Supplementary Material

Anonymous Author(s)
Affiliation
Address
`email`

A More details of Multimodal Pipeline

A.1 Image Encoder

In this work, we focus on patch features and apply vision transformer based models(ViTs) by [15] as our visual encoder backbones. We split input image into a sequence of patches and adopt the linear projection embedding of patch features v_q, which simplifies the step for fusing with text embedding. Following the vision-language models, we train our model with multiple popular ViTs to examine the influence of image encoder in OOD detection backpropagation process.

A.2 Knowledge Retrieval Module

Given the caption S, we parse it into triplets in the form of $T^{id} = \langle o(c), r(c), o'(c) \rangle$, where $o(c)$ and $o'(c)$ are concepts $\in C^{id}$,and $r(c)$ is the relation(s) between them, i.e., $\langle$man, riding, bicycle$\rangle$. In our example, the seed triplets(ID triplets) parsed from the caption are $\langle$man, riding, bicycle$\rangle$ and $\langle$bicycle, down, street$\rangle$. Then we construct knowledge graph by bridging these triplets with external open knowledge including domain and commonsense knowledge graphs, e.g., ConceptNet [13]. ConceptNet provides a large scale commonsense knowledge with over 21 million edges by 36 type of relations connecting 8 million nodes, i.e., IsA, UsedFor, AtLocation. In this study, to complete our knowledge graph, we collect concepts by querying from ConceptNet using $o(c)$, $o'(c)$ and rel_i where $i \in [0, 36]$ and integrate extracted triplets to seed triplets. For example, given "street" as $o(c)$ and "AtLocation" as rel_i, we will extract the related concepts are located at street to form triple t_i. Specifically, we query explicit knowledge triplets of $o(c)$ and $o'(c)$ from ConceptNet to form T^{cn}, i.e., $\langle$bicycle, used for, transport$\rangle$. Finally, these knowledge triplets $\in T = T^{cn} \cup T^{id}$ are encoded as language features, such as l_j using a language encoder(e.g., BERT[3]).

A.3 Multimodal Fusion Encoder

We consider our proposed OOD detection layer as a plug module to different vision-language architectures. In ViLT and CLIP, we sum the visual and textual embeddings with incorporating $s(\cdot)$ as in Eq 4 of the main paper, and pass them to a standard L-depth transformer. Considering the BLIP backbones, we perform a two-stream transformer pipeline consisting of stacked multiple layers to joint vision and knowledge textual representations. For each layer, we have self-attention unit and merged cross-attention unit which integrates vision and knowledge semantic information and the alignments across them, and a position-wise feed-forward network.

Moreover, we update the image and language embedding outputs of themselves previous layer as queries and concatenate them together as keys and values. To further improve the performance of attention function, we use a multi-head attention which is composed by multiple paralleled attention function in each head. The feed-forward layer transform the outputs of multi-head attention through two fully-connected layers with GeLU activation.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

B More details of Training objectives

We mainly introduce our training objectives in our pipeline in this section, including image text matching (ITM) and masked language modeling (MLM).

B.1 Image Text Matching

To incorporate both the vision and the language representations, we adopt ITM which is widely used in previous VL studies. Given an image and text of triple pair $\langle v_q, l_j \rangle$, ITM predicts whether they are matched as positive examples or not, and it is a binary classification problem with the loss function in Equation 5 of the main paper. We assume that each image and ID triple pair $\langle v_q, l_j \rangle$, as a positive example. The negative pairs are constructed through batch-sampling.

B.2 Masked Language Modeling

MLM utilizes vision features and text features of ID concepts and relations to predict the masked tokens in the caption sentence S. Following most VL models, we randomly masked some tokens in S replacing as y^{msk} and predict them with their visual and textual features. Since some tokens are replaced with “[mask]", the OOD score $s(\cdot)$ is changed based on the random masks. Thus, we incorporate $s(\cdot)$ to calculate the predicted probability for a masked token similar to Eq 4 of the main paper, and MLM loss is written as

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{(v,\hat{l})\sim D}\mathcal{H}(y_{mlm}, p(v, \hat{l})) \tag{1}$$

where $\mathcal{H}$ denotes the cross-entropy, y_{mlm} is a one-hot vector where the ground truth tokens are with probabilities of 1, $\hat{l}$ denotes the masked text.

C More Experiments

In this section, we show implement details, more experiments on ablation studies, and more qualitative analysis of our proposed VK-OOD models.

C.1 Implement details

Datasets. Following the practical settings, we adopt the training strategies of pre-training on more data and fine-tuning on downstream tasks. We pre-train on three datasets, including COCO [5], Visual Genome [4], and SBU Captions [9] with total of 1M images and 6.8M image-caption pairs, as approximate 30% less than the baseline(ViLT). Each caption is parsed to 1 - 3 triplets and augmented with 5 external knowledge triplets. For downstream datasts, we use Flickr30k [10] and COCO for image-text retrieval, VQAv2 [1] and OKVQA [8] for visual question answering and ablation studies, and NLVR2 [14] for visual reasoning. We resize each image to the size of 224×224 by center-cropping. In the merged attention module, each multimodal encoder layer consists of one multi-head self-attention block and one feedforward block, and total number of identical layers is 12.

