MACK: Multimodal Aligned Conceptual Knowledge for Unpaired Image-text Matching

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, the accuracy of image-text matching has been greatly improved by multimodal pretrained models, all of which are trained on millions or billions of paired images and texts. Different from them, this paper studies a new scenario as unpaired image-text matching, in which paired images and texts are assumed to be unavailable during model training. To deal with this, we propose a simple yet effective method namely Multimodal Aligned Conceptual Knowledge (MACK), which is inspired by the knowledge use in human brain. It can be directly used as general knowledge to correlate images and texts even without model training, or further fine-tuned based on unpaired images and texts to better generalize to certain datasets. In addition, we extend it as a re-ranking method, which can be easily combined with existing image-text matching models to substantially improve their performance.

1 Introduction

Image-text matching is one of the most representative technique in the field of vision & language understanding, which has wide applications such as online shopping, human-robot interaction and autonomous driving. Its major challenge lies in how to accurately measure the cross-modal similarity between images and texts. Recently, by training on very large-scale (millions or billions) paired images and texts, various multimodal pretrained models [18, 15, 7] alleviate the challenge in a supervised learning manner. Although the accuracy of image-text matching is greatly improved, collecting and annotating such large-scale data in real world applications is time-consuming and expensive.

Different from them, this work attempts to study the image-text matching in the context of a new scenario, namely *unpaired image-text matching*, in which paired images and texts are assumed to be unavailable during model training. It is motivated by the fact the human brain can well correlate arbitrary images with texts while does not need to learn from such large-scale paired images and texts. Instead, it stores semantic knowledge about objects, relations, actions, attributes, *etc.*, which is multimodal aligned and can be used to correlated visual and linguistic information [23, 2]. Inspired by these neuroscience evidences, this work tries to deal with the unpaired image-text matching by modeling humain brain-like knowledge.

Unimodal visual or linguistic knowledge has been widely used for vision & language understanding such as image captioning [34], visual reasoning [39] and visual question answering [26]. By directly combining these two types of knowledge together, multimodal knowledge shows more complementary advantages [40, 32]. However, existing knowledge-based methods are always designed to corporate with paired model training, which cannot well handle the unpaired image-text matching. Another alternative is multimodal knowledge graphs [41], most of which are extended from existing linguistic knowledge graphs by linking semantically related images to corresponding words. However, the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

image-word alignment might have the following issues: 1) images usually include redundant or word-unrelated contents, so the images and words carry unequivalent semantic information, and 2) word-related objects (or attributes) usually have diverse appearances in different images, so the alignment is one-to-many rather than one-to-one. In fact, obtaining well-aligned multimodal knowledge is a great challenge.

In this work, we propose a new method namely Multimodal Aligned Conceptual Knowledge (MACK) for unpaired image-text matching. To remove the word-unrelated content in images, it focuses on semantic concepts and collects pairs of words and semantically related image regions from public available datasets. Then it computes prototypical region representations and aligns them to the words, with the goal to alleviate the appearance variant. Based on the aligned conceptual knowledge (*i.e.*, word-prototypical region pairs), the MACK can bridge images and texts in the same feature space to measure their cross-modal similarities. To make the pre-computed general knowledge better suit certain datasets, we can further fine-tune it with the principle of region-level cycle-consistency, which does not need paired images and texts. Since the proposed MACK is simple yet effective, it can well corporate with existing image-text matching models as a re-ranking method to further improve their performance.

Our contributions are summarized as follows. 1) We study a rarely investigated problem as unpaired image-text matching, which is under great demand in practical applications. 2) We accordingly propose a simple yet effective method to model multimodal aligned conceptual knowledge, which can either perform test directly or perform training based on unpaired data. 3) The proposed method can be easily adopted by existing image-text matching models to re-rank their results to obtain better performance.

2 Related Work

2.1 Image-Text Matching

Image-text matching draws much attention recently, many effective models trained on paired images and texts have been proposed. Socher *et al.* [27] and Frome *et al.* [6] propose the early framework of Visual-Semantic Embedding (VSE) to correlate images and their labels. Kiros *et al.* [11] extend the VSE from image-label matching to image-text matching, which is later improved from various aspects such as adding intra-modal constraint [31], mining hardest negative sample [5], analyzing canonical correlation [36], improving sentence representation [24, 14], and reasoning visual relation [17, 9]. In addition to the VSE, another important framework is Cross-Modal Attention (CMA) proposed by Lee *et al.* [13], which uses pretrained object-level features [1] and then generates cross-modal attended representations for similarity measurement. The CMA is later improved from the directions of cross-modal memory [8], cross-modal message passing [33], context modeling [38], iterative matching [3], and graph structure [20]. Recently, many more complicated models [4, 15, 22, 18] based on multimodal versions of Transformer [30] are proposed, which rely on millions or billions of paired images and texts for model training and achieve very good performance. In contrast to the existing models, we attempt to study a new scenario of unpaired image-text matching that assumes paired images and texts are unavailable for model training, which is more challenging than conventional image-text matching.

