Towards Adversarial Attack on Vision-Language Pre-training Models

Jiaming Zhang∗
School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China jiamingzhang@bjtu.edu.cn

Qi Yi
School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China 21125273@bjtu.edu.cn

Jitao Sang†
1School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
2Peng Cheng Lab, China jtsang@bjtu.edu.cn

ABSTRACT
While vision-language pre-training model (VLP) has shown revolutionary improvements on various vision-language (V+L) tasks, the studies regarding its adversarial robustness remain largely unexplored. This paper studied the adversarial attack on popular VLP models and V+L tasks. First, we analyzed the performance of adversarial attacks under different settings. By examining the influence of different perturbed objects and attack targets, we concluded some key observations as guidance on both designing strong multimodal adversarial attack and constructing robust VLP models. Second, we proposed a novel multimodal attack method on the VLP models called Collaborative Multimodal Adversarial Attack (Co-Attack), which collectively carries out the attacks on the image modality and the text modality. Experimental results demonstrated that the proposed method achieves improved attack performances on different V+L downstream tasks and VLP models. The analysis observations and novel attack method hopefully provide new understanding into the adversarial robustness of VLP models, so as to contribute their safe and reliable deployment in more real-world scenarios. Code is available at https://github.com/adversarial-for-goodness/Co-Attack.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence.

KEYWORDS
Vision-and-Language Pre-training, multimodal, adversarial attack

1 INTRODUCTION
Vision-and-Language Pre-training (VLP) has revolutionized downstream Vision-and-Language (V+L) tasks very recently [6, 7, 22], such as image-text retrieval, visual grounding and visual entailment. This calls back attention to the research of multimodal neural networks over the last 10 years. While extensive studies have been conducted to achieve remarkable progress, only a few of them have investigated the adversarial robustness problem, which typically selects one single modality for perturbation to attack the multimodal task with standard adversarial attack method [20, 26, 28]. Regarding multimodal pre-training especially VLP models, so far as we know, there has been no study to systematically analyze the adversarial robustness performance and design a dedicated adversarial attack solution.

As still in its very early stage, many problems remain unexplored regarding the adversarial attack on VLP models. Among these, two critical issues are examined in this paper: (1) Standard adversarial attack is designed for classification tasks involved with only one modality. VLP models involve multiple modalities and typically many non-classification tasks like image-text cross-modal retrieval, making it impractical to directly employ the standard adversarial attack methods. A natural solution is to conduct adversarial attack on the embedding representation instead of the downstream task labels. However, due to the complex structure of VLP embedding representation, the problem turns to understanding how different attack settings affect the attack performance. (2) To attack the embedding representation of VLP model, the adversarial perturbations for different modalities should be considered collaboratively instead of independently. Figure 1 illustrates an example regarding adversarial attack on ALBEF [8] for the visual entailment task. It is shown that only perturbing image can successfully change the prediction.
Table 1: Attack settings along different attack targets and perturbed objects.

| Perturbed Object | Attack Target | Unimodal Embedding | Multimodal Embedding |
|------------------|---------------|--------------------|----------------------|
|                  |               | [CLS] of embedding | full embedding       |
|                  |               | Image@Uni_CLS      | Image@Uni_full       |
| Single-modal input | Image        | Text@Uni_CLS      | Text@Uni_full       |
| Bi-modal inputs  | Image&Text    | B1@Uni_CLS        | B1@Uni_full         |

from “entailment” to “contradiction”. By independently perturbing both image and text without considering their interaction, however, the attack fails as the two single-modal attacks may work in conflict with each other and lead to a countervailing $1 + 1 < 1$ effect. This study is devoted towards adversarial attack on VLP models, addressing these two issues by analyzing attack performance in different settings and developing collaborative multimodal adversarial attack solution.

