIP entity-aware pretraining improves performance for the KVQAE task.

Previous works aimed to improve the performance of pretrained CLIP models on downstream tasks without degrading their zero-shot performance after standard fine-tuning. [17] leverage new external knowledge by training new attention-based blocks introduced between the original layers of the CLIP vision encoder while freezing all the original weights to avoid catastrophic forgetting. Alternatively, [27] rely on a fine-grained contrastive loss that considers the interactions between image patches and textual tokens. [23] proposed a prompting-based approach that includes special prefix tokens during training to condition the language encoder on the input data type. Similarly, [28] augment CLIP with learned prompt vectors for image recognition tasks. In [26], besides combining auxiliary self-learning and CLIP plain objectives, they employ an image token removal data augmentation strategy that achieves significant performance gains. More in line with our work, [19] propose EI-CLIP, a CLIP model adapted to the e-commerce domain that handles the special meaning of entities. However, to the best of our knowledge, no existing approach consider the integration of general-domain entity-level information in PVLMs.

2 Method

2.1 Entity-aware cross-modal pretraining

Our approach leverages entity information during cross-modal passage retrieval by raising CLIP pretraining awareness to entity-level information. First, we explicitly formulate a contrastive objective for disentangling passage representations of distinct entities. Second, we extend the standard CLIP pretraining framework with two mask-based self-supervised learning objectives that implicitly inject entity information into CLIP representations while preserving text-image alignments with the standard contrastive loss (see Figure 3).

Entity-centric contrastive learning Given a collection of text-image pairs representing passages and their visual content associated with entities in a Knowledge Base (KB), we build a batch of randomly sampled passage-image pairs such that each pair is associated with a distinct entity. This enables training a CLIP model such that passages from the same entity are brought closer in the embedding space, while passages from different entities are pushed apart. Formally, given a batch of N text-image pairs, texts are tokenized into wordpiece sequences and images are transformed into vectors of 2D patches. The textual and visual encoders $\mathbf{E^T}(\cdot)$, $\mathbf{E^V}(\cdot)$ encode these inputs into global text and image embeddings (e^t, e^v). Finally, a CLIP dual encoder is trained to minimize a contrastive loss that align the positive text-image pairs and keep the negative pairs apart, leading to a joint embedding space of text and image representations.

Self-supervised entity-level masked language modeling Similar to the standard MLM objective used for pretraining LMs such as BERT [7], we pretrain the CLIP text encoder on Entity-Level Masked Language Modeling (EL-MLM). We perform entity-level masking using the special token `[MASK]` on each token from the identified entity span. We introduce an EL-MLM Head $g_{\text{MLM}}(\cdot)$ similar to $g_{\text{type}}(\cdot)$ that maps the hidden states of the input masked text to the token vocabulary space. Its weights are tied to the token embedding layer. The CLIP text encoder is trained with the following EL-MLM loss:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{|V|}(y_{i,k}\cdot\log(l_{i,k})) \tag{1}$$

where $|V|$: CLIP text encoder vocabulary size; $y_{i,k}$: index of the original tokens before masking; $l \in \mathbb{R}^{N\times|V|}$: logits after projection with $g_{\text{MLM}}(\cdot)$.

Mask entity type prediction Several works [4,8] have demonstrated that capturing latent entity-type information into embeddings may enhance their representation quality followed by improvement on downstream tasks. Intuitively, we suggest that entity-type information may help answer visual questions about entities. Thus, during pretraining, given an input passage P_i, tokenized into n textual tokens and from which we identified k entity spans (entity mentions), we replace each token t_i that belongs to an entity span s_j with the special token `[MASK]` we added to the CLIP text token vocabulary. The hidden states of the masked sequence $h^{t_{mask}}$ are generated using the CLIP text encoder $\mathbf{E^T}(\cdot)$. Let $\mathbb{V}_{type}$ be the vocabulary set of entity types, we design a mask prediction head $\mathbf{g}_{type}(\cdot)$ consisting of two linear layers with Gelu activation followed by a softmax layer that projects $h^{t_{mask}}$ to the entity type distribution space which yields the logits $l \in \mathbb{R}^{N\times|V_{type}|}$. We train CLIP text encoder to predict the entity type of each masked token given its context using the cross-entropy loss:

