Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Robust cross-modal retrieval with alignment refurbishment*

Jinyi GUO[1], Jieyu DING[‡2]
[1]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*
[2]*School of Mathematics and Statistics, Qingdao University, Qingdao 266071, China*
E-mail: jinyi_g@njust.edu.cn; djy@qdu.edu.cn

Received Oct. 27, 2022; Revision accepted Feb. 16, 2023; Crosschecked Sept. 19, 2023

Abstract: Cross-modal retrieval tries to achieve mutual retrieval between modalities by establishing consistent alignment for different modal data. Currently, many cross-modal retrieval methods have been proposed and have achieved excellent results; however, these are trained with clean cross-modal pairs, which are semantically matched but costly, compared with easily available data with noise alignment (i.e., paired but mismatched in semantics). When training these methods with noise-aligned data, the performance degrades dramatically. Therefore, we propose a robust cross-modal retrieval with alignment refurbishment (RCAR), which significantly reduces the impact of noise on the model. Specifically, RCAR first conducts multi-task learning to slow down the overfitting to the noise to make data separable. Then, RCAR uses a two-component beta-mixture model to divide them into clean and noise alignments and refurbishes the label according to the posterior probability of the noise-alignment component. In addition, we define partial and complete noises in the noise-alignment paradigm. Experimental results show that, compared with the popular cross-modal retrieval methods, RCAR achieves more robust performance with both types of noise.

Key words: Cross-modal retrieval; Robust learning; Alignment correction; Beta-mixture model
https://doi.org/10.1631/FITEE.2200514 **CLC number:** TP391

1 Introduction

In this paper, we focus on the robust image–text cross-modal retrieval problem, which involves searching an image (or text) for a given sentence (or image). It offers a broader range of applications and provides a better user experience than uni-modal retrieval, such as news search and product retrieval (Wang KY et al., 2016). State-of-the-art algorithms are trained with paired multi-modal data (e.g., Fig. 1a) and provide good results. Nonetheless, those clean paired data are modally aligned, which are expensive. With the explosive growth of multimedia data, the cross-modal data collected from the Internet are easily available, but most of them have some noise alignments, i.e., paired data but mismatched semantically. In general, these data exist in three forms: clean alignment, partial noise alignment, and complete noise alignment (Fig. 1). Experiments reveal that current methods perform badly in the context of noise-aligned data. Therefore, we propose a new method, named robust cross-modal retrieval with alignment refurbishment (RCAR), to solve the noise-alignment image–text retrieval problem.

Traditional cross-modal retrieval methods (Faghri et al., 2018; Li KP et al., 2019; Chen H et al., 2020; Diao et al., 2021) project different modal data into a shared semantic space, treat paired modal data as positive instances and unpaired ones as negative

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 12172186)

ORCID: Jinyi GUO, https://orcid.org/0000-0002-4439-8693; Jieyu DING, https://orcid.org/0000-0003-2085-5383

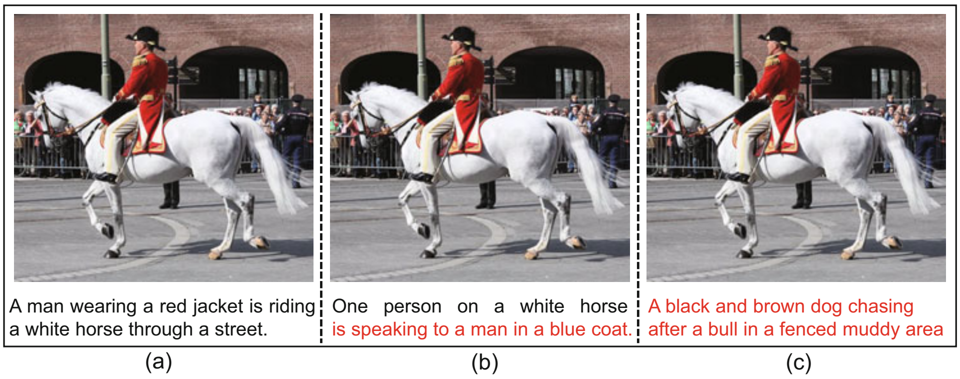

A man wearing a red jacket is riding a white horse through a street.
One person on a white horse is speaking to a man in a blue coat.
A black and brown dog chasing after a bull in a fenced muddy area
(a)
(b)
(c)

Fig. 1 Three types of data-alignment instances: (a) a clean instance is modally aligned, meaning that image and text have consistent semantics; (b) a partial noise-alignment instance denotes a pair with partially mismatched semantics; (c) a complete noise-alignment instance indicates a pair with entirely mismatched semantics. Noise semantics are marked in red. References to color refer to the online version of this figure

instances and are optimized by contrastive learning, as shown in Eq. (1), which maximizes the image–text similarity between positive instances (i,t) and minimizes the similarity between negative instances $(i,\hat{t})$ (Faghri et al., 2018):

$$\ell(i,t)=\sum_{\hat{t}}[\alpha-s(i,t)+s(i,\hat{t})]_{+}+\sum_{\hat{i}}[\alpha-s(i,t)+s(\hat{i},t)]_{+}, \tag{1}$$

where α is the similarity margin and generally takes a value of 0.2, $s(i,t)$ is the similarity between image and text, and $[x]_{+}$ takes the larger value between 0 and x. However, when the positive instance is unaligned, the model will still maximize the similarity incorrectly. Furthermore, cross-modal retrieval can also be reached by image–text matching (ITM), which concatenates the input of image and text to a transformer-based model and performs the binary classification using the classification [CLS] token (Lu et al., 2019; Chen YC et al., 2020; Li XJ et al., 2020). Despite the fact that this type of method has strong interaction capabilities, incorrect labels still degrade the model performance.

In contrast to image classification with noisy labels (Lin XY et al., 2021), we concentrate on cross-modal retrieval with noise-aligned multi-modal data, which takes mismatched multi-modal instance pairs into account rather than incorrectly labeled images. Note that many noisy label methods cannot be applied to the noise-alignment problem directly because these methods study class-level noise rather than instance-level noise in multi-modal data. However, there are still some methods that can be used, for example, sample selection (Han et al., 2018; Jiang et al., 2018) and label correction (Reed et al., 2015; Arazo et al., 2019). To make full use of noise-alignment pairs, we apply the method of refurbishing labels. To make this practicable, we adopt ITM instead of contrastive learning to train the cross-modal retrieval model, because changing the binary alignment label is not affected by the batch size and is easy to reach compared with finding an aligned text (image) to the image (text). Inspired by Arazo et al. (2019), we fit the ITM loss to a two-component beta-mixture model (BMM) to separate the cross-modal samples into clean and noisy samples. However, directly solving the noise-alignment problem with this method is not practical. According to our observations of ITM loss, noise-alignment data are quickly fitted due to the strong fitting ability of transformer-based models, in contrast to the slow decline in noise-labeled image classification loss. Consequently, noise-alignment instances have higher loss only during a narrow time window at the beginning, which results in lack of adequate time and makes it difficult to distinguish clean and noise alignments from the loss distribution. Therefore, it is necessary to slow down the model's fitting to the noise alignment, which can result in a larger time window for modeling a well-categorized BMM. We discover that learning with ITM and masked language modeling (MLM) makes it possible. On one hand, MLM is self-supervised and no additional noise is brought in. On the other hand, multi-task learning (MTL) consisting of these two tasks reduces the risk of overfitting on the single task of ITM as a regularization method (Ruder, 2017).