2.2 Multimodal Knowledge

There are only a few works that use multimodal knowledge for vision & language understanding tasks. For example, Zhu *et al.* [40] build heterogeneous graphs corresponding to visual, semantic and factual features for visual question answering. Wang *et al.* [32] combine visual and textual knowledge to find discriminative parts of objects and improve the performance of few-shot learning. In addition, different multimodal knowledge graphs [41, 21, 28] are proposed, most of which extend linguistic knowledge graphs by aligning images to words. Such cross-modal alignment could be quite noisy because images usually include redundant or word-unrelated contents. In fact, the multimodal knowledge used in these methods is either cross-modal unaligned or noisy aligned, which usually needs paired images and texts as strong supervisions for model training. Different from them, our multimodal aligned conceptual knowledge establishes more accurate one-to-one alignment in the fine-grained concept-level, which can handle the unpaired image-text matching without using strong supervisions.

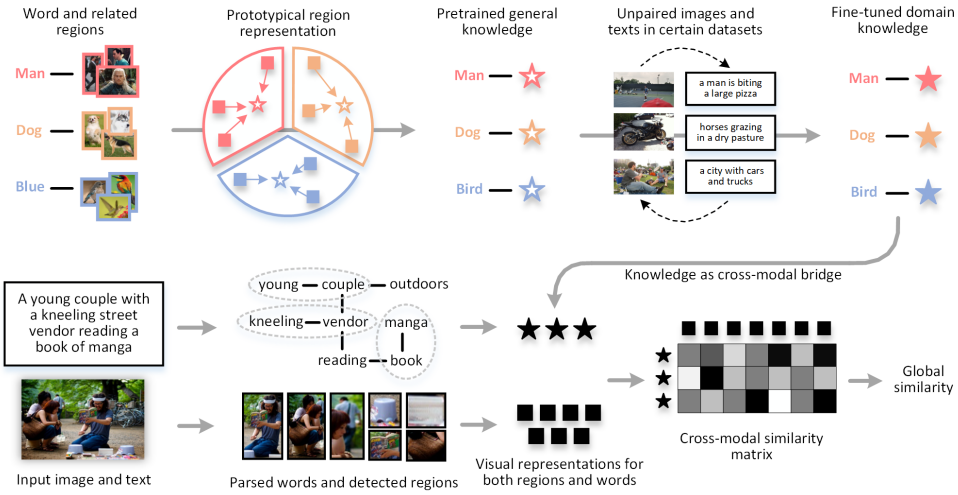

Figure 1: The proposed Multimodal Aligned Conceptual Knowledge (MACK) for unpaired image-text matching. The top figures illustrate how to obtain the knowledge in three steps: 1) collecting a set of conceptual words and their related image regions, 2) computing prototypical region representations to get the pretrained general knowledge, and 3) training on unpaired images and texts to obtain the fine-tuned domain knowledge. The bottom figures illustrate how to use the knowledge for unpaired image-text matching in three steps: 1) obtaining parsed words and detected regions from the input image and text, 2) aggregating related words and representing them by prototypical region representations in the knowledge, and 3) comparing word and region representations in the same feature space and pooling their cross-modal matrix to obtain a global similarity score.

3 Multimodal Aligned Conceptual Knowledge

As shown in Figure 1, we propose the Multimodal Aligned Conceptual Knowledge (MACK) for unpaired image-text matching. To obtain the knowledge, we first collect a set of conceptual words and their semantically related image regions from public datasets. For each word, obtaining a prototypical region representation by averaging the representations of all related regions. Thus, we obtain the pretrained general knowledge containing pairwise words and prototypical region representations. To make the knowledge better adapt to certain datasets, we then fine-tune the pretrained general knowledge by modeling the concept-level cycle consistency based on unpaired images and texts.

Give an image and a text, we first perform text parsing to obtain a word graph, and use the knowledge as a cross-modal bridge to obtain a set of prototypical region representations for all the words. Then, we perform object detection on the image to obtain another set of region representations. Thus, the representations of words and regions are in the same feature space, in which they can be directly compared as a cross-modal matrix. After pooling the values in row and column dimensions based on heuristic rules, we can finally obtain the global similarity score between the image and text.

