RoCOCO: Robust Benchmark MS-COCO to Stress-test Robustness of Image-Text Matching Models

Seulki Park[1] Daeho Um[1] Hajung Yoon [1] Sanghyuk Chun[2] Sangdoo Yun[2] Jin Young Choi[1]
[1]ASRI, ECE., Seoul National University [2]NAVER AI Lab

Abstract

Recently, large-scale vision-language pre-training models and visual semantic embedding methods have significantly improved image-text matching (ITM) accuracy on MS COCO 5K test set. However, it is unclear how robust these state-of-the-art (SOTA) models are when using them in the wild. In this paper, we propose a novel evaluation benchmark to stress-test the robustness of ITM models. To this end, we add various fooling images and captions to a retrieval pool. Specifically, we change images by inserting unrelated images, and change captions by substituting a noun, which can change the meaning of a sentence. We discover that just adding these newly created images and captions to the test set can degrade performances (i.e., Recall@1) of a wide range of SOTA models (e.g., 81.9% → 64.5% in BLIP, 66.1% → 37.5% in VSE∞). We expect that our findings can provide insights for improving the robustness of the vision-language models and devising more diverse stress-test methods in cross-modal retrieval task. Source code and dataset will be available at https://github.com/pseulki/rococo.

1. Introduction

Understanding the visual world with language is a crucial ability for artificial intelligence, which has inspired the research of image-text matching. Recently, the development of various methods [49, 16, 10] and large-scale vision-language pretraining models [60, 79, 47] have significantly improved image-text matching accuracy (i.e., recall@1). However, how much can we trust these numbers? How are we good, when using it in the wild?

Figure 1 shows an example of the test results, when adding additional images and texts to COCO [54] test set, tested with the recent state-of-the-art (SOTA), BLIP [47]. When we retrieve the top-ranked text from a given image, the "umbrella" is mistakenly recognized as a "gun". Meanwhile, given the text, the image partially mixed with unrelated image (i.e., skiing on the snow) is retrieved as top 1, instead of the correct clean image. These errors can pose serious risks, when deploying a model in real life. For example, an innocent citizen may be perceived as a threat when used in a defense industry (e.g., Fig. 1 Upper). Or, malicious images (e.g., pornographic pictures) can be inserted into other images to make them searchable on websites (e.g., Fig. 1 Lower). Lastly, these errors can greatly damage the trust of AI users. Therefore, the robustness test is important.

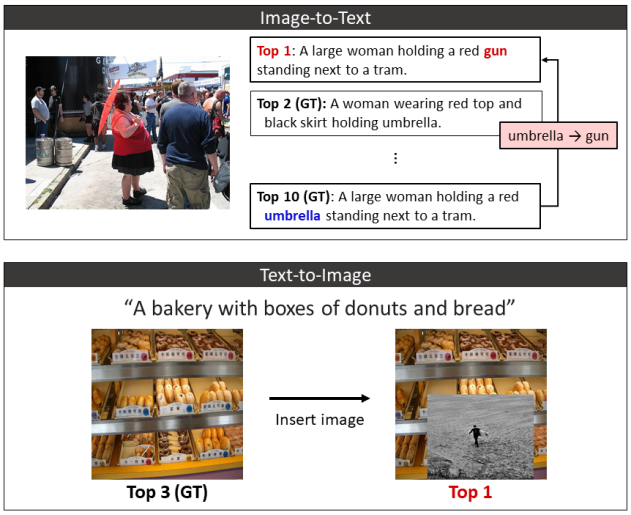


Figure 1: **Example of Image-text matching (ITM) results from the state-of-the-art BLIP [47].** When we add a new caption with only one word changed from "umbrella" to "gun", this new caption is retrieved as top 1 (Image-to-text). Likewise, when we add a new image created by inserting an unrelated image to the original one, this new image is ranked as top 1 (Text-to-image). In this paper, we discover the common weakness in ITM models and propose a novel robustness-evaluation benchmark.

To evaluate the robustness of the models, various attempts have been made in computer vision [33, 35, 32], and natural language processing (NLP) [37, 2, 21] areas, respectively. Recently, various robustness test methods have been introduced for multi-modal tasks, such as in visual question answering (VQA) [78, 28, 63, 26, 64, 51] and image cap-

tioning [59]. While stress-test datasets for fooling models have been actively proposed in the field of VQA, to the best of our knowledge, there has been no proposed dataset in the image-text retrieval task yet.

In this paper, we propose a novel evaluation benchmark to stress-test how robust the model is in image-text matching. While the existing image-text retrieval has been evaluated on a well-set query-image/text pool, we add various fooling text/images that can exist in real scenarios. Surprisingly, the models are easily fooled by simply changing words (e.g., umbrella $\rightarrow$ gun, man $\rightarrow$ pizza) or attaching a completely unrelated image to the original image, which would not confuse human at all. For example, by adding such examples to the original test set, the image-to-text retrieval accuracy (i.e., recall@1) has significantly dropped from $81.9\% \rightarrow 64.5\%$ in BLIP, and $66.1\% \rightarrow 37.5\%$ in VSE∞ [10] Also, for text-to-image retrieval, the recall@1 has dropped from $64.3\% \rightarrow 40.7\%$ in BLIP, and $51.6\% \rightarrow 34.3\%$ in VSE∞. We verify the consistent performance drop in other models as well regardless of using large-scale pre-training datasets or not. In addition, we observe the SOTA model could even confuse nonsensical captions when we further break the semantic meaning of the caption by replacing multiple words (e.g., "a grey computer mouse and a silver metal key." $\rightarrow$ "a baths light bent over wartime seo wasn."). From the observations, we conjecture that it is important to devise a robust learning algorithm, such that the model can better learn the word-level semantic meaning and its alignment to images.

Our key contributions can be summarized as follows:

- We provide various robustness-evaluation benchmark for image-text matching.
- We evaluate the state-of-the-art ITM models whose weights are publicly available on our newly proposed dataset and discover the significant performance drops across all models.
- Our results show that current models are paying attention to specific word or image part, rather than understanding the whole semantic meaning well.

We expect that our findings can provide insights for improving the robustness of the vision-language models and devising more diverse stress-test methods in cross-modal retrieval task.