To summarize, the contributions of this paper are as follows:

1. From a practical standpoint, we divide the noise-alignment problem into two categories, partial noise alignment and complete noise alignment, based on whether the noise-alignment modality contains the same semantics.

2. We present a robust cross-modal retrieval method, RCAR, which combines the noise correction theory with MTL.

3. We construct these two types of noise on two datasets, i.e., Microsoft Common Objects in Context (MS-COCO) and Flickr30K. We test our method and prove its robustness. Compared with popular methods, RCAR reaches the best retrieval efficiency.

2 Related works

2.1 Image classification with label noise

Image classification with noisy labels is a significant task in the field of computer vision, referring to the classification under noise supervision. Existing strategies, such as sample loss reweighting (Liu and Tao, 2016; Wang RX et al., 2018; Zhang et al., 2021), label refurbishing (Reed et al., 2015; Ma XJ et al., 2018; Arazo et al., 2019), and robust learning (Manwani and Sastry, 2013; Ghosh et al., 2017; Ma X et al., 2020), have been investigated from various perspectives to reduce the impact of noise on the model. Sample loss reweighting (Liu and Tao, 2016) defines the sample importance weight as the quotient of the joint probability of the true and false distributions, with the correct sample having the larger weight value. The "Active Bias" (Chang et al., 2017) method assumes that the prediction variance reflects the degree of inconsistency and sample difficulty and weights the loss accordingly.

In contrast to sample loss reweighting, label refurbishment attempts to avoid overfitting to incorrect labels by refurbishing a noisy label. Deep neuron network (DNN) prediction is used to update the labels (Song et al., 2020). These methods, in some ways, enable the model to build self-confidence and robustness. The first way to implement this idea is bootstrapping. Reed et al. (2015) established a bootstrapping method that uses the label confidence discovered during cross-validation to update the target label of training data. Dynamic bootstrapping (Arazo et al., 2019) uses the expectation-maximization (EM) algorithm to evaluate the likelihood of a sample being cleanly labeled dynamically. SELFIE (Song et al., 2019) corrects the high-confidence training sample by substituting the label with network prediction.

The purpose of the robust loss function is to provide loss functions that keep the risk of unseen test data low even when the data are noisy. Manwani and Sastry (2013) investigated the noise tolerance property of risk minimization (under various loss functions), theorized a sufficient condition for the loss function, and made the risk minimization of this function a noise tolerance for binary classification. The robust mean absolute error (MAE) (Ghosh et al., 2017) model, on the other hand, demonstrates that the MAE loss shows a better generalization since it satisfies the aforementioned requirement. The curriculum loss (CL) model in Lyu and Tsang (2020) shows that 0-1 loss offers some robustness; however, optimization is challenging. Hence, they proposed a very straightforward and effective loss. Additionally, it is demonstrated that CL provides a tighter upper bound for the 0-1 loss than the typical alternative loss based on summation. Rather than using a predetermined threshold or calculation to do curriculum learning, MentorNet (Jiang et al., 2018) applies a data-driven strategy. However, MentorNet is a self-training system that tends to accumulate errors. All these methods focus on image classification with noisy labels and cannot directly be applied in robust cross-modal learning because of modal heterogeneity.

2.2 Cross-modal retrieval

Cross-modal retrieval is the process of finding a common representation space for various modalities so that they can retrieve each other. The most important problem that needs to be solved is modal heterogeneity. For modal retrieval strategies, there are two approaches (Geigle et al., 2022). The first approach involves early interaction methods (Jia et al., 2021; Radford et al., 2021). This kind of method maps image regions and text words to the same dimension before concatenating the input to the transformer and then performs the binary classification task using the [CLS] token. Cross-modal retrieval methods are usually used to train several large-scale multi-modal pre-training models (Yang et al., 2022). The reason is that these are simple in principle, fast to train, and treat image regions and text words as equal tokens that can be fully interacted with inter-modal features while also fully interacting with intra-modal features, which is more beneficial to reducing inter-modal heterogeneity. The second approach is late interaction methods, e.g., visual-semantic embedding (VSE++) (Faghri et al., 2018), stacked cross attention network (SCAN) (Lee et al., 2018), and similarity graph reasoning and attention filtration (SGRAF) (Diao et al., 2021), which encode the modalities individually, project them into a shared latent semantic space, and then compute the similarity between the projected points for contrastive learning. According to the features used, this technique can be divided into two types. The first category mines the hardest negative for targeted training using the global features of the modal data (Faghri et al., 2018), with the image's global features retrieved using ResNet (He et al., 2016) and the text's global features extracted using gate recurrent unit (GRU) (Chung et al., 2014). The second category (Lee et al., 2018; Li KP et al., 2019; Chen H et al., 2020; Diao et al., 2021; Messina et al., 2021) uses local features of modal data, with the image's modal local features extracted using bottom-up attention and the text's modal local features extracted using GRU or BERT (Devlin et al., 2019). The most significant distinction between these methods is the method of calculating the image–text similarity. Lee et al. (2018) used stacked cross attention to find potential alignment between regions and words and thereby to infer image–text global similarity. Li KP et al. (2019) pointed out that simply using the features of image region lacks the semantic concept of the scene, and that directly calculating the image–text similarity is not the best option; they proposed the use of a graph convolutional neural network to infer the image region's relations, generating the region's features with a semantic concept of the scene. In fact, semantics can be complicated, such as shallow and confusing. Chen H et al. (2020) computed the image–text similarity using an iterative matching strategy to achieve semantic alignment for mining various semantic complexities. Diao et al. (2021) used the graph convolutional neural network to obtain the similarity. However, these methods are trained with clean image–text pairs and generate bad results under noise-alignment supervision.

3 Proposed method

Cross-modal retrieval can be formulated as the problem of learning a model $f(I,T)$ to predict the similarity of image I and text T from a set of multi-modal training instances $D=\{(I_i,T_i,y_i)\}_{i=1}^N$ with $y_i\in\{0,1\}$ being the binary ground-truth label that indicates whether the image–text pair (I_i,T_i) is aligned (1) or not (0). For the noise-alignment problem, it is defined that some image–text pairs (I_j,T_j) cannot be identified in the training data, which are unaligned but are labeled as positive incorrectly.

3.1 Model pipeline

As illustrated in Fig. 2, RCAR contains an image encoder, a text encoder, a single-stream transformer as a cross-modal encoder, and an alignment refurbisher. In this way, an input image I and input text T can be encoded into two sequences of embeddings $\{v_1,v_2,\cdots,v_O\}$ and $\{w_1,w_2,\cdots,w_L\}$, where O is the number of detected image regions and L is the length of the sentence. As the input of the cross-modal encoder, we concatenate the image and text embeddings into one sequence $\{[\mathrm{CLS}],v_1,v_2,\cdots,v_O,w_1,w_2,\cdots,w_L\}$. At the start of training, MTL is used with ITM and MLM to prevent the model from overfitting the noisy data. Then, ITM is conducted to do cross-modal retrieval. The refurbisher starts working after the warm-up period and it trains the network for m epochs.

3.2 Training objectives

3.2.1 ITM process

We use ITM as shown in Eq. (2) to predict whether a pair of image and text is aligned or not. Then, we make a binary classification according to the [CLS] token.

$$\ell_{\mathrm{ITM}}(I,T)=-\sum_{i=1}^{N}y_i^{\mathrm{T}}\log p(I_i,T_i), \tag{2}$$

where $p(I_i,T_i)$ denotes the binary softmax probability of the i^{th} pair.