Encoder backbones. First, we retrieve explicit knowledge triplets in pre-processing, by using ConceptNet Numberbatch[1]. Next, for ViLT, we use RoBERTa [6] as text encoder and ViT-B/32 by [11] as visual encoder. To scale with CLIP, we use CLIP-ViT-B/32 [11] as both backbones. Then, we follow the BLIP design with BERT-base as text encoder and ViT-B/16 as visual encoder.

Network training. We pre-train the model for 10 epochs, and use AdamW optimizer designed by [7] with the learning rate of 1e-4 and weight decay of 1e-2. The warm-up ratio of learning rate is 10% of the total training steps, and the learning rate was decayed linearly to 0 in the rest steps. Then, we finetune our model for 5 epochs with learning rate of 2e-4 for all downstream tasks. In addition, we apply RandAugment [2] as augmentation strategy in finetuning steps. We pre-train and fine-tune both on 8 NVIDIA RTX 2080Ti GPUs, and inference on 1 NVIDIA RTX 2080Ti GPU.

---

[1] `https://github.com/commonsense/conceptnet-numberbatch`

Model	COCO						F30k					
	TR			IR			TR			IR		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViLT	61.8	86.2	92.6	41.3	72.0	82.5	81.4	95.6	97.6	61.9	86.8	92.8
UNITER	64.4	87.4	93.1	50.3	78.5	87.2	85.9	97.1	98.8	72.5	92.4	96.1
ALBEF	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	98.4
VinVL	74.6	92.6	96.3	58.1	83.2	90.1	-	-	-	-	-	-
BLIP	80.6	**95.2**	**97.6**	**63.1**	**85.3**	91.1	**96.6**	**99.8**	**100**	**87.2**	**97.5**	**98.8**
Ours(ViLT)	73.8	91.4	96	52.4	81.3	90.1	85.9	97.1	97.6	80.1	94.6	96.7
Ours(CLIP)	69.8	87.5	93.6	48.8	78.5	82.5	92.3	98.4	99.5	79.8	92.1	96.4
Ours(BLIP)	**80.7**	95.1	96.8	62.9	84.8	**92.8**	96.4	99.6	99.8	86.3	97.1	**98.8**

Table 1: Detailed results of image-text retrieval tasks on COCO and Flickr30k datasets. Our model with different backbones outperforms other models and achieve the best and second-best results.

Model	Objectives	VQA	Flickr30k	
		test-dev	TR@1	IR@1
ViLT	ITM	70.6	82.1	65.6
ViLT	MLM	72.8	-	-
ViLT	ITM+MLM	74.2	88.1	74.1
VK-OOD	ITM	72.1	84.5	69.8
VK-OOD	MLM	73.4	-	-
VK-OOD	ITM+MLM	**74.8**	**89.0**	**77.2**

Table 2: Ablation study experiment results of VK-OOD model. ViLT is our implementation without explicit knowledge and OOD detection layer. ITM is image-text matching, and MLM is masked language modeling. Results on VQA are on test-dev set. Both downstream results are in zero-shot settings. The bold values mean the best model in the table. Comparing with the baselines, our model with OOD detection layer outperforms on all objectives with two datasets. Training on combinations of objectives improves model performance.

C.2 More experimental results

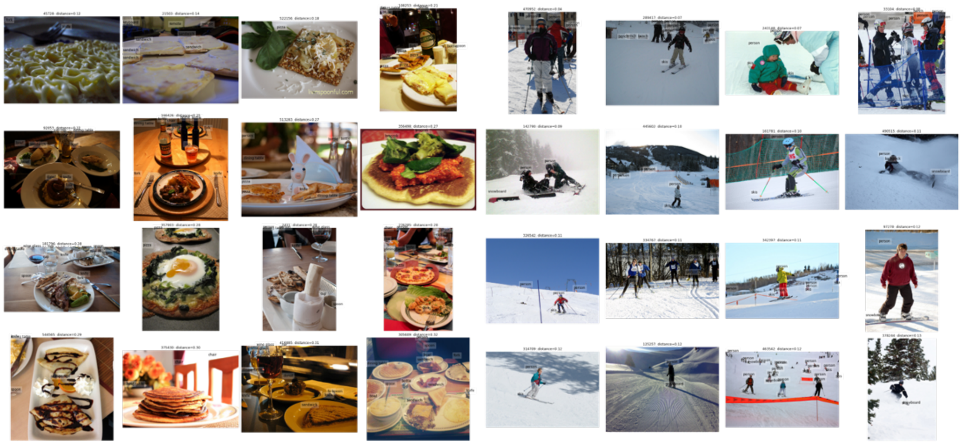
Detailed results of image-text retrieval tasks. We provided detailed results on COCO and F30K datasets, as shown in Table 1. The model of OOD detection layer with ViLT has significant improvements in image retrieval and text retrieval tasks. Overall, our model achieved the best and second-best results on both datasets comparing to other SOTA models.