$$\mathcal{L}_{\text{type}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{|V_{type}|}(y_{i,k}\cdot\log(l_{i,k})) \tag{2}$$

where N: size of the mini-batch; $|V_{type}|$: entity type set size; $y_{i,k}$: a label binary indicator of the k-th entity type class of i-th token. This loss encourages the model to capture the specific semantics of entity types by updating directly the text encoder weights. The final loss is $\mathcal{L}_{\text{combined}} = \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{type}} + \mathcal{L}_{\text{MLM}}$.

2.2 KVQAE task

To extract answers about named entities, we follow a two-step pipeline involving Information Retrieval (IR) followed by Reading Comprehension (RC). During the IR step, given a text-image pair (Q_T, Q_I) associated with a question, we retrieve the list of the k most relevant text-image pairs $\{(P_T, P_I)_1 \cdots (P_T, P_I)_k\}$ from a KB $\mathcal{KB} = \{(P_T, P_I)\}$ with respect to the query question. The KB includes passages in unstructured text, each associated with visual content. Consequently, the RC step aims at extracting the correct answers from the passages retrieved during the IR step. This extraction is performed using multi-passage BERT [24] on the top-K relevant passages.

Passage retrieval KVQAE benchmarks come with an external unstructured KB built from Wikipedia. Following [16], articles (entities) are split into passages of 100 words and each passage is headed with its corresponding Wikipedia article title. Given text and vision encoders, we map question and passage text-image pairs to dense vector representations and perform a cosine similarity-based dense search to retrieve the top-100 most relevant passages. Similarly to previous works [16,1,14], we assume that a passage is relevant if its text contains the answer.

Similarity-level fusion To efficiently address the KVQAE task, cross-modal retrieval needs to be combined with textual and visual mono-modal retrievals to consider all query-candidate multimodal interactions. Specifically, we use the state-of-the-art Dense Passage Retrieval (DPR) [13] dual encoder for mono-modal text retrieval (MM_T), and the pretrained EA-CLIP models for mono-modal image (MM_I) and cross-modal retrieval ($\text{CM}_{I\rightarrow T}$). Passages are ranked according to the final similarity scores S obtained after a late fusion that combines the retrieval results using linear interpolation: $S = \alpha_0 \cdot S_{MM_T} + \alpha_1 \cdot S_{MM_I} + \alpha_2 \cdot S_{CM_{I\rightarrow T}}$. Search scores are normalized to zero mean and unit variance to have comparable distributions. The interpolation hyperparameters α_i are determined through a grid search conducted on the validation set of each dataset to maximize the mean reciprocal rank at 100.

3 Experimental framework

Experiments are conducted on three recent KVQAE benchmarks, namely ViQuAE [16], EVQA [20], and InfoSeek [5].

Baselines. We compare this work with several state-of-the-art baselines. Late fusion approaches LF [16] and LF_{gcn} [1] combine text and image mono-modal retrieval scores using features from DPR [13], ArcFace [6], CLIP, and ImageNet-ResNet [9]; Early fusion baselines ECA (Early Cross-Attention) and ILF (Intermediate Linear Fusion) [14], where both perform retrieval using multimodal dense representations. Similar to our work, [15] employed different retrieval and training strategies for finetuning CLIP. We refer to their approach as CLIP_{FT} and to ours as EA-CLIP. Note that we do not include augmented-retrieval and large generative model-based approaches with billions of parameters in our baselines such as in [20,5,11,10].