2. Related Work

2.1. Image-Text Matching

Methods. Most image-text matching (ITM) methods [24, 23, 65, 36, 16, 71, 77] aim to learn joint visual-semantic embedding (VSE) such that paired image and text representation in the embedding space are close. Many VSE methods [44, 70, 20, 10] use region features extracted from Faster R-CNN [62] with bottom-up attention [3]. VSE∞ [10] also use grid features extracted from Faster R-CNN pre-trained on Visual Genome [42] and ImageNet [18] in [3], and Instagram pretrained ResNext-101 [72].

In recent years, large-scale pre-training models [12, 53, 79, 38, 40, 48, 47, 55, 13, 73, 1] have shown strong achievement in both zero-shot and fine-tuned performances. Most of these models adopt transformer architecture and can learn cross-modal representations benefiting from large-scale image-text pairs. For a more thorough study, we refer the reader to a recent survey [7]. In this paper, we re-evaluate the robustness of state-of-the-art ITM models.

Datasets. Recently, new ITM benchmark datasets [58, 15] have been proposed by extending MS COCO. Crisscrossed Captions (CxC) [58] add semantic similarity between all pairs to improve limited associations in MS COCO. Thus, CxC has enabled scoring between intra- and intermodality pairs. Meanwhile, ECCV caption [15] provides abundant positive image-caption pairs to correct the false negatives in MS COCO. While the previous works provided improved benchmark datasets, our main difference is that we aim to test the vulnerability of the models.

2.2. Robustness Test

Unimodal: After the initial finding [67] that deep learning (DL) models are vulnerable to imperceptible perturbations, robustness in deep learning methods has actively studied in both computer vision and natural language processing (NLP) areas. In computer vision, one research direction is data poisoning [5, 66, 34, 29, 11], which attacks the robustness of models during training by adding images with small perturbations. Meanwhile, adversarial attack studies [27, 43, 9, 17, 30] inject imperceptible noises to test images so that a model can make wrong predictions. For image retrieval task, Li et al. [46] showed that adding invisible noise to query image can make the model return incorrect images. Another line of research has proposed new ImageNet benchmarks for common robustness evaluation. For example, ImageNet-C [33] is applied with 75 common visual corruptions, and ImageNet-P [33] is implemented with common perturbations. Also, ImageNet-A [35] provides images belonging to ImageNet classes but more difficult, and ImageNet-R [32] introduces examples with various renditions. In NLP, research on data poisoning [69] and adversarial attacks [22, 2, 39, 25, 45, 21, 6] has also been actively studied to fool the prediction of models. Adversarial examples are produced by character-level modifications [4], paraphrasing sentences [37], or substituting a word with a synonym [61, 52].

The main difference between these methods and our work is that while the previous works generate human imperceptible noises and semantic-preserving texts, we rather generate perceptibly different images and semantic-breaking texts. Our intuition is straightforward: a robust

model should not be at least confused by the examples which are easy for human.

Multimodal: As vision-language models have generated growing research interest, robustness work for cross-modal domain has been actively studied [59, 50, 8]. Especially, in visual question answering (VQA) task, diverse robustness-evaluation benchmark [78, 28, 63, 26, 64, 51] has been proposed. For example, VQA-Rephrasings [63] generated dataset by rephrasing questions to evaluate the robustness in the input question. Adversarial VQA [51] and AdVQA [64] collected adversarial examples in human-in-the-loop manner. However, to the best of our knowledge, this is the first work to propose robustness-evaluation benchmark in ITM task. We hope that our work can inspire the future research to create more diverse stress-test benchmarks in ITM area.

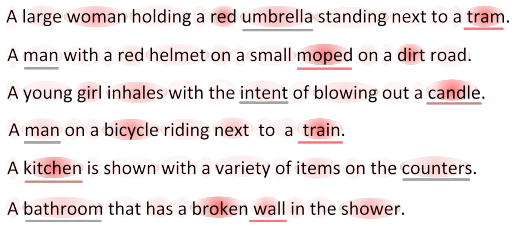


Figure 2: **Influence of a word in a caption.** The darker the red color of a word, the greater its influence. For each caption, the noun with the highest EI score is underlined in red, and the noun with the lowest EI score is underlined in gray. We can observe that some semantically important nouns such as 'man' and 'bathroom' have low EI scores, which can make a model not robust.

3. Robustness-Evaluation Benchmark

Our goal is to quantitatively measure how well ITM models understand both text and image. To this end, we add new fooling images and captions to COCO 5K test set. To effectively create confusing captions and images, we pay attention to how ITM models retrieve texts and images from a given query. ITM models [47, 10] find the most matching pair through a similarity measure (e.g., cosine similarity) between learned text embedding and image embedding. Thus, our assumption is that a new caption or an image with minimal changes to the original embedding will be able to fool a model since its similarity score is likely to remain similar. In the following sections, we describe the example-pair generation process in detail.

3.1. Caption Generation for given Image

To generate a fooling caption, we create a sentence whose meaning changes considerably, but whose embedding does not change much from the existing embedding. To this end, we replace one meaningful word in the sentence. Substituting a word in a sentence is a commonly used method for adversarial attacks in natural language processing (NLP) [21, 75]. Unlike the previous methods replacing a word with semantically similar words, we replace words with completely different or unrelated words (e.g., "umbrella" $\rightarrow$ "gun"). If the model gets confused and retrieves this new caption with different meaning as top 1, as in Figure 1, it is difficult to say that the model is robust. Through this, it is possible to check if the model successfully learned the alignment of image and text embedding by understanding semantic details of the image.

3.1.1 Embedding-Influence for Source Word Selection

To substitute a word in a caption, we need to choose which word (i.e., source word) to replace. First, in order to change the meaning of the caption significantly, we restrict the source words to noun. Among the nouns, we exclude the words whose substitution would not considerably change the meaning of the sentence. For example, from the caption, "A row of motorcycles parked in front of a building", we do not include the noun "row" in the source words, since the replacement of "row" does not meaningfully change the sentence.