3.1 Pretrained General Knowledge

The studied knowledge in this work has two major properties as follows. 1) Concept-level knowledge: the knowledge is built upon three major types of semantic concepts, which correspond to objects, attributes and actions in images, and nouns, adjectives and verbs in texts. It is more fine-grained than existing multimodal knowledge graphs that roughly link global images to words. 2) One-to-one alignment: rather than align a word to multiple possible images in a one-to-many manner, the knowledge focuses on cross-modal one-to-one alignment, with the goal to alleviate the appearance variations of objects, attributes and actions.

We define the knowledge as a set of semantic concepts having paired multimodal representations $\{w_k, v_k\}_{k=1,\dots,K}$, where $w_k \in \mathbb{N}^K$ and $v_k \in \mathbb{R}^F$ are the one-hot word representation and real-valued region representation of k-th semantic concept, respectively, and K is the total number of semantic concepts. To maximize the number of concepts, we collect all the words and their related regions in the Visual Genome (VG) dataset [12]. We could alternatively use other datasets, but the VG has

more diverse content that is useful for improving the generalization ability of the knowledge. For each word, we can easily obtain its semantically related image regions in the dataset, and different regions usually have diverse visual appearances. To avoid the impact of appearance variant, we compute a prototypical region representation v_k for each word w_k by averaging all its related region representations $\{r_j\}_{j=1,...,J_k}$:

$$v_k = \frac{1}{J_k} \sum\nolimits_{j=1}^{J_k} r_j \tag{1}$$

where each region representation r_j is obtained by feeding a bounding box and an image into the pretrained object detection model Faster-RCNN[1] [1]. Note that the numbers of regions for different words $\{J_k\}_{i=k,...,K}$ are quite imbalanced, which vary from tens to tens of thousands. Intuitively, frequently appeared words such as "man" and "dog" will have more related regions, and their computed prototypical region representations might be more robust than those of few-shot words such as "otter" and "nun". However, the few-shot words can still compute initial prototypical region representations with very limited numbers (*e.g.*, < 20) of regions and then refine them when more related regions are encountered.

3.2 Fine-tuned Domain Knowledge

To better adapt the pretrained general knowledge to certain datasets, we could further fine-tune the prototypical region representations based on the principle of concept-level cycle consistency, if unpaired images and texts in those datasets are available. In particular, given a batch of unpaired images and texts, we first obtain a set of detected regions using the Faster-RCNN above and a set of tokenized words. For the set of words, duplicated ones[2] are removed and the rest words are $\{w_m\}_{m=1,...,M}$. Then all the words can be represented by the corresponding prototypical region representations as $\{v_m\}_{m=1,...,M}$ or $V \in \mathbb{R}^{F\times M}$. For the set of regions, we compute similarity scores between the prototypical region representations (in the pretrained general knowledge) and the set of region representations, and then remove the redundant regions belonging to same concepts in a similar way as Non-Maximum Suppression (NMS). The resulting regions are denoted as $\{r_n\}_{n=1,...,N}$ or $R \in \mathbb{R}^{F\times N}$.

To fine-tune the prototypical region representations to better suit for certain datasets, we design a cycle consistent loss to learn a parametric transformation matrix $W \in \mathbb{R}^{F\times F}$. The main idea is to first measure similarities between each word and all regions, and use the similarities as weights to combine all regions to obtain a reconstructed word

$$\hat{V} = WR\sigma(S)^T,\ S = (WV)^T WR \tag{2}$$

where $S \in \mathbb{R}^{M\times N}$ is the similarity matrix between transformed word and region representations, $\sigma(\cdot)$ is the softmax operation in the column dimension, and $\hat{V} \in \mathbb{R}^{F\times M}$ contains the reconstructed word representations. Then, each original word and its reconstructed one are compared to generate a label indicating whether they are the same or not. By minimizing the cross-entropy loss L based on the predicted label and groundtruth one, we can optimize the W:

$$L = \sum\nolimits_{m=1}^{M} y_m^T \log(\hat{y}_m),\ \hat{Y} = \sigma((WV)^T \hat{V}) \tag{3}$$

where $\hat{Y} \in \mathbb{R}^{M\times M}$ includes the predicted labels, in which each column is denoted as $\hat{y}_m$, and y_m is the groundtruth label in which the m-th value is one and the rest values are zeros. After the training, we can use the W to transform all prototypical region representations to obtain the fine-tuned domain knowledge, denoted as $\{w_k, u_k\}_{k=1,...,K}$, where $u_k = Wv_k \in \mathbb{R}^F$.