3.2.2 MLM process

In addition to ITM, we apply MLM to motivate MTL. The input words are randomly masked off with a 15% probability and the masked ones are replaced

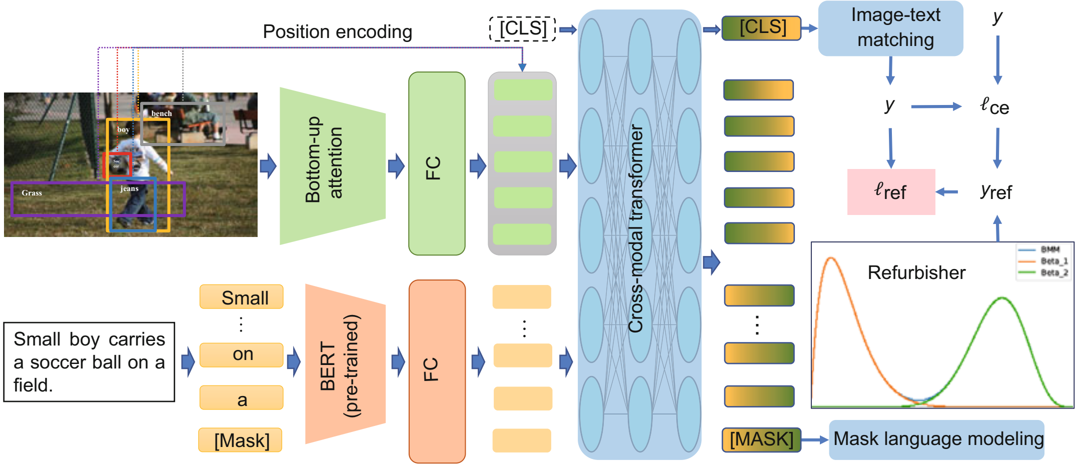


Fig. 2 Illustration of robust cross-modal retrieval with alignment refurbishment (RCAR). RCAR learns a robust cross-modal retrieval model by combining a modal alignment refurbisher with multi-task learning. Image–text matching (ITM) and masked language modeling (MLM) are used to motivate multi-task learning to alleviate overfitting to the noise. To make full use of noisy data, the refurbisher is used to correct the noise-alignment label y_i. FC: fully connected

with a special token [MASK]. The objective is to minimize the negative log-likelihood of these masked words by observing their context words $w_{\backslash m}$ and all image regions v:

$$\ell_{\text{MLM}}(\theta)=-\mathbb{E}_{(w,v)\sim D}\log P_\theta(w_m\mid w_{\backslash m},v),\tag{3}$$

where θ represents the trainable parameter.

3.3 Alignment refurbisher

For noise-alignment correction, we introduce an alignment refurbisher which builds a mixture distribution model. Although the Gaussian mixture model (GMM) is the most widely used, its performance in approximating the loss distribution of a mixture of clean and noisy samples is worse than that of BMM (Arazo et al., 2019), because BMM can model both symmetric and skewed distributions ranging in $[0,1]$ (Ma ZY and Leijon, 2011). By modeling the normalized ITM loss of the image–text pairs, the refurbisher fits a two-component BMM that can be defined as follows:

$$p(\ell)=\sum_{k=1}^{2}\mu_k f(\ell\mid\alpha_k,\beta_k),\tag{4}$$

where μ_k is the mixing coefficient of the k^{th} mixture component and $f(\ell\mid\alpha_k,\beta_k)$ is the probability density function of the k^{th} beta distribution:

$$f(\ell\mid\alpha_k,\beta_k)=\frac{\Gamma(\alpha_k+\beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)}\ell^{\alpha_k-1}(1-\ell)^{\beta_k-1},\tag{5}$$

where $\Gamma(\cdot)$ is the gamma function and $\alpha_k,\beta_k>0$.

To fit BMM to the ITM loss, we apply an EM algorithm. We define latent variable $\lambda_k(\ell)=p(k\mid\ell)$, which represents the posterior probability of the value ℓ being originated by mixture component k. In the expectation-step (E-step), the Bayes rule is used to update the latent variables $\lambda_k(\ell)$ with the other parameters μ_k,α_k, and β_k being fixed:

$$\lambda_k(\ell)=\frac{\mu_k f\left(\ell\mid\alpha_k,\beta_k\right)}{\sum_{k=1}^{K}\mu_k f\left(\ell\mid\alpha_k\beta_k\right)}.\tag{6}$$

After the E-step, we fix $\lambda_k(\ell)$ and use a weighted version of the method of moments to estimate the distribution parameters α_k,β_k:

$$\alpha_k=\bar{\ell}_k\left(\frac{\bar{\ell}_k\left(1-\bar{\ell}_k\right)}{\hat{\sigma}_k^2}-1\right),\beta_k=\frac{\alpha_k\left(1-\bar{\ell}_k\right)}{\bar{\ell}_k},\tag{7}$$

where $\bar{\ell}_k$ represents a weighted average of the losses $\{\ell_i\}_{i=1}^N$ of each training sample $\{I_i,T_i\}_{i=1}^N$ and $\hat{\sigma}_k^2$ represents the weighted variance estimate:

$$\bar{\ell}_k=\frac{\sum_{i=1}^N\lambda_k\left(\ell_i\right)\ell_i}{\sum_{i=1}^N\lambda_k\left(\ell_i\right)},\hat{\sigma}_k^2=\frac{\sum_{i=1}^N\lambda_k\left(\ell_i\right)\left(\ell_i-\bar{\ell}_k\right)^2}{\sum_{i=1}^N\lambda_k\left(\ell_i\right)}.\tag{8}$$

Then the updated mixing coefficients μ_k's can be calculated in the following way:

$$\mu_k=\frac{1}{N}\sum_{i=1}^N\lambda_k\left(\ell_i\right).\tag{9}$$

Finally, we can estimate the probability that the image–text pair is noise-aligned by calculating the posterior probability:

$$\lambda_t(\ell_i) = p\left(t \mid \ell_i\right) = \frac{p(t)p\left(\ell_i \mid t\right)}{p\left(\ell_i\right)}, \tag{10}$$

where t indicates the noise-alignment class, which is the beta component with a larger mean value.

We refurbish only the positive instance because the negative instance is manually constructed and clean. With the computation above, the alignment label y_i can be refurbished in the following manner:

$$y_{i_{\text{ref}}} = H((1-\lambda_t(\ell_i))\, y_i + \lambda_t(\ell_i) z_i), \tag{11}$$

where z_i is the one-hot class prediction and H uses the class with the highest probability after weighted summation as a hard label. The loss after alignment refurbishment can be denoted as follows:

$$\ell_{\text{ref}} = -\sum_{i=1}^{N} y_{i_{\text{ref}}}^{\mathrm{T}} \log\left(p(I_i, T_i)\right). \tag{12}$$

4 Experiments

4.1 Experimental settings

4.1.1 Noise-alignment type

From a practical standpoint, we propose two types of noise alignment with different proportions. The first type is partial noise alignment, which means that the image and text have matched semantics partially as shown in Fig. 1b. It is constructed by calculating the Jaccard similarity, as shown in Eq. (13), of the objects between different positive pairs, which measures the similarity between two sets of classes (Niwattanakul et al., 2013):

$$\mathcal{J}(A, B) = |A \cap B| / |A \cup B|. \tag{13}$$

Then, we replace the image or text randomly according to the similarity matrix. The second type is complete noise alignment, which means that the image and text are totally mismatched in terms of semantics, as shown in Fig. 1c, and this is constructed by replacing the captions of the images randomly.