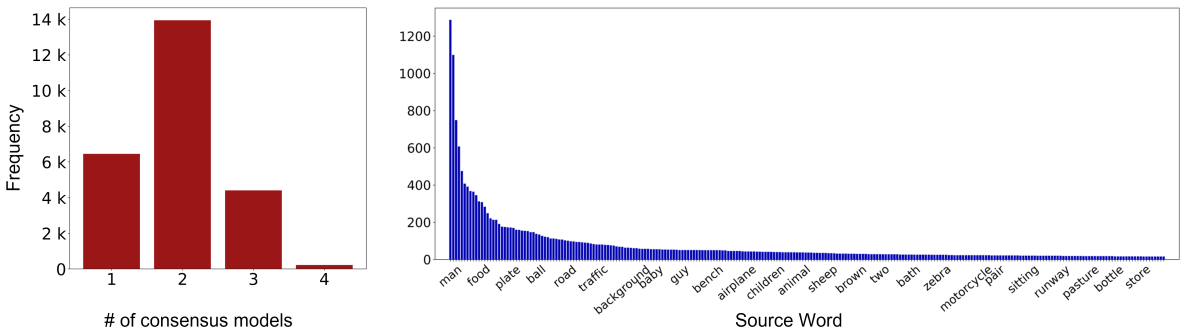
Then, we need to select one word to change among the meaningful source words (e.g., "motorcycles", "front", "building" in the caption above). To change the meaning of a caption with minimal changes to its embedding, we propose embedding-influence (EI) score that can estimate the influence of each word. EI sore measures the change in embedding when the word is removed from the caption. Changing a word with a low EI score will not show much change in embedding compared to other words. Thus, similar embedding output is likely to maintain a high similarity score and confuse a model.

To estimate the influence of each word, we measure the change in embedding when the word is removed from the caption. It is a classic technique in robust statistics to measure influence by estimating a change in prediction when a sample is removed [41]. Given a text encoder f_T, and a caption $C = \{c_m \mid m = 1, \cdots, M\}$ where M is the number of words in C, the embedding-influence (EI) score of a word, c_s, can be defined by

$$EI(c_s) = 1 - \frac{< f_T(C), f_T(C \setminus c_s) >}{\| f_T(C) \| \| f_T(C \setminus c_s) \|}, \qquad (1)$$

where $<, >$ denotes the dot(inner) product operation. A low EI score means that the embedding output of the caption without the word is similar to the original caption. That is, the deleted word has low influence to the embedding result for the caption.

Figure 2 shows the example of different influences of words in each caption. The darker the red color for a word,



(a) Consensus among models.

(b) Source word distribution.

Figure 3: **Statistics of source words selected by EI scores.** (a) Consensus among models. x-axis indicates the number of consensus models that pick the same word, y-axis the frequency of consensus for each case in x-axis over all captions. (b) The source word distribution exhibits a long-tailed distribution.

the greater the influence of the word. For each caption, we underline the noun with the highest EI score in red and the noun with the lowest one in gray. Since a noun such as "umbrella" (object) and "man" (subject) has seemingly important meaning, the meaning of the sentence with the noun(e.g., umbrella) replaced to another one (e.g., gun) can be changed considerably.

However, the model can hardly recognize the change when the noun (e.g. umbrella) has the lowest EI score that does not much affect the embedding output, compared to the noun (e.g. tram) with the highest EI score. This makes the model still select the changed sentence as top-1, which is a false matching. This provides our motivation that the noun with the lowest EI score becomes a source word to be replaced by a target word chosen according to the scheme described in the next section. In experimental section, we empirically verify the effectiveness of EI scores.

Because EI score of a word varies depending on a model, we employ four different representative models in Table 1 to measure EI score of a word. After gathering the word with the lowest score in each model, we choose one word that appear most frequently in the four models among the gathered words, whereas if the most frequent word is not one, the source word is randomly selected among the lowest-EI words chosen by the models. Figure 3 represents the statistics of source words selected based on the lowest EI scores across four models. Figure 3(a) shows the frequency of consensus among models: how many models pick the same word for each caption, where x-axis indicates the number of models that pick the same word, y-axis the frequency of consensus for each case in x-axis over all captions. Surprisingly, two or more models choose the same word in the percentage more than 70% though the models are trained using different architectures and datasets (e.g., pre-training dataset). A similar phenomenon has also been observed in image-only classifiers [57]. This implies that the currently proposed models have similar vulnerabilities against word-level attacks in common. In experiment section, we suggest necessity of new robust ITM models by various word-level modifications.

Table 1: **Models used for calculating EI scores.**

Model	COCO-trained	Text backbone
VSRN [49]	Finetuned	Bi-GRU [14]
CLIP [60]	Zero-shot	Transformer [68]
VSE∞ [10]	Finetuned	BERT [19]
BLIP [47]	Finetuned	BERT

Table 2: **Added Concept Groups.** We include new concepts, which are not included in GRIT [31]. Table shows added unique concepts and 3 random words from each group.

concept group	#concepts	concept lemmas (sampled)
material	32	metal, plastic, wooden
color	28	black, white, brown
direction	50	front, middle, bottom
vehicle_part	12	hood, wheel, tire
shape	15	round, square, octagon
event	11	Christmas, birthday, wedding
number	14	one, five, hundreds

3.1.2 Replacement with Target Word

Next, we need to decide a target word that replaces a selected source word. To create various confusing captions, we adopt three policies to generate new words.

First, we use concept groups from recently proposed GRIT benchmark [31]. GRIT has grouped a large number of nouns from popular computer vision datasets including COCO into 24 concept groups (e.g., food, people, places,