For the first issue, we analyzed the adversarial attack on VLP models in different settings along two dimensions of attack target and perturbed object. Two typical VLP architectures, fused VLP models (e.g., ALBEF [8], TCL [27]) and aligned VLP models (e.g., CLIP [16]) are examined with 3 downstream V+L tasks of image-text retrieval, visual entailment, and visual grounding. The extensive analyses derive some key observations regarding the different attacking performances among VLP architectures and V+L tasks, as well as the influence of attack settings on VLP robustness.

For the second issue, we propose a novel multimodal adversarial attack method on the VLP models called Collaborative Multimodal Adversarial Attack (Co-Attack), which collectively carries out the attack on the image modality and the text modality. Co-Attack is applicable to both fused VLP and aligned VLP models. The premise is to encourage the perturbed multimodal embedding away from the original multimodal embedding (for fused VLP models), or the perturbed image-modal embedding away from the perturbed text-modal embedding (for aligned VLP models). Experimental results demonstrate that the proposed method improves attack performance on different V+L downstream tasks.

The contributions of the paper can be summarized as follows.

1. We analyzed the performance of adversarial attacks on two typical architectures of VLP models and three downstream V+L tasks. The observations regarding attacking settings along attack targets and perturbed objects shed light on understanding the adversarial robustness of VLP models. (2) We developed a novel multimodal adversarial attack method on the VLP models. By considering the consistency between attacks of different modalities, it collaboratively combines multimodal perturbations towards a stronger adversarial attack.

2 PRELIMINARIES AND RELATED WORK

2.1 VLP models and Downstream V+L Tasks

2.1.1 VLP models.

Most early work on vision-language representation learning required pre-trained object detectors (e.g., Faster R-CNN [18]) and high resolution images [10, 24]. Rather than relying on visual features derived by computation-costly object detectors, recent methods use an end-to-end image encoder (e.g., ViT [3]) to speed up inference [8, 16, 27]. In this work, we consider CLIP [16], ALBEF [8] and TCL [27] for evaluation, which belong to the latter class of models. Among them, ALBEF and TCL model the interactions between image and text modality, containing both the unimodal encoder $E_i(\cdot)$, $E_t(\cdot)$, and a multimodal encoder $E_m(\cdot, \cdot)$ (illustrated in Figure 2(a)). An input image $x_i$ is encoded into the image embedding $e_i$ by the image encoder $E_i(\cdot)$, i.e., $e_i = E_i(x_i)$. An input text $x_t$ is encoded into the text embedding $e_t$ by the text encoder $E_t(\cdot)$, i.e., $e_t = E_t(x_t)$. The image embedding $e_i$ and text embedding $e_t$ are then fed into the multimodal encoder $E_m(\cdot, \cdot)$ to output a unified multimodal embedding $e_m$, i.e., $e_m = E_m(e_i, e_t)$. We denote this type of VLP models with multimodal encoder and unified multimodal embedding as fused VLP model. In contrast, CLIP focuses on learning unimodal image encoder and text encoder without considering the multimodal encoder. We denote this type of VLP models with only separate unimodal embeddings as the aligned VLP model (illustrated in Figure 2(b)).

2.1.2 Downstream V+L Tasks.

Image-Text Retrieval contains two sub-tasks: image-to-text retrieval (TR) and text-to-image retrieval (IR). For ALBEF and TCL, both for TR and IR, the feature similarity score is calculated at first between $e_i$ and $e_t$ for all image-text pairs to retrieve the Top-N candidates, and then the image-text matching score calculated by $e_m$ is used for ranking. TR and IR tasks on CLIP are carried out more directly. The ranking results are only based on the similarity between $e_i$ and $e_t$.

Visual Entailment (VE) is a visual reasoning task to predict whether the relationship between an image and a text is entailment, neutral, or contradiction. Both ALBEF and TCL consider VE as a three-way classification problem, and predict the class probabilities using a fully layer on the multimodal encoder’s representation of the [CLS] token [8].

Visual Grounding (VG) localizes the region in an input image based on the description of the corresponding input text. ALBEF extends Grad-CAM [19], and uses the derived attention map to rank the detected proposals [29].