3.3 Knowledge-based Unpaired Matching

To decide whether an image and a text is matched or not, we use the knowledge[3] to measure the cross-modal similarity. For the image, we use the Faster-RCNN to obtain a set of detected region representations $\{r_i\}_{i=1,...,I}$ or $R \in \mathbb{R}^{F\times I}$. For the text, we use the knowledge as a cross-modal

[1] We could alternatively use other pretrained models to detect regions. We select this model because it has been widely used and demonstrated to be effective for image-text matching.

[2] We also remove different tenses of verbs and the plural of nouns.

[3] The knowledge here could be the pretrained general knowledge or fine-tuned domain knowledge.

bridge to represent all the words as a set of prototypical region representations $\{u_j\}_{j=1,...,J}$. Even though these two set of representations are in the same feature space, they have different amount of semantic information, directly comparing their similarity might be sub-optimal. In fact, each u_j is concept-level corresponding a word, while each r_i usually describes not only an object but also its attribute or action. To make a better comparison, we infer the relation among words and aggregate related word representations before the similarity measurement.

In particular, for the words in the text, we use the Natural Language Toolkit (NLTK)[4] to predict their types and parse their dependencies. Then, we take each noun as an anchor and combine its representation with depended adjective representation or verb representation in an averaged manner. Thus, the set of word representations $\{u_j\}_{j=1,...,J}$ are re-computed as $\{u_l\}_{l=1,...,L}$ or $U \in \mathbb{R}^{F\times L}$. Then we can obtain the desired similarity score s for the image and text as:

$$s = \rho(U^T R) \tag{4}$$

where $\rho(\cdot)$ is the max-mean pooling that first performs max pooling on the column dimension and then mean pooling on the row dimension on the input matrix. Given multiple images and texts, we can similarly obtain their similarity scores for unpaired image-text matching.

3.4 Re-ranking Existing Models

Note that the knowledge in this work is inspired by human-like knowledge, which is quite different from the model knowledge learnt from paired data. So the proposed MACK tends to have complementary properties when combing with existing models. To demonstrate this idea, we try to use the knowledge to re-ranking the results by existing models to improve their performance.

Taking the text-driven image retrieval as an example, given a text as query and a set of N images as gallery, an existing model can produce a similarity vector $\mathbf{s} \in \mathbb{R}^{N\times 1}$. By sorting all the scores of $\mathbf{s}$ in the descending order, the model usually is able to rank the matched image in top-k. Then we can tune the top-k scores by adding the corresponding scores predicted by the knowledge with a balancing factor λ, and then re-rank the top-k images to improve the rank of matched image.

4 Experimental Results

To demonstrate the effectiveness of the proposed method, we perform extensive experiments of image-text matching on two publicly available datasets.

4.1 Datasets and Protocols

The details of datasets and protocols are as follows. 1) Flickr30k [37] consists of 31783 images collected from the Flickr website. Each image has 5 human annotated texts. We use the public training, validation and testing splits, which contain 29000, 1000 and 1000 images, respectively. 2) MSCOCO [19] consists of 123287 images, each of which is associated with 5 texts. We use the public training, validation and testing splits, with 113287, 1000 and 5000 images, respectively.

The image-text matching usually includes two sub-tasks in terms of: 1) image annotation: retrieving related texts given images, and 2) image retrieval: retrieving related images given texts. The commonly used evaluation criterions are "R@1", "R@5" and "R@10", *i.e.*, recall rates at the top-1, 5 and 10 results. Following other methods, we also use an additional criterion of "Rs" by summing all the recall rates to evaluate the overall performance.

4.2 Implementation Details

In the pretrained general knowledge, we collect all the words in the Visual Genome dataset, so the total number of semantic concepts is K = 27801. For each image, we use the pretrained Faster-RCNN to extract region representations, so the number of detected regions is I = 36 and the dimension of (prototypical) region representation is F = 2048. In the fine-tuned domain knowledge, we perform the model training in a minibatch manner. Each minibatch includes unpaired 128 images and 128 texts that are randomly sampled from the whole training set. When removing the redundant regions

[4] https://www.nltk.org/.

Table 1: Unpaired image-text matching by ablation models of MACK on the Flickr30k dataset. Details about these ablation models are explained in Section 4.3.

Method	Flickr30k dataset						
	Image Annotation			Image Retrieval			Rs
	R@1	R@5	R@10	R@1	R@5	R@10	
Pretrained general knowledge	10.8	26.6	35.8	4.1	10.8	18.0	106.2
- w/o region prototypes	1.5	6.1	9.3	1.1	4.5	9.0	31.7
- w/o adjectives or verbs	9.8	25.0	34.8	4.2	10.9	16.8	101.5
- w/o max-mean pooling	4.2	13.4	20.4	2.9	8.4	12.8	62.2
Fine-tuned domain knowledge	12.7	30.4	40.8	10.3	25.1	34.0	153.5

Table 2: Image-text matching by re-ranking two state-of-the-art models on the Flickr30k and MSCOCO datasets. "+" denotes the re-ranking.