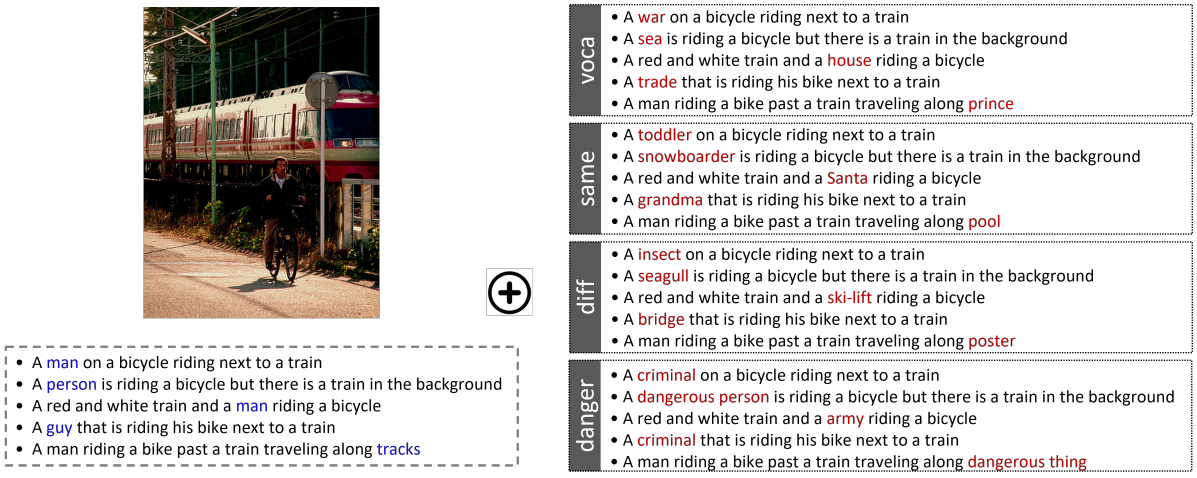


Figure 4: **Example of generated captions.** (Left) Original COCO image and captions. (Right) Our generated captions, Rand-voca, Same-concept, Diff-concept, and Danger from top to bottom. The model is to retrieve the most appropriate caption from a pool of both original and newly generated captions. Our assumption is that the robust model should be able to retrieve the original captions well without being confused by new captions with different meanings.

and so on). We add 7 concept groups (see Table 2) for words that are hard to be included in the given concept groups. Next, we categorize words according to the concept groups. Then, a source word is randomly replaced by any target word inside **Same-concept** or **Diff-concept**. For example, **Same-concept** replaces "umbrella" with any word in the same concept (i.e., tools), which can be "rope", "boxes", and so on. **Diff-concept** replaces "umbrella" with any word selected randomly from different concepts, such as "pizza" from "food" concept, and "monkey" from "animal" concept. We exclude cases when the meaning is not significantly changed (e.g., "umbrella" $\rightarrow$ "parasol").

Second, we replace a source word with a target word randomly selected from large-scale BERT [19] vocabulary (**Rand-voca**). We use words consisting of only English letters, excluding those in other languages or special characters. We also exclude cases when the meaning is not significantly changed.

Finally, as an example of a critical situation that can be applied in practice, we create a special case (**Danger**), by using words related to public security. For example, "umbrella" is replaced by "gun" or "weapon". Figure 4 shows the examples of generated captions. We show more examples in the Supplementary material.

3.2. Image Generation

To create images that are perceptible to human but confusing for models, we design a new image by mixing it with a fake image. By using the mixed version of the original image, we can bring minimal changes to the image embedding. The practical use case is that malicious contents can be inserted into normal images to confuse filtering algorithms and make it searchable on websites. These images are more challenging that the malicious image is hidden in the normal image.

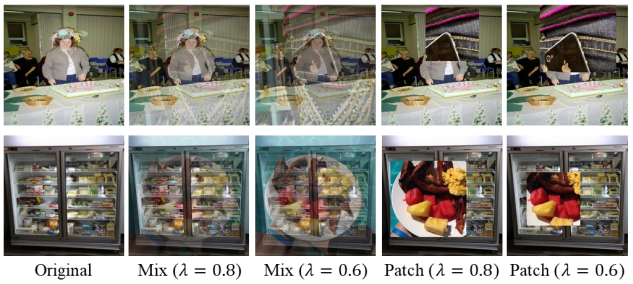


Figure 5: **Example of generated images with different λ.**

We mix two images in Mixup [76] (Mix) and Cutmix [74] (Patch) styles, respectively, as shown in Figure 5. When inserting a fake image x^f into an original image x^o, we test with different mixing ratios λ and $\mathbf{M}$ as in,

$$\text{Mix} : \tilde{x} = \lambda x^o + (1-\lambda)x^f$$
$$\text{Patch} : \tilde{x} = \mathbf{M} \odot x^o + (\mathbf{1}-\mathbf{M}) \odot x^f$$

where $\mathbf{M} \in \{0,1\}^{W \times H}$ denotes a binary mask, where W is the width and H is the height of the image. In Patch, λ is calculated as $\lambda = \frac{\sum_{i,j} \mathbf{M}_{i,j}}{W \times H}$.

4. Experiments and Results

4.1. Experimental setting

In this section, we evaluate the existing image-text matching (ITM) models on our new dataset. Since the pair between newly created images and captions is not of our interest, we calculate image-to-text scores between the original 5,000 images and new 50,000 captions. On the contrary, in the text-to-image task, 25,000 captions are to retrieve the best image from a pool of 10,000 images.

4.1.1 Evaluation Metrics

Recall@k, especially Recall@1 (R@1), is the most popular metric for evaluating the existing ITM methods. In this paper, we propose two metrics to evaluate *Drop Rate* and *False Recall*@1 (FR@1) in addition to R@1. Drop rate measures how much R@1 has dropped when the models are tested on the new retrieval pool, compared to the original COCO 5K testset. We calculate drop rate as $(\text{R@1} - \text{R}_{\text{New}}\text{@1})/\text{R@1}$.

False Recall@1 calculates the percentage that newly added incorrect captions/images are retrieved as top 1. This can quantitatively estimate the vulnerability of a model.

4.1.2 Models for Evaluation

We compare 14 state-of-the-art Vision-Language (VL) models, whose trained weights are available to the public. They can be categorized into two groups; large-scale vision-language(VL) pre-training and visual semantic embedding groups.Large-scale VL pre-training group includes CLIP with ViT-B/32, ViT-B/16 and ViT/L14 backbones [60], fine-tuned ALBEF [48], and zero-shot and fine-tuned BLIP with ViT-B and ViT-L backbones [47]. While 'zero-shot' and 'fine-tuned' models are both pre-trained on large-scale datasets, 'zero-shot' refers to not being fine-tuned with COCO train set.

Visual semantic embedding group includes models using region features based on bottom-up attention [3] and SCAN [44]: VSRN [49], SAF, SGR [20], and VSE∞ with BUTD region [10]. Also, we evaluate on VSE∞ [10] with BUTD grid and WSL grid. BUTD grid uses grid features pretrained on ImageNet and Visual Genome [3]. WSL grid is pre-trained on Instagram [56].

4.2. Main results

4.2.1 Image-to-Text Retrieval

Table 3 reports the image-to-text retrieval results on our new datasets. All models show a significant performance drop across the testsets. From the results, we have observed following points.