2.2 Adversarial Attack

2.2.1 In Computer Vision and Natural Language Processing.

Adversarial attack is first proposed in computer vision, which demonstrates the vulnerability of deep learning models [23]. The most common adversarial attacks in computer vision are gradient-based methods such as FGSM [4], PGD [13], MIM [2] and SI [11]. To obtain the image adversarial perturbations $\delta_i$ corresponding to input image $x_i$, we consider $\ell_p$-norm constrained perturbation in this work, where $\delta_i$ satisfies $\|\delta_i\|_p \leq \varepsilon_i$ with $\varepsilon_i$ being the maximum perturbation magnitude.
In computer vision, adversarial attack can be easily achieved by gradient-based methods in the continuous space. Unlike this, in natural language processing, adversarial attack in the discrete space is still a challenge. Current successful adversarial attacks in natural language processing usually modify or replace some tokens of input text $x_t$ to maximize the risk of outputting wrong embedding [9, 17].

2.2.2 In Multimodal Neural Network.
Regarding VLP models, so far as we know, there has been no related work to systematically analyze and design adversarial attack. Therefore, we discuss and employ these available adversarial attacks on multimodal neural networks involving V+L tasks as the baseline methods for comparison. Xu et al. [26] investigated attacking the visual question answering model by perturbing the image modality. Agrawal et al. [1] and Shah et al. [20] attempted to attack the vision-and-language model by perturbing the text modality. Yang et al. [28] studied the robustness of various multimodal models and proposed a defense method. We summarize three main differences between these multimodal attack studies and this work. First, the models in the above studies are based on the convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The methods and observations are not readily applicable to the transformer-based VLP models. Second, the above studies mainly addressed single V+L classification tasks, and the attacks used are not generalized to other non-classification tasks. Third, the adversarial attack methods they employed are essentially standard unimodal attack, while multimodal attacks specially for multimodal models remain unexplored.

3 ANALYZING ADVERSARIAL ATTACK IN VLP MODEL

3.1 Analysis Settings

3.1.1 Examined VLP Models, Downstream Tasks and Datasets.

Examined VLP Models. We evaluated the analyses and experiments on two types of models: the fused VLP model and the aligned VLP model. The fused VLP model contains an image encoder, a text encoder, and a multimodal encoder, and in this work we consider ALBEF [8] and TCL [27]. The image encoder is implemented by a 12-layer visual transformer ViT-B/16 [3]. Both the text encoder and the multimodal encoder are implemented by a 6-layer transformer. The aligned VLP model contains only image encoder and text encoder, and in this work we consider CLIP. There are different choices of image encoder for CLIP. We consider ViT-B/16 and ResNet-101 [5], denoted as CLIP\textsubscript{ViT} and CLIP\textsubscript{CNN}, respectively.

Downstream Tasks and Datasets. In this work, MSCOCO [12] and Flickr30K [15] are used to evaluate TR and IR tasks, and RefCOCO+ [30] is used to evaluate VG task, and SNLI-VE [25] is used to evaluate VE task. Note that CLIP can only tackle IR and TR tasks, TCL can handle VE, IR and TR tasks, and ALBEF can deal with all the above downstream tasks. For the VE task, since we are concerned with the performance of adversarial attack, we select only positive image-text pairs (with label of entailment) and discard the negative pairs (with label of neutral and contradictory) from the SNLI-VE testing dataset.
Table 2: Image-text retrieval results of ALBEF on Flickr30K and MSCOCO dataset. The reported value is attack success rate where a larger value indicates a stronger attack.