Method	Flickr30k dataset							MSCOCO dataset						
	Image Annotation			Image Retrieval			Rs	Image Annotation			Image Retrieval			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
CLIP [25]	65.4	87.2	91.7	85.4	97.1	98.7	525.6	35.3	60.0	70.1	55.2	78.7	86.7	386.1
+ MACK	66.8	88.2	92.6	86.2	97.2	98.9	530.0	36.9	61.6	71.7	55.7	79.6	87.1	392.8
ALBEF [16]	59.9	84.8	90.5	78.2	95.5	97.9	506.9	40.2	68.4	78.9	62.4	85.9	92.1	428.3
+ MACK	61.8	85.8	91.4	80.1	96.4	97.7	513.3	41.0	69.0	79.4	62.4	86.1	92.7	430.9

belonging to same semantic concepts, we keep only one region that has the largest similarity score compared with prototypical region representations. We use the Adam algorithm [10] to optimize the only parameter matrix W with a learning rate of $2e^{-4}$ for 30 epochs. When re-ranking existing models, we empirically set k = 15 and λ = 0.1. The knowledge-based unpaired image-text matching is very efficient, which does not need GPUs for acceleration.

4.3 Unpaired Image-text Matching

Since existing models are all proposed for paired image-text matching, we compare several ablation models of the proposed MACK to verify their effectiveness for unpaired image-text matching. The first ablation model is pretrained general knowledge, which does not need model training. Then we design another three models based on it by: 1) replacing the prototypical region representation with a random region representation, similar to existing multimodal knowledge graphs, 2) removing the adjectives or verbs when re-computing the word representations, and 3) replacing the max-mean pooling with mean-mean pooling. We also present another ablation model as fine-tuned domain knowledge, which additionally performs a training stage on unpaired images and texts using the concept-level cycle consistent loss.

We compare the performance of unpaired image-text matching by all these models on the Flickr30k dataset in Table 1. It is reasonable that the pretrained general knowledge achieves much worse performance than existing image-text matching models trained on paired images and texts. The unpaired image-text matching still needs more research and there is still a lot of room for performance improvement. By removing the region prototypes, adjectives or verbs, and max-mean pooling, the performance all becomes worse. And we find that the region prototypical representations make the largest contributions to the performance of pretrained general knowledge. By training on the unpaired images and texts, the fine-tuned domain knowledge shows significant performance improvements, especially on the sub-task of image retrieval. In the following experiments, we use the pretrained general knowledge as the default MACK due to its simplicity.

4.4 Re-ranking State-of-the-art Models

The recent state-of-the-art models for image-text matching are variants of multimodal Transformers, all of which are pretrained on very large-scale datasets containing millions or billions of paired images and texts. To demonstrate that our knowledge-based MACK is complementary with data-based existing models, we use the MACK to re-rank the retrieval results of two state-of-the-art models

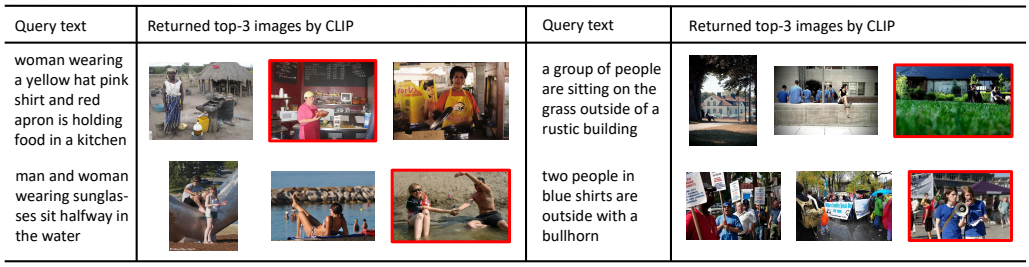

Figure 2: Examples of retrieved top-3 images based on text queries by the CLIP. Groundtruth matched images are marked by red bounding boxes, which can be re-ranked higher by the MACK.

Table 3: Cross-dataset image-text matching by re-ranking existing models on the Flickr30k and MSCOCO datasets. "+" denotes the re-ranking. MSCOCO -> Flickr30k: training existing models on the MSCOCO dataset and testing them on the Flickr30k dataset. Flickr30k -> MSCOCO: training existing models on the Flickr30k dataset and testing them on the MSCOCO dataset.