Data perspective. Evaluation on Rand-voca shows the most performance degradation as can be seen from Table. We conjecture that this is because Rand-voca includes many unexpected words that do not commonly co-appear in captions, whereas in Same-concept and Diff-concept, both the source words and the target words belong to the same dataset, COCO. The above observation means that models seem more vulnerable to sentences made up of unfamiliar combinations of words that have rarely appeared in the trained captions. This conjecture questions the model's ability to understand the word-level meaning within the sentence.

Model perspective. We expect Large-scale VL pre-training models should be more robust since they are trained on additional pre-training datasets (e.g., 400M image pairs in CLIP [60], 129M in BLIP [47], 14M in ALBEF [48]). However, the results of drop rate or false recall@1 do not show superior robustness to the conventional Visual Semantic Embedding models. We assume that since Large-scale VL pre-training models learn multimodal representation by minimizing the distance of the matched image-text pair (i.e., ITM loss), they could be vulnerable to a single-word change in caption. Therefore, we argue that it is important to devise a robust learning algorithm, such that the model can better learn the word-level semantic meaning and its alignment to images.

4.2.2 Text-to-Image Retrieval

We evaluate VL methods on new image set with $\lambda = 0.9, 0.8$ in Table 4. We have generated images with three random seeds and report the averaged results. We can observe consistent degradation for all VL methods.

We display the examples of retrieving incorrect images with BLIP ViT-B when $\lambda = 0.8$ in Figure 6. While human would not prefer to mixed images to the original images, we observe that models are easily confused two images. We presume that models attend to specific region in an image without understanding the whole context and details of images.

We believe that this is a simple but effective way to test the model's robustness. We show the results on more λ and more visualization of images in Supplementary material.

Figure 6: **Image retrieval examples.**

Table 3: **Image-to-Text retrieval results.** Models are re-evaluated on four new benchmark datasets: Rand-voca, Same-concept, Diff-concept, and Danger. Recall@1 (R@1)(↑), drop rate(↓), False Recall@1 (FR@1)(↓) are shown. We can see consistent degradation across all vision-language models regardless of using pre-training datasets and different methods. The biggest performance drops are marked in bold.

	COCO 5K	Rand-voca			Same-concept			Diff-concept			Danger		
	R@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1
Large-scale VL pre-training models													
CLIP ViT-B/32 (zero-shot) [60]	50.10	36.44	27.27	34.63	**35.77**	**28.60**	**36.64**	37.48	25.18	32.27	42.18	15.81	19.69
CLIP ViT-B/16 (zero-shot) [60]	52.44	**38.18**	**27.19**	**34.87**	38.36	26.85	34.40	40.23	23.28	30.57	44.67	14.81	18.19
CLIP ViT-L/14 (zero-shot) [60]	56.04	**39.90**	**28.81**	33.95	40.90	27.02	**34.86**	42.66	23.88	24.07	46.48	17.06	30.16
ALBEF [48]	77.58	**60.13**	**22.49**	**26.07**	60.55	21.95	25.09	61.84	20.29	23.75	63.37	18.32	20.43
BLIP ViT-B (zero-shot) [47]	70.54	**35.28**	**49.98**	**54.58**	47.77	32.28	37.45	45.58	35.39	40.89	42.39	39.90	43.99
BLIP ViT-B [47]	81.90	**64.50**	**21.25**	**23.72**	68.74	16.07	18.74	69.20	15.51	17.36	67.81	17.21	18.92
BLIP ViT-L (zero-shot) [47]	73.66	**45.96**	**37.60**	**40.49**	55.38	24.82	28.27	55.69	24.39	27.56	55.93	24.07	26.54
BLIP ViT-L [47]	82.36	**66.84**	**18.85**	**21.18**	71.16	13.60	16.02	72.70	11.72	13.86	72.37	12.13	13.73
Visual Semantic Embedding models													
VSRN [49]	52.66	**42.22**	**19.82**	**22.14**	44.56	15.38	18.06	46.12	12.41	14.47	46.78	11.17	12.77
SAF [20]	55.46	**39.30**	**29.14**	**31.54**	42.04	24.20	28.35	45.00	18.85	22.24	42.77	22.88	26.35
SGR [20]	57.22	**41.69**	**27.14**	**30.43**	43.61	23.79	28.02	46.56	18.63	22.07	44.90	21.53	24.72
VSE∞ (BUTD region) [10]	58.02	**31.71**	**45.34**	**47.99**	39.79	31.42	35.12	36.91	36.38	39.86	37.66	35.09	37.38
VSE∞ (BUTD grid) [10]	59.40	**32.24**	**45.72**	**48.75**	41.12	30.77	33.58	38.71	34.84	38.40	39.71	33.15	35.32
VSE∞ (WSL grid) [10]	66.06	**37.54**	**43.17**	**46.07**	48.76	26.19	29.59	44.86	32.09	35.06	45.39	31.29	33.07

Table 4: **Text-to-Image retrieval.** Models are evaluated with our new benchmark: Mix and Patch with different λ. Recall@1 (R@1)(↑), drop rate(↓), False Recall@1 (FR@1)(↓) are shown. The results are averaged over image generations with three different random seeds. We can see consistent degradation across all vision-language models.

