Deep Learning for Corrupted Image-Text Matching

Sunwoo Kim, Seongwook Shin, and Byonghyo Shim
Seoul National University

{swkim, swshin, bshim}@islab.snu.ac.kr

Abstract

Image-text matching plays a vital role in connecting clean images and texts. In many real-world applications, since a large number of corrupted images exists, it is very difficult to match images and texts accurately. In this paper, we propose a novel corrupted image-text matching scheme, referred to as robust corrupted image-text matching (RCIT), suitable for various real-world scenarios.

Ⅰ. Introduction

In recent years, deep learning (DL) has achieved great success in a variety of real-world problems such as network localization [1]. The application of DL techniques in the field of image-text matching has also emerged as an active topic in the last three years [2]. Image-text matching is a task to search for images that match with their corresponding texts (i.e., sentences) and vice versa. Although great progress has been made, image-text matching remains challenging in many real-world applications such as google image search due to a large number of corrupted images. When image quality degrades, it is highly likely that the conventional image-text matching schemes cannot perform image-text matching effectively.

In this paper, we propose a novel image-text matching scheme suitable for various real-world scenarios. The main idea of the proposed scheme, henceforth referred to as *robust corrupted image-text matching* (RCIT), is to gather the matching triplets (clean and corrupted images and text) in the common embedding space. It is now well-known that many image-text matching schemes rely on common embedding space, where the distance between paired image and text instances is reduced, to infer the image-text similarity [3]. By gathering matching triplets together in the embedding space, we can achieve accurate image-text matching even with corrupted images.

Ⅱ. Visual Semantic Embedding

In this section, we briefly review the basics of visual semantic embedding (VSE) approach used in image-text matching. First, we use the image feature extractor (e.g., CNN) and text feature extractor (e.g., RNN) to obtain the set of visual and text features from image and text (i.e., sentence), respectively. Then, the extracted features are aggregated by feature aggregators into the common embedding space:

$$\mathbf{v} = f_{VISUAL}(\{\phi_n\}_{n=1}^{N})$$
$$\mathbf{u} = f_{TEXT}(\{\psi_n\}_{n=1}^{N})$$

where $\mathbf{v}$ and $\mathbf{u}$ are visual and text instance vectors of same dimension in the common embedding space.

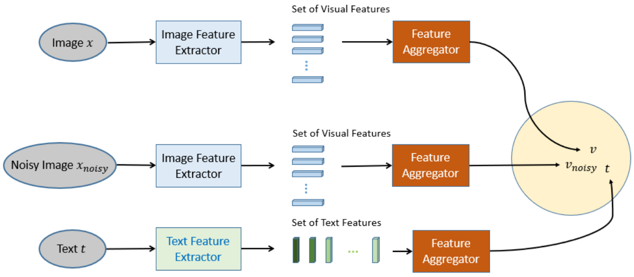


Figure 1. Overall Training Process of RCIT

Finally, the image-text similarity score is defined as the dot product between $\mathbf{v}$ and $\mathbf{u}$. In the test phase, the similarity scores are used to rank a query text against all possible images and vice versa.

Ⅲ. Corrupted Image-Text Matching

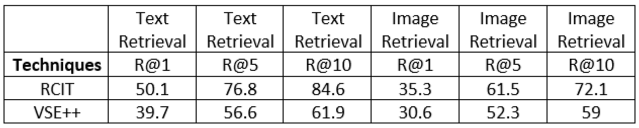
Fig. 1 depicts the overall training process of RCIT scheme. The key operation of RCIT is to jointly minimize the distance between clean image-text pair and corrupted image-text pair. In doing so, we can learn the proper embeddings representation of corrupted images in the common embedding space Distinctive feature of RCIT over the conventional image-text matching scheme is that matching clean and corrupted images and texts are used simultaneously in the training process. To jointly train the three feature extractors, we combine the two bidirectional ranking losses used for matching clean image-text and corrupted image-text pairs with the loss used for matching corrupted image-clean image pair [2]. That is,

$$L = L_{CI-T} + L_{I-T} + L_{CI-I}$$

Since three modules are trained together, the similarity representations are shared, allowing greater capacity in characterizing the complex alignments between corrupted images, clean images, and sentences.

Ⅳ. Simulation Result

In this section, we present the simulation results to evaluate the image-text matching performance of the proposed RCIT. For dataset, we use Flickr30k, which



	Text Retrieval	Text Retrieval	Text Retrieval	Image Retrieval	Image Retrieval	Image Retrieval
Techniques	R@1	R@5	R@10	R@1	R@5	R@10
RCIT	50.1	76.8	84.6	35.3	61.5	72.1
VSE++	39.7	56.6	61.9	30.6	52.3	59

Table 1. Corrupted Image-Text Matching performance

consists of 29,783 training images, 1,000 images for validation and testing [2]. To generate the corrupted image from the given clean image, we contaminate with the clean image with either speckle or salt-pepper noise with noise rate of 0.6. Note that noise rate is defined as the fraction of number of unclean pixels over the total number of pixels of an image. For fair comparison, we train VSE++ with corrupted images. As performance measure, Recall at K (R@K), which is defined as the proportion of queries whose ground-truth exists within the top K chosen candidates, is considered. In Table 1, we evaluate the image-text matching performance of RCIT and compare it with that of VSE++, a state-of-the-art VSE-based image-text matching technique [3]. From the result, we observe that R@1, R@5, and R@10 of RCIT are much higher than those of VSE++.

V. Conclusion

In this paper, we proposed a novel image-text matching scheme suitable for many real-world applications where a large number of images with pool quality exists. We demonstrated from the numerical evaluations that the proposed scheme is very effective in terms of matching image and text accurately regardless of image quality.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2021-2017-0-01637) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

REFERENCE

[1] S. Kim and B. Shim, "Localization of internet of things network via deep neural network based matrix completion," in Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC), pp. 1766-1770, Oct. 2020.

[2] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," arXiv:2101.01388, 2021.

[3] F. Faghri, D. J. Fleet, R. Kiros, and S. Fidler, “VSE++: Improving visual-semantic embeddings with hard negatives,” in Proc. Brit. Mach. Vis. Conf. (BMVC), pp. 1-13, 2018.