Training Objectives with OOD deteation Layer. To evaluate our proposed model, we perform more ablations with the default training settings of the baseline and our model mentioned in Section 3.3 of the main paper. We consider different combinations of train objectives and evaluate in zero-shot settings. We observe our model performance on training objectives. Note that, ViLT is our implementation with the same subset of training datasets. IOur raw results are presented in Table 2. We train on pre-train datasets with $\mathcal{L}_{\text{itm}}$ in Equation 7, $\mathcal{L}_{\text{mlm}}$ in Equation 8 and $\mathcal{L}$ in Equation 9. The results in Table 2 show that training on image-text matching and masked language modeling is beneficial for both downstream tasks comparing to the baseline model, especially, there is promising improvements in image retrieval and text retrieval tasks. Thus, it is beneficial to train on both ITM and MLM for filtering outlier concepts and improve performance on downstream tasks.

C.3 More Qualitative Analysis

As the feature maps shown in Fig 2 of the main paper, our model result demonstrates more clusters can be identified over the multimodal features extracted by VK-OOD. Thus, it illustrates that our model is able to detect outliers and cluster images closest to the corresponding μ_k with image and explicit knowledge triplets. Figure 1a and Figure 1b are examples that the nearest images in each cluster.

Figure 2 and Figure 3 show more qualitative examples of our multimodal alignment results of our models. We visualize the multimodal attention maps on images corresponding to concept triplets using Grad-cam designed by [12]. Following our model architecture, the caption is parsed and integrated with knowledge triplets. The right bottom subfigure in our model in Figure 2 and Figure 3



(a) example cluster 1

(b) example cluster 2

Figure 1: Example images in the clusters on COCO val set.

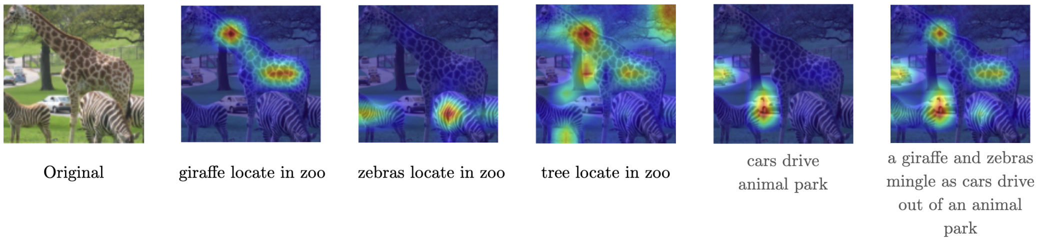
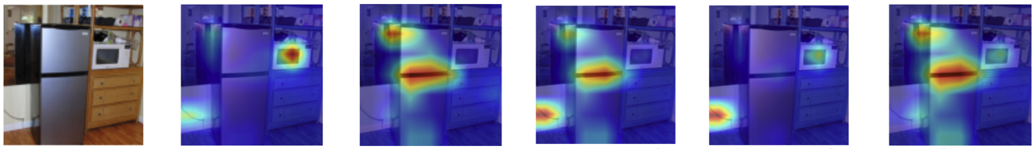


Figure 2: Visualization of the attention maps of image feature v_q and language features l_j of external knowledge alignment. The results are from our VKOOD-ViLT model. The original sample caption is “a giraffe and zebras mingle as cars drive out of an animal park". We highlight areas in the example image corresponding to different knowledge triplets. Comparing with the attention maps of the baseline model, our model learns object shapes such as zebras and localize those objects correctly.

are the multimodal alignment of original captions from MSCOCO [5] dataset. Other subfigures show the alignments of extracted triplets on the image.

Interestingly, we find that our model is able to capture concept “plug” as a part of “refrigerator” or “microwave” in Figure 2. The heatmap area of “plug” and “microwave” in Figure 2 clearly suggest that our model has the capability to exploit different relevance between visual and corresponding conceptual text features. By contrast, the baseline results have not shown the relation between plug and microwave. In Figure 3, it shows that we detect three zebras comparing with baseline, but counting cars is not performing well as we expected – since the size (or scale) of cars is not sufficiently high, and moreover some parts of them are occluded.



Original

microwave locate in room

refrigerator locate in room

plug is part of refrigerator

plug is part of microwave

a refrigerator freezer next to a microwave oven

Figure 3: Visualization of the attention maps of image and knowledge concept triplets alignment. The results are from our VKOOD-ViLT model. The original sample caption is “a metallic refrigerator freezer next to a microwave oven". We highlight areas in the example image corresponding to different knowledge triplets. Comparing with the attention maps of the baseline model, our model learns the relations between the parts (i.e., plug) of the objects correctly.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[8] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[10] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[13] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[14] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.