	COCO 5K	Mix ($\lambda = 0.9$)			Mix ($\lambda = 0.8$)			Patch ($\lambda = 0.9$)			Patch ($\lambda = 0.8$)		
	R@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1
Large-scale VL pre-training models													
CLIP ViT-B/32 (zero-shot) [60]	30.14	20.29	32.68	33.55	22.79	24.39	26.03	22.49	25.38	28.63	24.15	19.87	23.69
CLIP ViT-B/16 (zero-shot) [60]	33.03	20.05	39.30	39.00	23.57	28.64	29.88	22.58	31.64	35.18	24.70	25.22	29.41
CLIP ViT-L/14 (zero-shot) [60]	36.14	25.49	29.47	28.99	27.75	23.22	24.29	27.56	23.74	27.64	29.09	19.51	23.97
ALBEF [48]	60.67	44.13	27.27	26.60	48.02	20.85	21.11	48.86	19.47	19.58	51.80	14.62	15.30
BLIP ViT-B (zero-shot) [47]	56.36	39.03	30.75	31.54	43.94	22.04	22.28	41.96	25.55	27.56	45.05	20.07	22.79
BLIP ViT-B [47]	64.31	40.71	36.70	39.93	46.97	26.96	30.84	48.40	24.74	42.57	52.61	18.19	21.45
BLIP ViT-L (zero-shot) [47]	58.18	44.29	23.87	25.13	47.61	18.17	19.96	46.79	19.58	21.07	49.50	14.93	16.50
BLIP ViT-L [47]	65.06	41.87	35.64	42.45	48.92	24.81	33.91	48.55	25.38	29.17	49.50	23.92	22.10
Visual Semantic Embedding models													
VSRN [49]	40.34	27.04	32.97	39.05	31.36	22.26	28.87	30.08	25.43	31.11	32.50	19.43	24.80
SAF [20]	40.11	30.90	22.96	27.84	33.37	16.80	22.87	32.50	18.97	23.78	34.03	15.16	19.69
SGR [20]	40.45	30.71	24.08	28.08	33.41	17.40	22.57	32.40	19.90	23.95	34.08	15.75	19.90
VSE∞ (BUTD region) [10]	42.46	31.57	25.65	30.74	35.61	16.13	20.45	34.17	19.52	23.51	36.48	14.08	17.28
VSE∞ (BUTD grid) [10]	44.07	30.22	31.43	36.68	35.26	19.99	25.00	35.70	18.99	23.52	38.75	12.07	15.82
VSE∞ (WSL grid) [10]	51.55	34.31	33.44	38.60	40.40	21.63	26.26	43.67	15.29	18.39	46.87	9.08	11.31

4.3. Ablation studies

4.3.1 Effects of EI scores

To verify our word selection with embedding-influence (EI) scores, we analyze the effects of using different methods: Random, High EI scores, Low EI scores. Random refers to randomly select a noun, and Large EI selects a noun with the largest EI score. To offset the effect of the changed word (i.e., target word), we construct new captions by deleting the source word, without replacement.

As shown in Table 5, low EI word deletion is the most effective way to fool models. On the other hand, High EI word deletion shows limited performance drops. This verifies our assumption that exploiting the word's influence on embedding feature can effectively confuse models. We believe manipulating words with low EI scores can be one effective way to test robustness of newly trained models.

4.3.2 Substituting more words

To further analyze the vulnerability of the VL models, we conduct experiments by replacing multiple words. We wonder if the model would confuse even when the original semantic meaning is more broken. Thus, we randomly select between 2 and 5 words and substitute them with words in Bert vocabulary. Since many captions are not long, words are not limited to nouns and are randomly selected.

We show the results in Table 6. Although it is presumed to be an easy task, meaningful performance degradation occurs in the entire model when multiple words are changed. When more than two words are substituted, large-scale VL pre-training models show more robust performance com-

Table 5: **Effects of using EI scores.** Deleting a source word with the lowest EI score shows the largest performance drop.

	COCO	Random Deletion			High EI Deletion			Low EI Deletion		
	R@1(↑)	R@1(↑)	drop rate(↓)	FR@1(↓)	R@1(↑)	drop rate(↓)	FR@1(↓)	R@1(↑)	drop rate(↓)	FR@1(↓)
CLIP ViT-B/32 (zero-shot) [60]	50.10	38.58	22.99	29.66	42.76	14.65	21.84	**36.04**	**28.06**	**32.30**
CLIP ViT-L/14 (zero-shot) [60]	56.04	42.54	24.09	30.4	48.58	13.31	20.42	**39.22**	**30.01**	**33.74**
BLIP ViT-B (zero-shot) [47]	70.54	45.58	35.38	40.54	57.14	19.00	25.80	**36.34**	**48.48**	**52.48**
BLIP ViT-B [47]	81.90	65.54	22.46	19.98	72.74	11.18	14.06	**59.28**	**27.62**	**30.10**
VSRN [49]	52.66	44.7	15.12	18.02	43.46	17.47	22.56	**38.56**	**26.78**	**29.36**
VSE∞ (BUTD region) [10]	58.02	34.2	41.05	45.58	40.58	30.06	38.06	**30.02**	**48.26**	**50.72**
VSE∞ (BUTD grid) [10]	59.40	34.3	42.26	46.46	39.92	32.79	39.78	**30.46**	**48.72**	**51.54**
VSE∞ (WSL grid) [10]	66.06	40.8	38.24	41.68	47.32	28.37	33.76	**36.56**	**44.66**	**47.14**

Top1: motorcyclist on permits white and blue developers illumination permits suspiciously.
Top2: it's a cloudy night for a ride on the motorcycle.

Top1: a train station sad an awning klan concurrency sad a train on the right platform.
Top2: a train platform with trains on the other side.

Top1: a baths light bent over wartime seo wasn.
Top2: a grey computer mouse and a silver metal key.

Top1: a person umm a partial as viscount ride a named board.
Top2: a man stands next to a very small plane.

Top1: three men in baseball uniforms simplicity fighting yourselves hospitals crowd shuttle them.
Top2: interesting antics by two men during a baseball game.

Figure 7: **Example of substituting multiple random words.** We further break the semantic meaning of the caption by randomly replacing four words. The examples show the cases that wrong captions are retrieved as top 1, evaluated with BLIP (ViT-B) [47]. We discover that the model could be confused by these nonsensical sentences, that human would not be confused with. We assume that the model is likely to pay more attention to specific words (e.g., "motorcycle" in the first image) than to understanding the language.

Table 6: **Image-to-Text retrieval on dataset with multiple words substitutions.** We generate captions by randomly replacing more words and add to COCO test set. The results are averaged over generations with three different random seeds. Recall@1 (R@1)(↑), drop rate(↓), False Recall@1 (FR@1)(↓) are shown. Models can confuse sentences even when the semantic meaning is more largely damaged.