Method	ViQuAE				InfoSeek			EVQA (1-hop)			
	MRR	P@1	EM	F1	MRR	P@1	SoftMatch	MRR	P@1	BEM	F1
BM25	19.0	13.1	–	–	4.2	2.4	–	**21.1**	**15.6**	–	–
DPR	32.8	22.8	–	–	8.4	5.2	–	19.7	13.2	–	–
LF [16]	37.9	27.8	22.1	25.4	–	–	–	–	–	–	–
LF_{gcn} [1]	**38.3**	**29.0**	–	–	–	–	–	–	–	–	–
ILF [14]	37.3	26.8	21.3	25.4	–	–	–	–	–	–	–
ECA [14]	37.8	26.7	20.6	24.4	–	–	–	–	–	–	–
$CLIP_{FT}$ [15]	*37.6*	*28.6*	30.9	34.3	–	–	(12.4)	–	–	29.1	26.6
DPR+$CLIP_{zs}$	33.2	23.2	28.5	32.2	8.7	5.3	4.8	20.1	13.6	29.5	26.8
DPR+EA-CLIP	37.3	27.6	28.5	32.2	**9.0**	**5.4**	5.2	**20.5**	**13.8**	30.2	27.7

Table 1: Passage retrieval (MRR, P@1) and reading comprehension (EM, F1, SoftMatch, BEM) results in the multimodal fusion setting.

Evaluation metrics. We evaluate the retrieval performance using Precision@1 (P@1) and Mean Reciprocal Rank at 100 (MRR) metrics. Passages/articles are considered relevant using text matching with the ground truth answers. Answer extraction is evaluated using F1-score and Exact Match metrics for the ViQuAE dataset, BERT Matching metric (BEM) [3] for InfoSeek and Soft matching score for EVQA.

Experiment Settings. We first pretrain a CLIP model following our approach on 12M ViQuAE KB text-image passage pairs annotated with entity-level information [1]. Note that we did not consider the text-to-image matching loss term $\mathcal{L}_{t \to i}$ when computing the contrastive loss. Indeed, our experiments showed poor performance on text-to-image retrieval. We believe that CLIP struggles in this setting because textual questions are quite different from its original caption-based pretrained data. We then continue finetuning on each downstream KVQAE dataset. We used the released pretrained CLIP model with the ViT-B/32 visual encoder and its corresponding transformer-based text encoder. We consider 22k entity types obtained using the Wikidata *is_instance* relation.

4 Results

Passage retrieval. We report in Table 1 (cols MRR and P@1) the performance of our approach in the late fusion setting against state-of-the-art (SOTA) baselines. Overall, our approach (EA) achieves comparable performance on the ViQuAE dataset compared to the best-performing ones [16,1]. Indeed combining mono-modal text (DPR), mono-modal image (CLIP), and cross-modal retrieval (CLIP) is sufficient to reach the SOTA performance without the need for specialized features such as face embeddings. Results are confirmed on EVQA and InfoSeek while emphasizing the importance of finetuning CLIP. However, passage retrieval results must be put into perspective with respect to the KVQAE objective namely, extracting answers. Regarding how passages are considered relevant, we experimentally observed that the best-performing retrieval system will not necessarily yield the best answer extraction results, as the sole text-matching criterion

Meth.	Retr.	ViQuAE		InfoSeek		EVQA	
		MRR	P@1	MRR	P@1	MRR	P@1
CLIP$_{ZS}$	$i_q - i_e$	29.3	22.0	7.3	4.4	13.6	7.6
	i_q - t_e	32.7	22.0	5.5	2.5	14.5	7.7
	fusion	40.8	32.9	7.0	3.9	16.0	9.4
EA-CLIP	$i_q - i_e$	29.4	21.5	7.9	4.0	14.3	8.2
	$i_q - t_e$	39.0	29.0	6.6	3.0	15.2	8.7
	fusion	**44.4**	**35.5**	**8.1**	**4.5**	**16.8**	**10.1**

Table 2: Cross-modal entity retrieval evaluation on ViQuAE and EVQA test sets, and InfoSeek validation set. Results show mono-modal image retrieval ($i_q - i_e$), cross-modal retrieval (i_q - t_e), and their score-level fusion.

for passage retrieval relevance may generate noisy passages. An alternative IR approach followed by [15] is to perform entity-level retrieval first, then map the retrieved entities to the corresponding passages. Thus, we evaluate our approach on cross-modal entity retrieval for further assessment. Similarly to passage retrieval, given a visual question, retrieved entities are relevant if their corresponding Wikipedia document contains the ground truth answer. Table 2 shows that our approach yields competitive performance consistently across all datasets and metrics with significant gains in cross-modal retrieval (i_q - t_e).