| Attack          | Flickr30K (1K test set) | MSCOCO (5K test set) |
|-----------------|--------------------------|----------------------|
|                 | TR | IR | TR | IR | TR | IR |
|                 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| Text@UniCLS     | 10.10 | 1.70 | 1.10 | 24.12 | 13.84 | 10.78 |
|                 | 21.40 | 15.94 | 10.46 | 63.40 | 66.18 | 61.54 |
| Image@UniCLS    | 48.70 | 36.30 | 29.40 | 48.88 | 42.78 | 37.08 |
|                 | 45.56 | 41.34 | 36.86 | 48.56 | 48.25 | 43.18 |
| Bi@MultiCLS     | 64.80 | 50.60 | 43.20 | 62.40 | 57.80 | 52.90 |
|                 | 65.02 | 69.56 | 66.34 | 63.40 | 66.18 | 61.54 |
| Text@MultiCLS   | 13.70 | 2.70 | 1.30 | 24.02 | 15.10 | 11.46 |
|                 | 26.82 | 16.92 | 11.50 | 27.74 | 25.78 | 21.75 |
| Image@MultiCLS  | 20.50 | 13.90 | 11.20 | 27.36 | 25.16 | 22.32 |
|                 | 24.43 | 33.68 | 34.34 | 24.43 | 33.68 | 34.34 |
| Bi@MultiCLS     | 35.90 | 24.30 | 19.70 | 44.54 | 41.34 | 36.86 |
|                 | 41.86 | 41.62 | 37.66 | 38.10 | 48.10 | 48.24 |
| Text@Unifull    | 7.30 | 1.10 | 0.60 | 16.02 | 10.72 | 8.64 |
|                 | 15.86 | 9.40 | 6.24 | 19.70 | 18.85 | 16.49 |
| Image@Unifull   | 59.20 | 48.60 | 42.70 | 56.54 | 54.96 | 50.54 |
|                 | 58.40 | 61.88 | 59.18 | 60.54 | 65.78 | 62.72 |
| Bi@Unifull      | 66.40 | 55.00 | 49.60 | 63.76 | 62.68 | 58.18 |
|                 | 65.02 | 69.56 | 66.34 | 51.79 | 66.19 | 67.35 |
| Text@Multiuniform | 7.70 | 1.40 | 1.00 | 16.78 | 11.40 | 8.98 |
|                 | 15.00 | 8.72 | 5.74 | 19.11 | 18.39 | 15.92 |
| Image@Multiuniform | 60.00 | 50.20 | 42.90 | 58.74 | 56.90 | 52.02 |
|                 | 54.30 | 58.58 | 55.32 | 60.54 | 65.78 | 62.72 |
| Bi@Multiuniform | 65.70 | 55.40 | 48.10 | 62.92 | 62.80 | 58.28 |
|                 | 65.04 | 65.78 | 62.72 | 51.08 | 65.86 | 67.09 |

classification are gradient-based methods. Among them, we selected FGSM[4], which requires only one gradient calculation, to represent this family of methods:

$$\delta_i = \epsilon_i \cdot \text{sign}(\nabla_{x_i} \mathcal{L}(C(x_i), \hat{y})), \quad (1)$$

where $C$ denotes the classification model, $\mathcal{L}$ is cross-entropy loss, and $\hat{y}$ is ground-truth label corresponding to the original image $x_i$. The cross-entropy loss requires logit-wise representation, but many V+L downstream tasks are non-classification tasks, such as image-text retrieval. So we used the method proposed by Zhang et al. [31] that maximizes KL (Kullback-Leibler) divergence loss $\mathcal{L}$ of embedding-wise representation to conduct adversarial attack:

$$\delta_i = \epsilon_i \cdot \text{sign}(\nabla_{x_i} \mathcal{L}(E_i(x_i), \hat{y})), \quad (2)$$

For perturbing the text-modal input, $T(\cdot)$ represents the modification or replacement of tokens in the input text $x_i$, i.e., $x_i' = T(x_i)$. Then the text adversarial perturbation $\delta_t$ can be represented as follows:

$$\delta_t = \arg \max_{x_t} \|E_t(x_t') - E_t(x_t)\| - x_t, \quad (3)$$

where the maximum perturbation $\epsilon_t$ is constrained to the token level, i.e., how many tokens are modified/replaced with semantic-consistent. In this work, we use BERT-Attack [9] to conduct adversarial attack on text modality.