	COCO	2 words substitution			3 words substitution			4 words substitution			5 words substitution		
	R@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1
Large-scale VL pre-training models													
CLIP ViT-B/32 (zero-shot) [60]	50.10	42.89	14.39	19.71	46.07	8.04	12.67	47.45	5.29	8.15	48.37	3.45	5.46
CLIP ViT-B/16 (zero-shot) [60]	52.44	45.35	13.52	19.07	48.43	7.65	11.89	49.97	4.71	8.01	50.61	3.49	5.95
CLIP ViT-L/14 (zero-shot) [60]	56.04	47.35	15.51	22.18	50.22	10.39	15.78	51.99	7.23	11.56	53.07	5.30	8.27
ALBEF [48]	77.58	72.43	6.64	2.40	73.03	5.86	0.88	73.23	5.61	0.43	73.26	5.57	0.32
BLIP ViT-B (zero-shot) [47]	70.54	53.04	24.81	30.75	62.99	10.70	14.72	67.95	3.67	5.44	69.73	1.15	1.86
BLIP ViT-B [47]	81.90	73.62	10.11	12.76	77.45	5.43	7.10	79.54	2.88	4.05	80.48	1.73	2.51
BLIP ViT-L (zero-shot) [47]	73.66	60.35	18.07	21.66	67.99	7.70	10.16	71.63	2.76	3.93	72.87	1.07	1.61
BLIP ViT-L [47]	82.36	73.93	10.24	12.65	77.93	5.38	7.45	79.81	3.10	4.23	80.98	1.68	2.54
Visual Semantic Embedding models													
VSRN [49]	52.66	45.07	14.41	17.79	47.89	9.06	11.33	49.89	5.26	7.08	50.99	3.17	4.29
SAF [20]	55.46	44.06	20.56	20.29	47.22	14.86	26.71	50.02	9.81	15.12	51.71	6.76	10.85
SGR [20]	57.22	43.57	23.86	28.53	46.98	17.90	22.79	49.81	12.95	17.49	51.91	9.28	13.09
VSE∞ (BUTD region) [10]	58.02	33.94	41.50	46.81	37.15	35.98	42.66	40.39	30.39	37.79	43.17	25.60	33.01
VSE∞ (BUTD grid) [10]	59.40	34.79	41.44	45.95	38.03	35.98	41.75	41.17	30.68	37.14	44.97	24.30	30.57
VSE∞ (WSL grid) [10]	66.06	39.95	39.52	43.79	44.04	33.33	38.44	48.29	26.90	32.85	51.73	21.69	27.51

pared to VSE models. Especially, VSE∞ shows the vulnerability even for captions with 5 words changed. We think that VSE∞'s simple pooling operator can be overfitted to COCO dataset.

Meanwhile, Figure 7 displays the examples of newly created captions which BLIP (ViT-B) has retrieved as top 1. The figure shows the results when four words are replaced. We observe that Top 1 retrieved caption includes at least one correct key word, such as "motorcyclist" in the first image. These results lead us to suspect that the model seems to be paying more attention to certain words than whole sentences. We include the examples of substituting two and

three words in Supplementary Material.

5. Conclusion

In this paper, we propose a robust-evaluation benchmark that can measure the robustness of image-text matching (ITM) models. To the best of our knowledge, it is the first benchmark to test robustness in image-text matching task. Unlike existing studies for the robustness test in computer vision and natural language processing (NLP) area, which generate semantic-preserving texts and images with imperceptible changes, we propose a strategy in the opposite direction to the existing adversarial attack strategy. Our main idea is to create fooling captions and images by minimal changes in embedding feature. From evaluation on various state-of-the-art vision language (VL) models, we discover that both models with and without large-scale pre-training data show significant performance degradation and retrieve the incorrect caption/image at a high rate. Our empirical results raise up necessity of new robust ITM models and our benchmark dataset could promote further robustness studies in ITM task.

Limitations. In the process of randomly replacing words, some unnatural sentences such as "A war on bicycle riding next to a train (man $\rightarrow$ war)" are created. However, these sentences do not violate our intention to test how well the ITM model understands both visual and semantic meaning. Creating benchmarks is a very challenging but important study that can boost improvements of the existing algorithms. We hope that our study can inspire researchers in ITM task and more robustness benchmarks can be created.

Supplementary Material

F. Text-to-Image Retrieval

We report the results on new image set with $\lambda = 0.7, 0.6$ in Table 7. We can still observe meaningful performance drop, when the added images are more significantly perturbed that seemed less confusing. In most cases, False Recall@1 exceeded 10%. In BLIP [47], performance degradation occurred more in fine-tuned models than in zero-shot models. We conjecture that this is because the models overfitted to COCO dataset during finetuning. We display the examples of retrieving incorrect images with BLIP ViT-B when $\lambda = 0.6, 0.7$ in Figure 8.

G. Substituting more words

Figure 9 shows examples of newly added captions that BLIP ViT-B model has retrieved as top 1. While the created captions are not natural, they include some keywords. Thus, we can conclude that the model is focusing on some nouns rather than the whole sentence.

Table 7: **Text-to-Image retrieval.** Models are evaluated with our new benchmark: Mix and Patch with different λ. Recall@1 (R@1)(↑), drop rate(↓), False Recall@1 (FR@1)(↓) are shown. The results are averaged over image generations with three different random seeds. We can see consistent degradation across all vision-language models.

	COCO 5K	Mix ($\lambda = 0.7$)			Mix ($\lambda = 0.6$)			Patch ($\lambda = 0.7$)			Patch ($\lambda = 0.6$)		
	R@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1	R@1	drop rate	FR@1
Large-scale VL pre-training models													
CLIP ViT-B/32 (zero-shot) [60]	30.14	25.24	16.25	19.16	26.87	10.84	14.34	25.18	16.45	20.27	25.96	13.86	17.97
CLIP ViT-B/16 (zero-shot) [60]	33.03	26.60	19.46	22.48	28.67	13.19	16.90	26.14	20.85	25.64	27.12	17.88	22.76
CLIP ViT-L/14 (zero-shot) [60]	36.14	30.45	15.75	18.59	32.01	11.43	15.17	30.33	16.08	21.16	30.96	14.34	19.30
ALBEF [48]	60.67	52.53	13.42	14.44	55.91	7.85	9.19	53.71	11.47	12.54	54.75	9.76	10.92
BLIP ViT-B (zero-shot) [47]	56.36	48.12	14.62	16.52	50.81	9.85	11.70	46.97	16.66	18.74	48.38	14.16	16.33
BLIP ViT-B [47]	64.31	53.56	16.72	20.15	57.77	10.18	13.45	55.20	14.17	16.79	56.68	11.87	14.82
BLIP ViT-L (zero-shot) [47]	58.18	51.21	11.98	13.82	53.69	7.71	9.38	51.05	12.25	13.96	52.28	10.14	11.95
BLIP ViT-L [47]	65.06	52.06	19.98	24.43	57.08	12.27	16.19	55.58	14.57	18.05	57.19	12.10	15.41
Visual Semantic Embedding models													
VSRN [49]	40.34	34.80	13.72	19.24	37.04	8.17	12.69	34.01	15.68	21.31	34.99	13.25	18.59
SAF [20]	40.11	35.55	11.37	16.57	36.91	7.98	12.32	35.22	12.20	16.81	35.75	10.87	15.24
SGR [20]	40.45	35.59	12.01	16.54	37.23	7.96	11.96	35.23	12.90	17.12	35.85	11.37	15.53
VSE (BUTD region) [10]	42.46	38.74	8.76	11.99	40.38	4.90	7.20	38.18	10.08	13.57	38.99	8.17	11.42
VSE (BUTD grid) [10]	44.07	39.01	11.47	15.39	41.22	6.46	9.26	40.10	9.00	12.12	40.94	7.09	9.94
VSE (WSL grid) [10]	51.55	45.13	12.46	15.70	47.97	6.95	9.30	48.37	6.17	8.02	48.93	5.08	6.58