4.1.2 Data sources

We construct complete noise alignment on two public datasets, i.e., MS-COCO (Lin TY et al., 2014) and Flickr30K (Huiskes and Lew, 2008), and adopt partial noise alignment on only MS-COCO because the image–text pairs in MS-COCO have class information in the form of 80-dimensional one-hot vectors. For each type of noise, we validate our method's robustness at four different noise ratios, i.e., 0%, 20%, 40%, and 60%, and report the results of other experiments at the 40% noise ratio. For the original dataset, MS-COCO contains 123 287 images and five captions for each image. Flickr30K consists of 31 000 images collected from the Flickr website, and here also each image is associated with five captions. We follow the split in Karpathy and Li (2015).

4.1.3 Evaluation metrics

We use the recall at K ($R@K$), which is defined as the fraction of queries for the correctly retrieved item among the closest K points to the query to measure the performance of image retrieval and text retrieval.

4.1.4 Implementation details

The entire network is trained on a TITAN RTX GPU. Following the method of Messina et al. (2021), we adopt faster regions with convolutional neural networks (Faster R-CNN) (Ren et al., 2017) as the image encoder and a pre-trained BERT (Devlin et al., 2019) as the text encoder, to extract local features. An eight-layer transformer is used with eight heads per layer. We train RCAR with MTL for 10 epochs and with ITM for 20 epochs. The model is warmed-up for seven epochs. The batch size is set to 64. We use the Adam (Kingma and Ba, 2015) optimizer with a learning rate initialized by 3×10^{-5} and use the cosine annealing strategy to update parameters.

4.2 Retrieval results on noisy cross-modal datasets

We provide the results of representative models, including VSE, VSE++, visual semantic reasoning network (VSRN), transformer encoder reasoning and alignment network (TERAN), SCAN, iterative matching with recurrent attention memory (IMRAM), and SGRAF. These methods represent four distinct technical paths: (1) global-feature-based methods: VSE, VSE++; (2) transformer-based model: TERAN; (3) local-feature-based methods without inter-modal attention: VSRN; (4) local-feature-based methods with inter-modal attention: SCAN, IMRAM, and SGRAF. Tables 1 and 2 present the quantitative results of comparison between these methods on two datasets with different ratios of noise-alignment data.

Table 1 Comparison of performance of RCAR with state-of-the-art methods in the context of partial noise-alignment data (part) on the MS-COCO dataset

Noise	Method	Image2Text			Text2Image		
		R@1	R@5	R@10	R@1	R@5	R@10
0%	VSE	42.0	76.3	86.3	32.0	67.3	80.4
	VSE++	64.6	90.0	95.7	52.0	84.3	92.0
	VSRN	76.2	94.8	98.2	62.8	89.7	95.1
	TERAN	77.7	95.9	98.6	**65.0**	91.2	96.4
	SCAN	70.4	94.1	**98.9**	56.8	87.5	94.0
	IMRAM	72.3	94.6	98.3	60.6	88.8	95.0
	SGRAF	**78.3**	**96.2**	98.7	62.8	90.3	95.7
	RCAR	74.3	95.8	98.3	61.7	**91.4**	**96.8**
20%	VSE	34.5	70.1	83.0	27.4	62.1	76.8
	VSE++	33.2	68.4	81.0	27.2	61.9	76.7
	VSRN	66.1	90.9	96.1	55.7	86.0	92.7
	TERAN	69.1	91.9	96.5	58.2	85.8	89.1
	SCAN	66.4	91.1	95.8	51.6	83.9	82.2
	IMRAM	69.1	93.2	97.1	56.1	86.0	93.2
	SGRAF	68.3	93.1	96.0	56.1	86.5	92.9
	RCAR	**71.0**	**93.7**	**97.2**	**58.6**	**88.4**	**95.6**
40%	VSE	31.6	64.9	79.0	24.5	57.7	72.6
	VSE++	33.3	62.8	77.0	24.1	56.8	71.3
	VSRN	58.2	86.3	93.0	45.6	77.6	86.5
	TERAN	17.2	65.9	73.1	14.9	33.1	43.5
	SCAN	62.6	91.5	96.6	49.2	82.0	90.6
	IMRAM	66.2	91.1	96.7	51.9	84.7	92.1
	SGRAF	67.8	92.4	96.4	52.7	80.5	83.5
	RCAR	**68.5**	**93.2**	**96.9**	**55.6**	**85.9**	**94.6**
60%	VSE	24.5	57.5	73.9	19.2	51.5	68.1
	VSE++	22.9	49.7	65.4	18.9	49.4	64.8
	VSRN	32.6	59.3	68.5	23.3	48.2	60.1
	TERAN	0.1	0.5	1.0	0.1	0.5	1.0
	SCAN	35.2	61.4	72.7	25.9	49.6	59.6
	IMRAM	42.3	78.5	89.6	40.2	73.4	86.3
	SGRAF	42.9	69.7	85.1	37.8	63.4	71.8
	RCAR	**49.5**	**82.3**	**90.7**	**42.8**	**78.9**	**88.3**

Evaluation criterion is $R@K$. The best results are in bold

The experiments reveal the following: (1) Complete noise alignment is more harmful for models to learn cross-modal consistency than partial noise alignment because models can still learn the object information in partially noisy data. (2) Hard negative mining (VSE++) has poor robustness compared with the traditional loss function (VSE) model because the hardest negative is likely to be a positive instance for noise-aligned data. (3) Using intra-modal attention to optimize modal features, i.e., VSRN, has little effect on robustness improvement because the cross-modal attention mechanism is not optimized. On the contrary, using cross-modal attention to compute image–text similarity, i.e., SCAN and SGRAF, can increase model robustness. The reason is that the model focuses attention on the aligned regions and reduces the learning of non-aligned regions. However, performance drops significantly on 60% complete noise. (4) Transformer-based model, i.e., TERAN, has bad performance because it overfits the noise alignment easily due to its excellent fitting ability. (5) Traditional methods have some robustness because some of them still have a good performance in the context of 20% complete noise and all of them suffer from a "cliff-like drop" in the context of 60% complete noise. The reason is that these methods cannot learn a good semantic common space of those two modals on high-ratio noise. (6) RCAR is more robust because it reduces overfitting to the noise alignment and can still learn correct knowledge from the refurbished noisy instances.

4.3 Ablation study

Table 3 provides the results of ablation studies. To explore the effect of MTL and the refurbisher, we validate our approach by revisiting each term in Flickr30K with 40% complete noise alignment. The results reveal the following: (1) Baseline, i.e., single-stream transformer with ITM, has a little worse performance than SCAN. (2) Both MTL and the refurbisher contribute to model robustness, and RCAR acquires better improvements by considering both of them. For example, the improvements of Image2Text and Text2Image are 21.9 and 13.6 respectively in terms of the R@1 score.

4.4 Sensitivity to parameters

To explore the influence of the warm-up epochs after which the refurbisher begins to work, i.e., the parameter m, we tune m in $\{6, 7, 8, 9, 10, 11\}$ and show their performance in Fig. 3. We find that the retrieval results are the best when $m = 7$, because the model is affected by the noisy sample when m is large, while the losses are not separated because of the underfitting of the clean sample when m is small.