Reading Comprehension. Table 1 (cols. EM, F1, BEM and SoftMatch) shows the RC results, where answers are extracted from the retrieved passages provided by the aforementioned IR systems. We can see that EA-CLIP contributes positively to the KVQAE task by consistently outperforming all baseline methods across all evaluation metrics on ViQuAE and EVQA. This illustrates the benefit of considering entity-level information to retrieve passages with more relevant answers. Specifically, the best performance gain is achieved on the ViQuAE dataset, with respectively +1.9 and +2.1 for EM and F1 scores. Several factors impact the level of improvement across datasets. Obviously, the size of the KB matters since IR and RC performances are affected by the number of distractors. Moreover, despite the entity overlap between EVQA and ViQuAE KBs (both built from Wikipedia), our EA-CLIP pretraining approach was performed on the ViQuAE KB and passages, which favors the performance on the ViQuAE dataset. One can also observe that CLIP$_{zs}$ slightly outperforms CLIP$_{FT}$ on EVQA, showcasing the potential drawbacks of CLIP finetuning. While EA-CLIP achieves a Soft Match score of 5.2 on InfoSeek, CLIP$_{FT}$ performs better (12.4). However, these results are not directly comparable due to differences in the KBs size used for evaluation. This highlights even more the challenging nature of the KVQAE task with very large KBs.

5 Conclusion

The proposed entity-aware CL framework leverages the capabilities of CLIP for cross-modal retrieval. By explicitly considering entity-level information in textual data, CLIP

learns to align texts and images while considering entity information through auxiliary prediction tasks. Our approach achieves SOTA results on the KVQAE task, demonstrating the benefit of raising entity-information awareness for training cross-modal models.

Acknowledgment This work was supported by by the ANR-19-CE23-0028 ANR MEERQAT project and the ANR-23-PEIA-0008 ANR SHARP project, supported by France 2030. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013719 made by GENCI. It also relied on the use of the FactoryIA cluster, financially supported by the Ile-de-France Regional Council.

References

1. Adjali, O., Grimal, P., Ferret, O., Ghannay, S., Le Borgne, H.: Explicit knowledge integration for knowledge-aware visual question answering about named entities. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. pp. 29–38 (2023)
2. Bassani, E., Romelli, L.: ranx.fuse: A python library for metasearch. In: CIKM. pp. 4808–4812. ACM (2022). https://doi.org/10.1145/3511808.3557207
3. Bulian, J., Buck, C., Gajewski, W., Boerschinger, B., Schuster, T.: Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. arXiv preprint arXiv:2202.07654 (2022)
4. Chen, S., Wang, J., Jiang, F., Lin, C.Y.: Improving entity linking by modeling latent entity type information. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 7529–7537 (2020)
5. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can pre-trained vision and language models answer visual information-seeking questions? arXiv preprint arXiv:2302.11713 (2023)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR June 16-20. pp. 4690–4699. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019). https://doi.org/10.1109/CVPR.2019.00482
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
8. Dong, X., Yu, W., Zhu, C., Jiang, M.: Injecting entity types into entity-guided text generation. arXiv:2009.13401 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, June 27-30. pp. 770–778. IEEE Computer Society, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.90, https://doi.org/10.1109/CVPR.2016.90
10. Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N.A., Luo, J.: Promptcap: Prompt-guided task-aware image captioning. arXiv preprint arXiv:2211.09699 (2022)
11. Hu, Z., Iscen, A., Sun, C., Chang, K.W., Sun, Y., Ross, D.A., Schmid, C., Fathi, A.: Avis: Autonomous visual information seeking with large language model agent. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
12. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data **7**(3), 535–547 (2019)

13. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.550, https://aclanthology.org/2020.emnlp-main.550
14. Lerner, P., Ferret, O., Guinaudeau, C.: Multimodal inverse cloze task for knowledge-based visual question answering. In: European Conference on Information Retrieval. pp. 569–587. Springer (2023)
15. Lerner, P., Ferret, O., Guinaudeau, C.: Cross-modal retrieval for knowledge-based visual question answering (2024)
16. Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besançon, R., Moreno, J.G., Lovón Melgarejo, J.: Viquae, a dataset for knowledge-based visual question answering about named entities. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3108–3120. Association for Computing Machinery, Madrid, Spain (2022)
17. Liu, H., Son, K., Yang, J., Liu, C., Gao, J., Lee, Y.J., Li, C.: Learning customized visual models with retrieval-augmented knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15148–15158 (2023)
18. Luo, Z., Xi, Y., Zhang, R., Li, G., Zhao, Z., Ma, J.: Conditioned masked language and image modeling for image-text dense retrieval. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 130–140 (2022)
19. Ma, H., Zhao, H., Lin, Z., Kale, A., Wang, Z., Yu, T., Gu, J., Choudhary, S., Xie, X.: Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18051–18061 (2022)
20. Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., Ferrari, V.: Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. arXiv preprint arXiv:2306.09224 (2023)
21. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS 2017 Workshop on Autodiff. MIT Press, Long Beach, CA, USA (2017)
22. 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: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR, Virtual Event (2021), http://proceedings.mlr.press/v139/radford21a.html
23. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Prefix conditioning unifies language and label supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2861–2870 (2023)
24. Wang, Z., Ng, P., Ma, X., Nallapati, R., Xiang, B.: Multi-passage bert: A globally normalized bert model for open-domain question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5878–5882 (2019)
25. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association

for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
26. Yang, Y., Huang, W., Wei, Y., Peng, H., Jiang, X., Jiang, H., Wei, F., Wang, Y., Hu, H., Qiu, L., et al.: Attentive mask clip. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2771–2781 (2023)
27. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
28. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)

A Entity-centric Contrastive Learning Details

$\mathbf{E^T}(\cdot)$, $\mathbf{E^V}(\cdot)$ encode tokenized inputs yielding respectively the hidden state representations $h^t \in \mathbb{R}^{N\times L\times D}$ and $h^v \in \mathbb{R}^{N\times L\times D}$, where *D* is the embedding dimension and *L* is the max token sequence length. To perform instance level matching, the hidden state of the special "Beginning of sequence" token `[BOS]` in the final layer $h^t_{\texttt{[BOS]}} \in \mathbb{R}^{N\times D}$ and $h^v_{\texttt{[BOS]}} \in \mathbb{R}^{N\times D}$ are used for dense vector representation. $p^t_{\texttt{[BOS]}} = \mathbf{g}(h^t_{\texttt{[BOS]}})$ and $p^v_{\texttt{[BOS]}} = \mathbf{g}(h^v_{\texttt{[BOS]}})$ are respectively the `[BOS]` text and visual features mapped to a common multimodal space using the projection head $\mathbf{g}(\cdot)$. The final global text and image embeddings (e^t, e^v) are the $L2$-normalization of $(p^t_{\texttt{[BOS]}}, p^v_{\texttt{[BOS]}})$. We train CLIP on minimizing the following contrastive loss:

$$\mathcal{L}_{i\to t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(e^t_i \cdot e^v_i/\tau)}{\sum_{j=1}^{N}\exp(e^t_i \cdot e^v_j/\tau)} \tag{3}$$

where τ is the temperature hyperparameter controlling the level of penalties on hard negative pairs.

B Details about the experimental framework

In Table 3, we report the main features of the three datasets used for evaluation.

B.1 Evaluation metrics

Concerning the EVQA dataset, when specified, the evaluation is done on the 1-hop questions, which adds up the templated and automatic subsets.