**Attacking multimodal embedding.** For the attack on the input text, we substitute the text embedding $E_t(\cdot)$ in Equation (3) with the multimodal embedding $E_m(\cdot)$:

$$\delta_t = \arg \max_{x_t} \|E_m(E(x_i), E_t(x_i)) - E_m(E(x_i), E_t(x_i))\| - x_t. \quad (4)$$

Similarly, for the attack on the input image, we substitute the image embedding $E_i(\cdot)$ in Equation (2) with the multimodal embedding $E_m(\cdot)$:

$$\delta_i = \epsilon_i \cdot \text{sign}(\nabla_{x_i} \mathcal{L}(E_m(E(x_i), E_t(x_i)), E_m(E_i(x_i), E_t(x_i)))), \quad (5)$$

### 3.3 Observations

**Observations on Image-Text Retrieval.** The results of ALBEF on the TR and IR tasks shown in Table 2. In Table 2, we have the following main findings: (1) Perturbing bi-modal inputs (Bi@) is consistently stronger than perturbing any single-modal input (both Text@ and Image@). This suggests that stronger adversarial attacks are expected if multiple modal inputs are allowed to be perturbed. (2) For perturbing image-modal input, attacking full embedding outperforms attacking [CLS] of embedding. But for perturbing text-modal input, we observed the opposite result. This suggests that for text encoders, the [CLS] of embedding represents the sentence-level significance, and attacking it has a greater impact than attacking the full embedding. However, for image encoders, attacking the full embedding has a more significant impact, which is consistent with the observation in [14]. They replaced the [CLS] of embedding with a "global average pooling", and found that removing the [CLS] of embedding contributes a negligible influence on accuracy and robustness. (3) The attack performance of Bi@Multiuniform is much better than Bi@MultiCLS. This indicates that for perturbing bi-modal inputs to attack multimodal embedding, the influence of [CLS] of embedding is limited and even weakens the adversarial attack. (4) Bi@Multi and Bi@Uni achieve the similar performance. This suggests that for tasks like TR and IR that require intermediate outputs (unimodal embedding) involved for inference, the impact of attacking unimodal embedding and attacking full embedding is comparable.

The results of attacking CLIPViT and CLIPCNN are shown in Table 3 and Table 4, respectively. Main observations include: (1) It is found that despite the large difference in the architecture, consistent with the observation in ALBEF, perturbing bi-modal inputs is stronger than perturbing any single-modal input. (2) Both for perturbing image-modal input and perturbing bi-modal inputs, the attack success rate on CLIPCNN (Table 4) is higher than CLIPViT (Table 3). But it is hard to tell the difference between CLIPCNN and CLIPViT for perturbing text-modal input. This suggests that ViT is a more robust image encoder than ResNet-101 (CNN) on
Table 3: Image-text retrieval results of CLIP\_\text{ViT} on Flickr30K and MSCOCO dataset. The reported value is attack success rate.

| Attack | Flickr30K (1K test set) | MSCOCO (5K test set) |
|--------|-------------------------|-----------------------|
|        | TR                     | IR                    | TR                      | IR                  |
|        | R@1  R@5  R@10         | R@1  R@5  R@10       | R@1  R@5  R@10         | R@1  R@5  R@10     |
| Text@\text{Uni}\_\text{CLS} | 19.60 10.50 6.30 | 20.90 19.00 14.90 | 24.92 25.26 22.08 | 16.61 23.68 24.34 |
| Image@\text{Uni}\_\text{CLS} | 56.20 49.90 42.20 | 43.68 47.00 42.58 | 41.34 52.68 52.74 | 26.28 40.79 44.47 |
| Bi@\text{Uni}\_\text{CLS} | 64.00 60.90 52.90 | 51.08 60.34 58.54 | 46.34 62.10 64.34 | 29.75 48.78 54.88 |

Table 4: Image-text retrieval results of CLIP\_\text{CNN} on Flickr30K and MSCOCO dataset. The reported value is attack success rate.