4.5 Computation time

We record the computation time of representative methods (i.e., VSE++, VSRN, SCAN, TERAN, SCAN, IMRAM, SGRAF, and RCAR). The results in Table 4 reveal the following: (1) The global-feature-based model, i.e., VSE++, has

Table 2 Comparison of performance of RCAR with state-of-the-art methods in the context of complete noise-alignment data (cmp) on the Flickr30K and MS-COCO datasets

Noise	Method	Flickr30K						MS-COCO*					
		Image2Text			Text2Image			Image2Text			Text2Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
0%	VSE	27.8	56.9	68.9	20.1	45.8	57.1	42.0	76.3	86.3	32.0	67.3	80.4
	VSE++	52.9	80.5	87.2	39.6	70.1	79.5	64.6	90.0	95.7	52.0	84.3	92.0
	VSRN	71.3	90.6	96.0	54.7	81.8	88.2	76.2	94.8	98.2	62.8	89.7	95.1
	TERAN	71.8	90.5	94.7	55.7	83.1	89.3	77.7	95.9	98.6	**65.0**	91.2	96.4
	SCAN	67.4	90.3	95.8	48.6	77.7	85.2	70.4	94.1	**98.9**	56.8	87.5	94.0
	IMRAM	72.0	92.1	96.8	53.5	80.4	87.4	72.3	94.6	98.3	60.6	88.8	95.0
	SGRAF	**76.2**	**93.4**	**97.3**	**57.5**	82.4	88.8	**78.3**	**96.2**	98.7	62.8	90.3	95.7
	RCAR	71.3	90.5	95.3	54.6	**83.7**	**89.5**	74.3	95.8	98.3	61.7	**91.4**	**96.8**
20%	VSE	16.1	34.2	43.3	10.6	27.6	36.2	31.1	64.7	78.4	23.3	56.4	69.6
	VSE++	15.3	34.5	43.0	9.8	27.5	36.3	25.1	51.0	60.4	20.1	45.2	54.5
	VSRN	50.0	76.4	84.0	33.9	62.0	71.8	50.3	80.5	90.4	42.0	77.9	87.7
	TERAN	35.1	51.1	54.8	27.4	43.9	46.8	69.8	92.1	94.8	55.5	86.1	91.9
	SCAN	57.5	84.3	91.4	39.8	67.7	76.6	65.1	90.5	95.3	47.4	80.5	90.3
	IMRAM	56.0	85.3	92.1	40.1	68.7	76.3	68.9	92.5	96.5	56.6	86.0	92.0
	SGRAF	52.8	84.7	92.1	42.9	70.4	76.7	67.4	92.8	95.3	54.4	86.2	92.8
	RCAR	**66.5**	**88.2**	**92.4**	**50.6**	**79.9**	**87.6**	**70.4**	**93.1**	**97.5**	**57.8**	**88.8**	**95.5**
40%	VSE	8.2	19.3	25.6	6.1	14.9	20.3	19.0	48.1	61.5	15.0	43.1	58.4
	VSE++	7.9	26.4	39.5	6.7	16.5	21.4	15.8	29.4	35.3	14.6	44.8	56.7
	VSRN	31.8	58.0	69.0	21.9	44.1	54.8	28.3	59.2	71.0	26.2	58.2	71.3
	TERAN	0.4	3.2	4.9	2.0	7.6	13.2	37.7	63.4	70.9	30.3	58.3	70.2
	SCAN	37.7	65.8	74.9	24.7	50.0	60.0	59.2	89.4	94.8	48.8	78.9	86.0
	IMRAM	36.8	62.4	73.6	23.4	46.0	52.0	57.7	89.2	94.4	42.6	76.0	81.7
	SGRAF	26.1	56.5	68.9	20.9	46.2	58.0	38.1	72.0	83.5	28.5	61.0	74.8
	RCAR	**61.0**	**85.0**	**90.6**	**44.5**	**75.2**	**83.6**	**64.7**	**90.9**	**95.8**	**52.9**	**85.8**	**93.8**
60%	VSE	3.5	9.3	12.3	2.3	7.1	10.4	7.0	18.3	23.5	5.9	16.1	20.5
	VSE++	2.7	8.9	12.6	1.3	3.8	5.7	6.5	15.8	21.3	5.2	15.2	20.9
	VSRN	11.1	27.3	37.9	7.7	20.1	26.6	9.1	25.6	36.7	8.7	26.8	37.7
	TERAN	0.1	0.6	1.0	0.1	0.6	1.1	7.6	11.4	13.0	6.3	11.7	14.5
	SCAN	21.1	58.2	66.9	10.1	26.1	33.4	28.7	59.3	72.6	15.4	34.2	46.5
	IMRAM	7.0	22.0	31.1	5.0	14.2	20.5	30.0	63.6	74.4	23.0	46.8	51.7
	SGRAF	20.1	54.2	65.1	13.6	32.2	45.5	23.4	53.6	66.2	17.1	44.7	58.6
	RCAR	**51.2**	**81.0**	**89.9**	**38.7**	**69.8**	**78.9**	**55.3**	**83.9**	**91.6**	**44.3**	**79.7**	**90.5**

* MS-COCO's 1k testing set is used. Evaluation criterion is $R@K$. The best results are in bold

fewer parameters and shorter computation time compared with local-feature-based and transformer-based models, i.e., VSRN, SCAN, TERAN, SCAN, IMRAM, SGRAF, and RCAR. The global-feature-based method cannot fit the training data well, which leads to the fact that the model does not achieve the best performance on clean data and also does not achieve the worst performance on data with high percentage of noise. (2) RCAR has more parameters because RCAR uses the BERT-based model as the text feature extractor, which has 109M parameters. (3) RCAR has the longest inference time, because RCAR uses the pre-interaction method and needs to concatenate different image–text pairs and input them into the transformer layer when calculating the similarity, which increases the inference time. However, the training time of RCAR is the shortest among the local-feature-based methods. Because RCAR uses the pre-trained BERT-based model for parameter initialization and a robust strategy for label correction, the training time is significantly reduced. For example, VSRN and RCAR have a similar size of parameters, but the training of VSRN takes 25.40 h, while RCAR takes only 7.50 h, which indicates that RCAR can converge faster.

4.6 Visualization and analysis

To illustrate the effect of MTL, we draw the boxplots shown in Fig. 4, which demonstrates the distribution of 90% ITM loss of clean and noisy instances over the first 15 epochs. The remaining 10% loss

Table 3 Ablation study in the context of 40% complete noise-alignment data on the Flickr30K dataset

Method	Image2Text			Text2Image		
	R@1	R@5	R@10	R@1	R@5	R@10
SCAN	43.3	72.8	80.9	33.0	59.7	69.2
Baseline	39.1	69.7	79.3	30.9	62.4	73.0
w/o MTL	46.1	75.0	84.8	36.0	67.2	78.4
w/o ref	46.3	74.3	83.4	34.9	64.1	73.1
RCAR	**61.0**	**85.0**	**90.6**	**44.5**	**75.2**	**83.6**

Evaluation criterion is $R@K$. w/o: without. MTL: multi-task learning. The best results are in bold

Table 4 Comparison of methods in the context of model size and computation time for 40% complete noise-alignment data on the Flickr30K dataset