Method	MSCOCO -> Flickr30k							Flickr30k -> MSCOCO						
	Image Annotation			Image Retrieval			Rs	Image Annotation			Image Retrieval			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VSRN [25]	42.3	69.3	78.1	53.1	79.5	87.5	409.9	37.8	68.5	79.3	50.2	79.5	87.8	403.4
+ MACK	42.6	69.8	78.5	53.3	79.7	87.7	411.7	38.0	68.8	79.5	50.4	79.6	87.9	404.4
SAEM [35]	52.1	81.3	88.7	70.1	90.6	95.4	478.3	34.5	65.6	77.8	46.7	77.3	86.4	388.5
+ MACK	52.5	81.8	88.7	70.0	91.0	95.5	479.6	34.7	65.6	77.9	46.9	77.4	86.5	389.2

including CLIP [25] and ALBEF [16]. The original and re-ranked performance of these two models on the Flickr30k and MSCOCO datasets are shown in Table 2. It should be noted that, we use the zero-shot versions of these two models with the goal to make fair comparisons with the proposed MACK. The zero-shot means that they are pretrained on large-scale external datasets while directly perform test on the Flickr30k and MSCOCO datasets. From the table, we can see that although their original accuracies are very high, re-ranking with our proposed MACK can further improve them. For the CLIP, the re-ranking can lead to 1.4% and 1.6% improvements in R@1 of image annotation on the two datasets, respectively. For the ALBEF, the re-ranking can lead to overall 6.4% improvement in Rs on the Flickr30k dataset.

These evidences demonstrate that the MACK does show complementary properties when combing with existing models. To further explain this, we show some examples of retrieved images based on text queries by the CLIP in Figure 2. These examples are selected based on the following rules: 1) ranks of groundtruth matched images (marked by red bounding boxes in the figure) by the CLIP are not top-1, and 2) our proposed MACK can re-rank these matched images higher. In the first example, it seems that the CLIP cannot well understand the "yellow hat", "red apron" or "pink shirt" in the query text, since the first retrieved image does not contain either of these semantic concepts but ranks higher than the groundtruth one. While our MACK especially focuses on these semantic concepts and can improve the corresponding similarity between the matched image and query text.

4.5 Cross-dataset Image-text Matching

We also try to re-rank existing image-text matching models when applying them to the scenario of cross-dataset image-text matching. The experimental settings are explained as follows: 1) selecting two representative image-text matching models (VSRN [25] and SAEM [35]) and training them on a source dataset (*e.g.*, Flickr30k or MSCOCO), 2) performing model test on a different target dataset (*e.g.*, MSCOCO or Flickr30k), and 3) using the proposed MACK to re-rank their results on the target dataset. Note that either the two models or our MACK does not use the target dataset for model training.

In Table 3, the results of two directions of cross-dataset generalization, *i.e.*, MSCOCO -> Flickr30k and Flickr30k -> MSCOCO, are both presented. We can find that using the MACK to re-rank the VSRN and SAEM can improve their generalization abilities in the cross-dataset scenario. Although

Table 4: Image-text matching by ALBEF + MACK on the Flickr30k and MSCOCO datasets. f is the maximum number of used regions when computing a prototypical region presentation.

f	Flickr30k dataset							MSCOCO dataset						
	Image Annotation			Image Retrieval			Rs	Image Annotation			Image Retrieval			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
20	60.0	85.1	91.2	79.7	96.0	97.7	509.8	39.7	68.0	79.1	62.2	85.6	92.7	427.5
50	60.2	85.3	91.1	79.9	96.0	97.7	510.3	39.7	68.1	79.1	62.2	85.7	92.7	427.8
100	60.7	85.3	91.2	80.1	95.9	97.7	510.9	39.9	68.3	79.2	62.5	86.0	92.7	428.8
∞	61.8	85.8	91.4	80.1	96.4	97.7	513.3	41.0	69.0	79.4	62.4	86.1	92.7	430.9

the relative improvements might be a little bit small, they are substantial in all 7 evaluation criterions. Comparing Table 3 with Table 2, it seems that the performance improvements of re-ranking partially depend on the performance of base models. When the base models are more accurate, the performance improvements of re-ranking are larger.

4.6 Prototypical Representation Analysis

In the proposed MACK, one of the most important part is the prototypical region representation. Even the same object could have completely different appearances in images, so using how many regions to compute each region prototype is crucial. We count the number of regions for each word in the Visual Genome dataset and show the statistical result in Figure 3, in which the horizontal axis is word index and vertical axis is logarithmic transformed number of regions. From the figure, we can see that about 80% words have very small numbers of regions, *i.e.*, less than 20 (denoted by the dashed line). Therefore, for most semantic concepts in the knowledge, computing their prototypical region representations is data-efficient.