(a) $\lambda = 0.6$

(b) $\lambda = 0.7$

Figure 8: **Image retrieval examples.**

Top 1: a pile disguised teddy star and dolls in a toy box.

Top 2: a pile of teddy bears and dolls in a toy box.

Top 1: arrival cat looking annoyance arrival large group of pigeons.

Top 2: a cat looking at a large group of pigeons.

Top 1: coincidentally metallic refrigerator freezer sitting in coincidentally muscles.

Top 2: two refrigerators side by side in a kitchen.

Top 1: a television that is on with a white man talking neighbors disgusting signs.

Top 2: a television that is on with a white man talking and campaign signs.

Top 1: a young zebra cynical an adult zebra standing on a elevators brown landscape.

Top 2: dishes zebra standing next to katie zebra in dishes dry grass field.

(a) Two words substitution

Top 1: daughters mound tease cake in their dining room while moms get marriott.

Top 2: daughters frosting a cake in their dining room while moms get water.

Top 1: rounding yellow fire hydrant eliot mommy sidewalk in an urban area.

Top 2: a yellow fire hydrant near the curb of a street.

Top 1: a she chalmers four women walking down a virtual in the rain.

Top 2: a group of four women walking down a street in the rain.

Top 1: a guy cutting off another guy inflicted dazzling from his assure.

Top 2: a guy cutting off another guy's cast from his arm.

Top 1: a detrimental player in peripheral much with striped shorts.

Top 2: undercover facto player swings his racket at undercover facto ball lithuania undercover facto court.

(b) Three words substitution

Top 1: grow boy sits freely bed beds leans over nigel metal laptop.

Top 2: a curious toddler reaches out to touch a laptop computer.

Top 1: mohamed glide grazing together fatally pissed green grassy appealing.

Top 2: two brown horses in a pasture eating grass.

Top 1: comb young curtains in comb pat uniform throwing comb ball.

Top 2: little league baseball player throwing a baseball from the mound.

Top 1: governmental male skate-boarder give governmental white fortification doing governmental conquer.

Top 2: a young skate board rider on top of a metal box.

Top 1: a skier violin projection shouting snow looking lecturer shouting profit.

Top 2: skier in austen flight mundane crossed skiis above composed nat sweating.

(c) Five words substitution

Figure 9: **Example of substituting multiple random words.**

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[2] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. 1, 2

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 2, 6

[4] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018. 2

[5] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012. 2

[6] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2022. 2

[7] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022. 2

[8] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022. 3

[9] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017. 2

[10] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 1, 2, 3, 4, 6, 7, 8, 10

[11] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, 2020. 2

[13] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. Vista: vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[14] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *8th Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST 2014*, 2014. 4

[15] Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, 2022. 2

[16] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1, 2

[17] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. 2

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019. 4, 5

[20] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, 2021. 2, 6, 7, 8, 10

[21] Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. Towards robustness against natural language word substitutions. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[22] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018. 2

[23] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2018. 2

[24] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 2013. 2

[25] Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[26] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of

logic. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 2020. 1, 3

[27] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[28] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 1, 3

[29] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2

[30] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 2

[31] Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: general robust image task benchmark. *arXiv preprint arXiv:2204.13653*, 2022. 4

[32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 2

[33] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 1, 2

[34] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems*, 2018. 2

[35] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[36] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[37] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL-HLT*, 2018. 1, 2

[38] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 2

[39] Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 2

[40] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021. 2

[41] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894, Sydney, Australia, 2017. PMLR. 3

[42] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 2017. 2

[43] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2

[44] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2, 6

[45] Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and William B Dolan. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. 2

[46] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. Universal perturbation attack against image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4899–4908, 2019. 2

[47] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 2022. 1, 2, 3, 4, 6, 7, 8, 9, 10

[48] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 2021. 2, 6, 7, 8, 10

[49] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 4, 6, 7, 8, 10

[50] Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020. 3

[51] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 1, 3

[52] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. In *Proceedings of the 2020 Conference on Em-*

pirical Methods in Natural Language Processing (EMNLP), 2020. 2

[53] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 2020. 2

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014. 1

[55] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15692–15701, 2022. 2

[56] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 6

[57] Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. Trivial or impossible — dichotomous data difficulty masks model differences (on imagenet and beyond). In *International Conference on Learning Representations*, 2022. 4

[58] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. *arXiv preprint arXiv:2004.15020*, 2020. 2

[59] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2, 3

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 2021. 1, 4, 6, 7, 8, 10

[61] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019. 2

[62] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 2015. 2

[63] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3

[64] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 2021. 1, 3

[65] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[66] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 2017. 2

[67] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017. 4

[69] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on nlp models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. 2

[70] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 2020. 2

[71] Yun Wang, Tong Zhang, Xueya Zhang, Zhen Cui, Yuge Huang, Pengcheng Shen, Shaoxin Li, and Jian Yang. Wasserstein coupled graph learning for cross-modal retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[72] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2

[73] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[74] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 5

[75] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 3

[76] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 5

[77] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text

matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15661–15670, 2022. 2

[78] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[79] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2