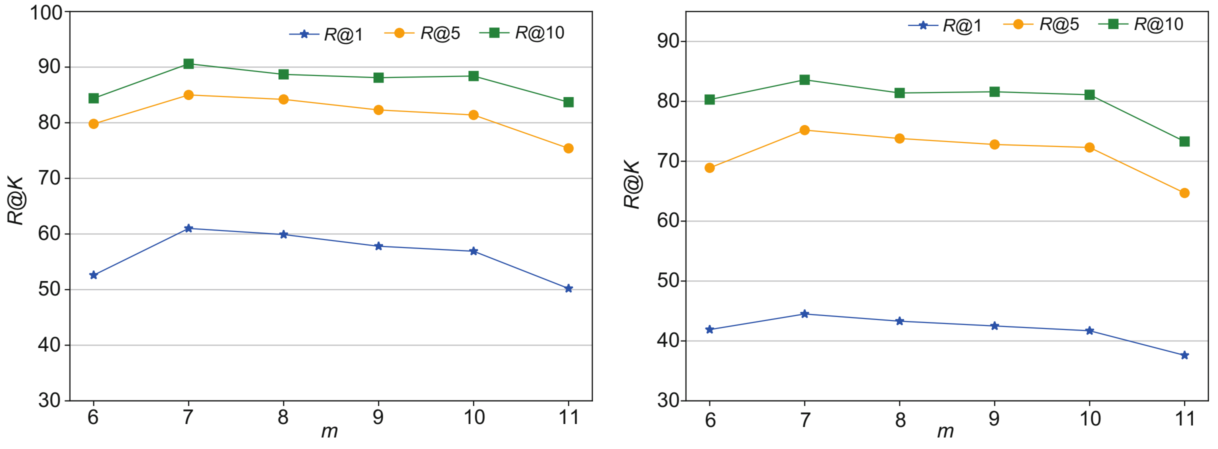
Method	Model size	Training time (h)	Inference time (h)	Total time (h)
VSE++	10.8M	2.40	0.03	2.43
VSRN	137.6M	25.40	0.08	25.48
TERAN	215.4M	18.60	0.10	18.70
SCAN	12.7M	8.20	0.06	8.26
IMRAM	17.1M	9.00	0.13	9.13
SGRAF	18.1M	10.30	0.15	10.45
RCAR	136.3M	7.50	0.40	7.90

data that are too large or too small are regarded as outliers. When the two distributions do not overlap, the data become more divisible. From observation, MTL creates a larger time window (4–14 epochs) for data separation.

Meanwhile, as shown in Fig. 5a, we visualize the effect of the refurbisher. By fitting the sample losses to a beta mixture distribution, we can find the following: (1) The loss of most noisy instances is larger than the loss of the clean instances. (2) The sample losses are clustered into two classes, with the small mean value being the clean cluster (blue curve) and the larger mean value being the noisy cluster (gray curve).

A t-distributed stochastic neighbor embedding (t-SNE) figure is often used to visualize the data distribution by the downscaling technique (van der Maaten and Hinton, 2008), and we demonstrate the distribution of training data, as shown in Fig. 5b. Note that to use the large amount of image–text multi-modal data with noise (i.e., data from the Web), the influence of noise-aligned image–text pairs must be reduced. In other words, in the noise cross-modal retrieval task, the term "noisy data" refers to negative samples that are incorrectly marked as positive. We construct these data by randomly replacing the aligned text (or image) with an incorrectly aligned text (or image). Therefore, noisy data are negative samples in fact. Figs. 5b and 5c demonstrate the data distribution after dimensionality reduction by the t-SNE method, revealing the following: (1) For the SCAN method, most of the noisy samples and positive samples are clustered into one class, which shows that SCAN overfits the noisy data and has poor robustness. (2) For our RCAR method, a large amount of noisy data and a large number of negative samples are clustered into one class, which illustrates that our model does not overfit the noisy data in the end, demonstrating the robustness of our model.

Fig. 6 illustrates the qualitative results of text retrieval for the given image queries. Most of the



(a)

(b)

Fig. 3 Parameter sensitivity of m in Image2Text (a) and Text2Image (b)

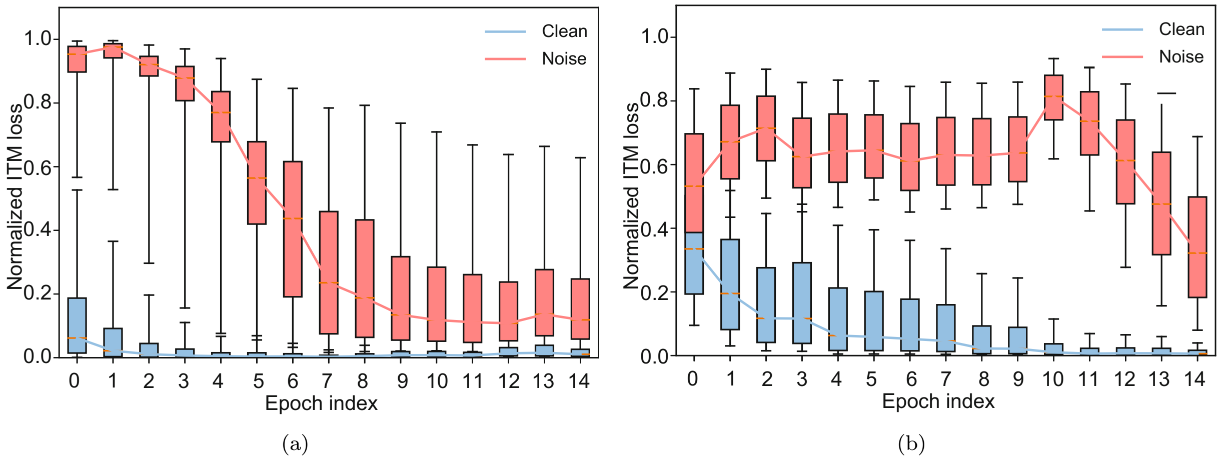


Fig. 4 Visualization of the effect of multi-task learning (MTL): (a) without MTL; (b) with MTL (the refurbisher is not involved)