To verify whether using small numbers of regions has large impact on the performance, we define a threshold f that controls the maximum number of regions. We use the ALBEF + MACK as the base model and select f from {20, 50, 100, ∞ }, where ∞ is the default setting of ALBEF + MACK. The experimental results on the Flickr30k and MSCOCO datasets are shown in Table 4. We can see that as the f becomes larger the performance becomes slightly better. It seems that different numbers of regions do not have a significant impact on the performance. It is reasonable since computing the mean might not need too many samples. This is also a good property for the knowledge, since if we want to add a new semantic concept, we only have to annotate a limited number of regions.

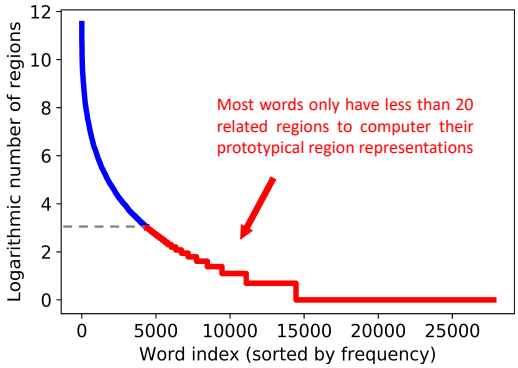

Figure 3: Logarithmic number of regions *v.s.* word indexes. (best viewed in colors).

4.7 Knowledge Visualization

At last, we visualize the pretrained general knowledge and fine-tuned domain knowledge (trained on the Flickr30k dataset) in Figure 4, to illustrate the major change after the training. In the figure, there are two low-dimensional word distributions (surrounded by solid boxes) corresponding to the pretrained general knowledge and fine-tuned domain knowledge, respectively. The two word distributions are obtained by using the t-SNE method [29] to embed two set of prototypical region presentations (*i.e.*, $\{v_k\}$ or $\{u_k\}$, k = 1,...,K). For clear illustration, we set K = 100 and only select the most frequently used 100 semantic concepts in the Flickr30k dataset. For each semantic concept, we show the corresponding word in the two-dimensional coordinate. We also group semantically related words and show them with different colors.

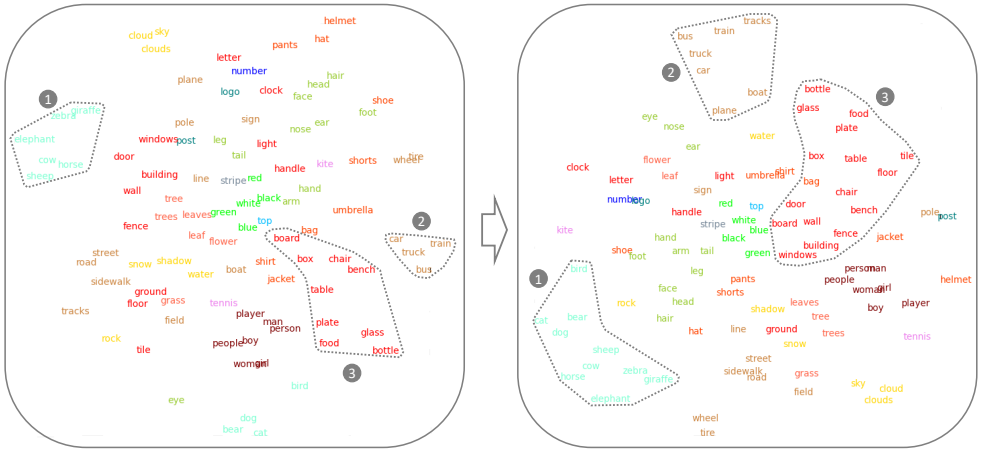

Figure 4: Visualization of pretrained general knowledge and fine-tuned domain knowledge. Each word indicates its corresponding prototypical region representation embedded by the t-SNE method. In the two word distributions, we group semantically related words and represent them using same colors. We also mark three representative groups by dashed lines, which contain words about animals, indoor objects and transports, respectively.