B.2 Hyperparameters and training settings

The 12M ViQuAE KB text-image passage pairs used for training our models are available on the Hugging Face data hub: https://huggingface.co/datasets/usr256864/viquae_passage_linked_entities.

Table 3: KVQAE benchmark datasets and KB statistics.

KB-VQA Benchmarks	Dataset			Knowledge Base	
	Train	Val	Test	#Entities	#Passages
ViQuAE [16]	1,190	1257	1,250	1,495,352	11,885,968
EVQA [20]	212,338	2,950	5,750	2,004,561	31,341,692
InfoSeek [5]	934,048	73,620	347,980	6,084,491	42,529,637

The initial learning rate is set to 2e-6 and the total number of finetuning epochs is set to 20. For the sake of fair comparison, all CLIP models were trained with a batch size of 1,000 using gradient check-pointing. Models selection is done based on the in-batch mean reciprocal rank on the validation set. Our implementation relies on PyTorch [21], Transformers [25], Faiss library [12], and ranx [2]. Our code will be made publicly available after the anonymity period.

C Details about the results

C.1 RC results

The number of retrieved passages, K, that feed the RC stage is set to 24 (ViQuAE) or 100 (InfoSeek and EVQA) according to a grid-search on the validation set.

C.2 Illustration of the difference between the original CLIP (ZS-CLIP) and our entity-aware CLIP (EA-CLIP)

A qualitative analysis of the embedding space shows that EA-CLIP text encoder produces more qualitative passage representations compared to ZS-CLIP, where passages belonging to the same entity are better clustered with less overall overlapping than ZS-CLIP embeddings (see Figure 2). This highlights the potential benefit of considering entity information for cross-modal retrieval.

D Pretraining Architecture Figure

Figure 3 illustrates the pretraining stages in the proposed approach.

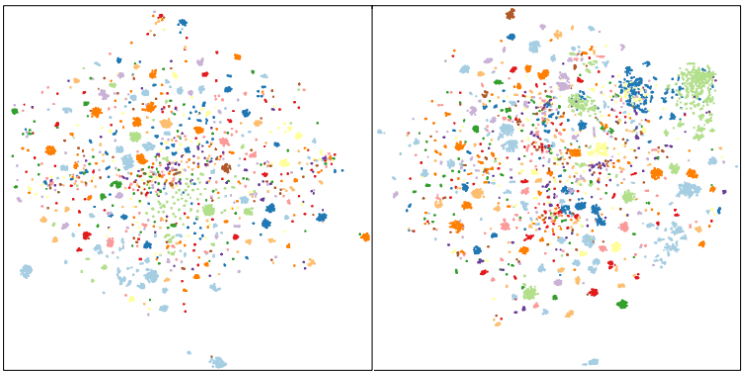

Fig. 2: t-SNE visualization of 25,000 randomly selected passages covering 1,000 entities (articles) in the ViQuAE KB. Passage embeddings are obtained with text encoders of our EA-CLIP (left) and zero-shot CLIP (right). Colors represent entity labels.

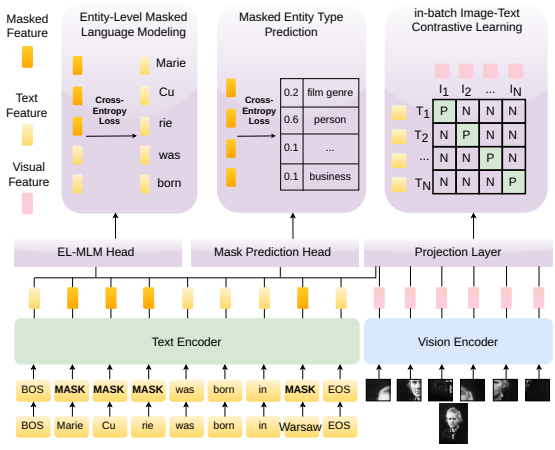

Fig. 3: Entity-aware cross-modal pretraining.