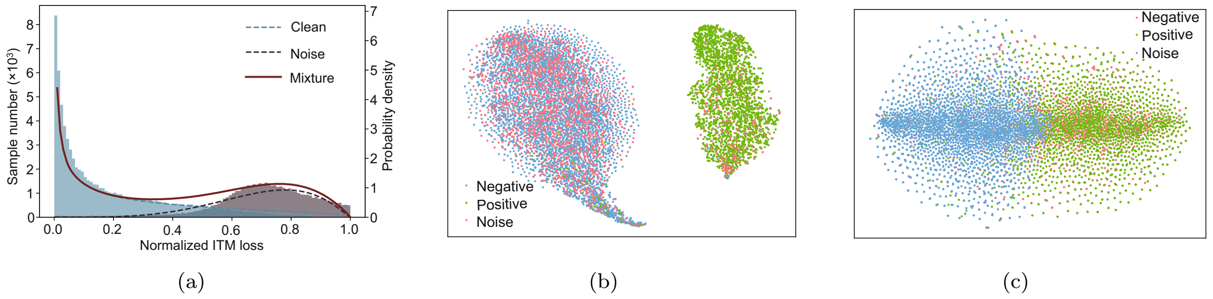
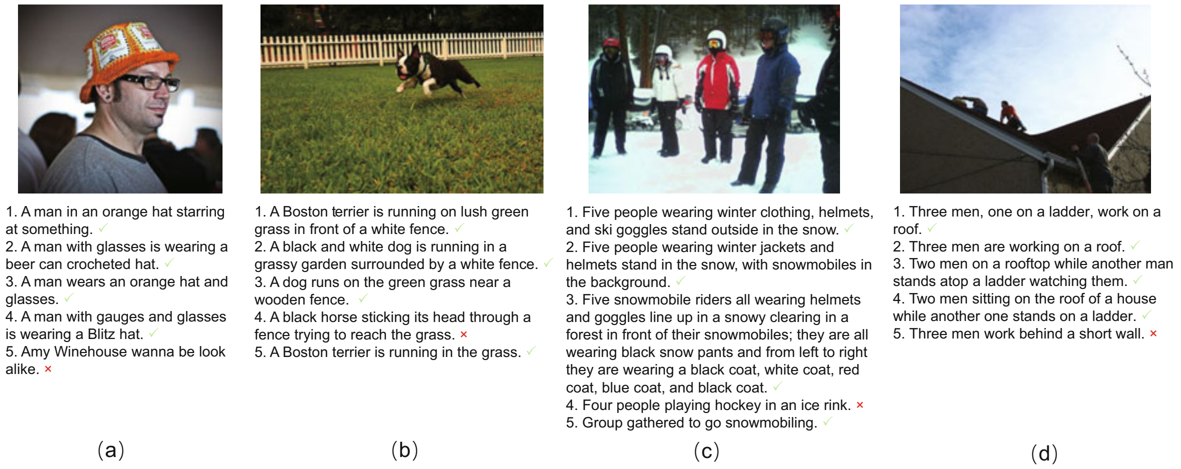


Fig. 5 Visualization of the refurbisher's effect (a), t-SNE result of RCAR (b), and t-SNE result of SCAN (c). In (a), the x-axis is the normalized loss values. The left scale of the y-axis is the sample number of the loss values in different intervals corresponding to the histogram and the right scale is the probability density for the given loss values corresponding to the three curves. In (b), most of the noise-alignment data are clustered into the negative category. In (c), SCAN overfits the noise, and most of the noise-alignment data are clustered into the positive category. References to color refer to the online version of this figure



1. A man in an orange hat starring at something. ✓
2. A man with glasses is wearing a beer can crocheted hat. ✓
3. A man wears an orange hat and glasses. ✓
4. A man with gauges and glasses is wearing a Blitz hat. ✓
5. Amy Winehouse wanna be look alike. ×

(a)

1. A Boston terrier is running on lush green grass in front of a white fence. ✓
2. A black and white dog is running in a grassy garden surrounded by a white fence. ✓
3. A dog runs on the green grass near a wooden fence. ✓
4. A black horse sticking its head through a fence trying to reach the grass. ×
5. A Boston terrier is running in the grass. ✓

(b)

1. Five people wearing winter clothing, helmets, and ski goggles stand outside in the snow. ✓
2. Five people wearing winter jackets and helmets stand in the snow, with snowmobiles in the background. ✓
3. Five snowmobile riders all wearing helmets and goggles line up in a snowy clearing in a forest in front of their snowmobiles; they are all wearing black snow pants and from left to right they are wearing a black coat, white coat, red coat, blue coat, and black coat. ✓
4. Four people playing hockey in an ice rink. ×
5. Group gathered to go snowmobiling. ✓

(c)

1. Three men, one on a ladder, work on a roof. ✓
2. Three men are working on a roof. ✓
3. Two men on a rooftop while another man stands atop a ladder watching them. ✓
4. Two men sitting on the roof of a house while another one stands on a ladder. ✓
5. Three men work behind a short wall. ×

(d)

Fig. 6 Qualitative results of text retrieval for the given image queries. For each image query, we show the top-five ranked sentences (or expressions) in (a)–(d). We observe that our RCAR retrieves the correct results in the top-ranked sentences. References to color refer to the online version of this figure

retrieved sentences are correct (shown as tick). Some outputs are mismatched (shown as fork), but reasonable, for example, 4 in Fig. 6b and 4 in Fig. 6c contain similar semantic meaning to the image. On the other hand, there are semantically incorrect outputs such as 5 in Fig. 6a, possibly due to the influence of noise-alignment data. Fig. 7 shows the qualitative results of image retrieval for the given sentence queries. Each sentence corresponds to a ground-truth image. For each sentence query, we display the top-three retrieved images, ranking from left to right. As indicated in these examples, our model retrieves the ground-truth image successfully and other top-ranking results are also reasonable.

5 Conclusions

This paper presented the RCAR method for robust cross-modal retrieval with noise alignment. It combines the noise classification theory with MTL, increasing the model's robustness by adaptively refurbishing the label of the noise-alignment data in cross-modal learning. Experimental results showed that RCAR has better performance than the current popular methods on two types of noise-alignment data.

Contributors

Jinyi GUO and Jieyu DING designed the research. Jinyi GUO processed the data and drafted the paper. Jieyu DING helped organize the paper. Jinyi GUO and Jieyu DING revised and finalized the paper.

Compliance with ethics guidelines

Jinyi GUO and Jieyu DING declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Arazo E, Ortego D, Albert P, et al., 2019. Unsupervised label noise modeling and loss correction. Proc 36th Int Conf on Machine Learning, p.312-321.

Chang HS, Learned-Miller E, McCallum A, 2017. Active bias: training more accurate neural networks by emphasizing high variance samples. Proc 31st Int Conf on Neural Information Processing Systems, p.1003-1013.

Chen H, Ding GG, Liu XD, et al., 2020. IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12652-12660. https://doi.org/10.1109/CVPR42600.2020.01267

Chen YC, Li LJ, Yu LC, et al., 2020. UNITER: universal image-text representation learning. Proc 16th European Conf on Computer Vision, p.104-120. https://doi.org/10.1007/978-3-030-58577-8_7

Chung J, Gulcehre C, Cho KH, et al., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. https://arxiv.org/abs/1412.3555

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Proc Conf of the North American Chapter of

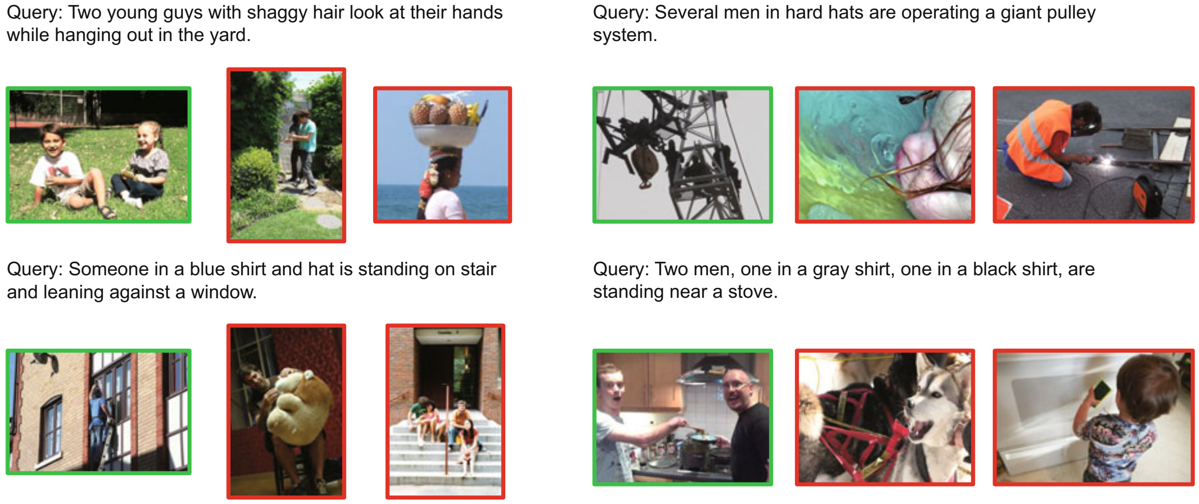

Query: Two young guys with shaggy hair look at their hands while hanging out in the yard.
Query: Several men in hard hats are operating a giant pulley system.
Query: Someone in a blue shirt and hat is standing on stair and leaning against a window.
Query: Two men, one in a gray shirt, one in a black shirt, are standing near a stove.