In the left word distribution, we can see that semantically related words stay nearby. For example, the words in the three groups (marked by dash lines) are about animals, indoor objects and transports, respectively. In the center area, there are many words (marked by the green color) about colors. However, there are still some related words that have remote distances. For example, in the bottom there are also several words about animals such as bird, dog, bear and cat, which are far from the words in the first group. After fine-tuning the knowledge with unpaired data, we find the word distribution is semantically more compact. We can see that all the words about animals are very close staying together in the first group. In the second group, more words about transports such as plane and boat are added. These evidences indicate that by training with the concept-level cycle consistency, related semantic concepts are able to leverage their potential relations to make the prototypical region representations more discriminative.

5 Conclusion and Future Work

This work have studied a practically important but seldom investigated problem as unpaired image-text matching, in which paired images and texts are not available during model training. To deal with this problem, inspired by the knowledge use in human brain, we have proposed a simple yet effective method namely Multimodal Aligned Conceptual Knowledge (MACK). Compared with existing multimodal knowledge methods, the MACK has two majors properties in terms of conceptual level and one-to-one alignment. It can either perform test directly or perform training based on unpaired data, as well as re-rank existing image-text matching models to improve their performance. We have demonstrated its effectiveness with extensive experiments.

To the best of our knowledge, this might be the first work to study the unpaired image-text matching. So the goal of this initial work is not pursuing high performance but to explore the modeling of knowledge. In the future, we will study effective strategies to improve the performance. In addition, we will also consider to mine more useful information from the knowledge from the aspects of graphic knowledge structure, dynamic knowledge updating and advanced knowledge reasoning.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[2] Y. Bi. Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, 25(10):883–895, 2021.

[3] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12655–12663, 2020.

[4] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120, 2020.

[5] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017.

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Neural Information Processing Systems*, pages 2121–2129, 2013.

[7] G. Geigle, J. Pfeiffer, N. Reimers, I. Vulić, and I. Gurevych. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *arXiv preprint arXiv:2103.11920*, 2021.

[8] Y. Huang and L. Wang. Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In *IEEE International Conference on Computer Vision*, pages 5774–5783, 2019.

[9] Y. Huang, Q. Wu, and L. Wang. Learning semantic concepts and order for image and sentence matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics*, 2015.

[12] 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.

[13] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In *European Conference on Computer Vision*, pages 212—228, 2018.

[14] G. Lev, G. Sadeh, B. Klein, and L. Wolf. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*, pages 833–850, 2016.

[15] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI Conference on Artificial Intelligence*, pages 11336–11344, 2020.

[16] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Neural Information Processing Systems*, 34, 2021.

[17] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu. Visual semantic reasoning for image-text matching. In *IEEE International Conference on Computer Vision*, pages 4654–4662, 2019.

[18] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137, 2020.

[19] 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, 2014.

[20] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang. Graph structured network for image-text matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930, 2020.

[21] Y. Liu, H. Li, A. Garcia-Duran, M. Niepert, D. Onoro-Rubio, and D. S. Rosenblum. Mmkg: multi-modal knowledge graphs. In *European Semantic Web Conference*, pages 459–474, 2019.

[22] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-task vision and language representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[23] K. Patterson, P. J. Nestor, and T. T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12):976–987, 2007.

[24] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[25] 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, 2021.

[26] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI Conference on Artificial Intelligence*, pages 8876–8884, 2019.

[27] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. *Neural Information Processing Systems*, 26, 2013.

[28] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng. Multi-modal knowledge graphs for recommender systems. In *ACM International Conference on Information & Knowledge Management*, pages 1405–1414, 2020.

[29] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Neural Information Processing Systems*, 30, 2017.

[31] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016.

[32] S. Wang, J. Yue, J. Liu, Q. Tian, and M. Wang. Large-scale few-shot learning via multi-modal knowledge discovery. In *European Conference on Computer Vision*, pages 718–734, 2020.

[33] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *IEEE/CVF International Conference on Computer Vision*, pages 5764–5773, 2019.

[34] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel. What value do explicit high level concepts have in vision to language problems? In *IEEE conference on computer vision and pattern recognition*, pages 203–212, 2016.

[35] Y. Wu, S. Wang, G. Song, and Q. Huang. Learning fragment self-attention embeddings for image-text matching. In *ACM International Conference on Multimedia*, pages 2088–2096, 2019.

[36] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3441–3450, 2015.

[37] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[38] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li. Context-aware attention network for image-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3536–3545, 2020.

[39] Y. Zhang, M. Jiang, and Q. Zhao. Explicit knowledge incorporation for visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1356–1365, 2021.

[40] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, Y. Xiao, and N. J. Yuan. Multi-modal knowledge graph construction and application: A survey. *arXiv preprint arXiv:2202.05786*, 2022.

[41] Y. Zhu, C. Zhang, C. Ré, and L. Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work does not have potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sections 4.1 and 4.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.1.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]