Fig. 7 Qualitative results of image retrieval for the given sentence queries. For each sentence query, we show the top-three ranked images, ranking from left to right. We outline the true matches in green boxes and false matches in red boxes. References to color refer to the online version of this figure

the Association for Computational Linguistics: Human Language Technologies, p.4171-4186. https://doi.org/10.18653/v1/n19-1423

Diao HW, Zhang Y, Ma L, et al., 2021. Similarity reasoning and filtration for image-text matching. Proc AAAI 35th Conf on Artificial Intelligence, p.1218-1226. https://doi.org/10.1609/aaai.v35i2.16209

Faghri F, Fleet DJ, Kiros JR, et al., 2018. VSE++: improving visual-semantic embeddings with hard negatives. British Machine Vision Conf, Article 12.

Geigle G, Pfeiffer J, Reimers N, et al., 2022. Retrieve fast, rerank smart: cooperative and joint approaches for improved cross-modal retrieval. *Trans Assoc Comput Ling*, 10:503-521. https://doi.org/10.1162/tacl_a_00473

Ghosh A, Kumar H, Sastry PS, 2017. Robust loss functions under label noise for deep neural networks. Proc 31st Conf on Artificial Intelligence, p.1919-1925.

Han B, Yao QM, Yu XR et al., 2018. Co-teaching: robust training of deep neural networks with extremely noisy labels. Proc 32nd Int Conf on Neural Information Processing Systems, p.8536-8546.

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Huiskes MJ, Lew MS, 2008. The MIR flickr retrieval evaluation. Proc 1st ACM Int Conf on Multimedia Information Retrieval, p.39-43. https://doi.org/10.1145/1460096.1460104

Jia C, Yang YF, Xia Y, et al., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. Proc 38th Int Conf on Machine Learning, p.4904-4916.

Jiang L, Zhou ZY, Leung T, et al., 2018. MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. Proc 35th Int Conf on Machine Learning, p.2309-2318.

Karpathy A, Li FF, 2015. Deep visual-semantic alignments for generating image descriptions. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3128-3137. https://doi.org/10.1109/CVPR.2015.7298932

Kingma DP, Ba J, 2015. Adam: a method for stochastic optimization. Proc 3rd Int Conf on Learning Representations.

Lee KH, Chen X, Hua G, et al., 2018. Stacked cross attention for image–text matching. Proc 15th European Conf on Computer Vision, p.212-228. https://doi.org/10.1007/978-3-030-01225-0_13

Li KP, Zhang YL, Li K, et al., 2019. Visual semantic reasoning for image-text matching. IEEE/CVF Int Conf on Computer Vision, p.4653-4661. https://doi.org/10.1109/ICCV.2019.00475

Li XJ, Yin X, Li CY, et al., 2020. Proc 16th European Conf on Computer Vision, p.121-137. https://doi.org/10.1007/978-3-030-58577-8_8

Lin TY, Maire M, Belongie S, et al., 2014. Proc 13th European Conf on Computer Vision, p.740-755. https://doi.org/10.1007/978-3-319-10602-1_48

Lin XY, Bhattacharjee D, El Helou M, et al., 2021. Fidelity estimation improves noisy-image classification with pretrained networks. *IEEE Signal Process Lett*, 28:1719-1723. https://doi.org/10.1109/LSP.2021.3104769

Liu TL, Tao DC, 2016. Classification with noisy labels by importance reweighting. *IEEE Trans Patt Anal Mach Intell*, 38(3):447-461. https://doi.org/10.1109/TPAMI.2015.2456899

Lu JS, Batra D, Parikh D, et al., 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Proc 33rd Int Conf on Neural Information Processing Systems, p.13-23.

Lyu YM, Tsang IW, 2020. Curriculum loss: robust learning and generalization against label corruption. Proc 8th Int Conf on Learning Representations.

Ma X, Huang H, Wang Y, et al., 2020. Normalized loss functions for deep learning with noisy labels. Proc 37th Int Conf on Machine Learning, p.6543-6553.

Ma XJ, Wang YS, Houle ME, et al., 2018. Dimensionality-driven learning with noisy labels. Proc 35th Int Conf on Machine Learning, p.3361-3370.

Ma ZY, Leijon A, 2011. Bayesian estimation of beta mixture models with variational inference. *IEEE Trans Patt Anal Mach Intell*, 33(11):2160-2173. https://doi.org/10.1109/TPAMI.2011.63

Manwani N, Sastry PS, 2013. Noise tolerance under risk minimization. *IEEE Trans Cybern*, 43(3):1146-1151. https://doi.org/10.1109/TSMCB.2012.2223460

Messina N, Amato G, Esuli A, et al., 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Trans Multim Comput Commun Appl*, 17(4):128. https://doi.org/10.1145/3451390

Niwattanakul S, Singthongchai J, Naenudorn E, et al., 2013. Using of jaccard coefficient for keywords similarity. Proc Int MultiConf of Engineers and Computer Scientists, p.380-384.

Radford A, Kim JW, Hallacy C, et al., 2021. Learning transferable visual models from natural language supervision. Proc 38th Int Conf on Machine Learning, p.8748-8763.

Reed SE, Lee H, Anguelov D, et al., 2015. Training deep neural networks on noisy labels with bootstrapping. Proc 3rd Int Conf on Learning Representations.

Ren SQ, He KM, Girshick R, et al., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Patt Anal Mach Intell*, 39(6):1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

Ruder S, 2017. An overview of multi-task learning in deep neural networks. https://arxiv.org/abs/1706.05098

Song H, Kim M, Lee JG, 2019. SELFIE: refurbishing unclean samples for robust deep learning. Proc 36th Int Conf on Machine Learning, p.5907-5915.

Song H, Kim M, Park D, et al., 2020. Learning from noisy labels with deep neural networks: a survey. https://arxiv.org/abs/2007.08199

van der Maaten L, Hinton G, 2008. Visualizing data using t-SNE. *J Mach Learn Res*, 9(86):2579-2605.

Wang KY, Yin QY, Wang W, et al., 2016. A comprehensive survey on cross-modal retrieval. https://arxiv.org/abs/1607.06215

Wang RX, Liu TL, Tao DC, 2018. Multiclass learning with partially corrupted labels. *IEEE Trans Neur Netw Learn Syst*, 29(6):2568-2580. https://doi.org/10.1109/TNNLS.2017.2699783

Yang J, Duan J, Tran S, et al., 2022. Vision-language pretraining with triple contrastive learning. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15650-15659. https://doi.org/10.1109/CVPR52688.2022.01522

Zhang HY, Xing XM, Liu L, 2021. DualGraph: a graph-based method for reasoning about label noise. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9654-9663. https://doi.org/10.1109/CVPR46437.2021.00953