| Attack | Flickr30K (1K test set) | MSCOCO (5K test set) |
|--------|-------------------------|-----------------------|
|        | TR                     | IR                    | TR                      | IR                  |
|        | R@1  R@5  R@10         | R@1  R@5  R@10       | R@1  R@5  R@10         | R@1  R@5  R@10     |
| Text@\text{Uni}\_\text{CLS} | 18.50 10.90 6.30 | 20.40 20.46 17.14 | 22.40 24.00 22.16 | 15.77 24.08 25.25 |
| Image@\text{Uni}\_\text{CLS} | 63.30 65.60 59.70 | 51.58 64.24 63.70 | 43.12 58.62 61.62 | 27.56 46.67 52.87 |
| Bi@\text{Uni}\_\text{CLS} | 70.50 73.50 69.10 | 54.10 70.96 71.20 | 46.68 65.62 70.82 | 28.97 50.74 58.95 |

Table 5: Visual entailment results of ALBEF and TCL on SNLI-VE dataset. The reported value is attack success rate.

| Model       | Text@\text{Uni}\_\text{CLS} | Image@\text{Uni}\_\text{CLS} | Bi@\text{Uni}\_\text{CLS} | Text@\text{Multi}\_\text{CLS} | Image@\text{Multi}\_\text{CLS} | Bi@\text{Multi}\_\text{CLS} | Text@\text{Uni}\_\text{full} | Image@\text{Uni}\_\text{full} | Bi@\text{Uni}\_\text{full} | Text@\text{Multi}\_\text{full} | Image@\text{Multi}\_\text{full} | Bi@\text{Multi}\_\text{full} |
|-------------|-------------------------------|-------------------------------|-----------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| ALBEF       | 26.65                         | 25.52                         | 38.99                       | 34.36                         | 39.51                         | 50.90                         | 20.74                         | 31.78                         | 39.53                         | 33.17                         | 42.28                         | 52.72                         |
| TCL         | 28.34                         | 29.78                         | 42.39                       | 37.00                         | 32.33                         | 48.87                         | 27.20                         | 18.98                         | 36.60                         | 37.23                         | 32.91                         | 49.92                         |

defending the image-modal attack, which is also consistent with some existing observations [21]. (3) Note that the image encoder for ALBEF (Table 2) and CLIP\_\text{ViT} (Table 3) are the same (ViT-B/16). By comparing the differences in the attack success rates between them, we can observe that there is no distinct superiority or inferiority between attacking the unimodal embedding of the ALBEF and CLIP\_\text{ViT}. This demonstrates that regarding different VLP models, different pre-training objectives of ALBEF (fused VLP model) and CLIP\_\text{ViT} (aligned VLP model) have no significant impact on the adversarial robustness.

Observations on Visual Entailment. The results of ALBEF and TCL on the VE task are shown in Table 5. We have the following main findings: (1) Consistent with the results on the previous tasks, perturbing bi-modal inputs is stronger than perturbing any single-modal input, and attacking full embedding outperforms attacking [CLS] of embedding for perturbing image-modal input. (2) Bi@\text{Multi}\_\text{full} achieved the best performance in all settings of attacks. This further demonstrates the superiority of attacking multimodal embedding.

Observations on Visual Grounding. The results of ALBEF on VG task are shown in Table 6. We have the following main findings: (1) Consistent with the results on the previous tasks, perturbing bi-modal inputs is stronger than perturbing any single-modal input. This suggests that stronger adversarial attacks are expected if multiple modal inputs are allowed to be perturbed. In addition, for perturbing the single-modal input, [CLS] of embedding has less impact on image modality compared to the text modality. For perturbing the bi-modal inputs, attacking [CLS] of embedding is limited and even harmful compared to attacking the full embedding. (2) Regarding the observations on different downstream V+L tasks, for the tasks like image-text retrieval that require intermediate outputs (unimodal embedding) involved in the inference, the result of attacking the full embedding is consistent with attacking [CLS] of embedding. But overall, attacking the multimodal embedding has a superior or comparable performance than the unimodal embedding along all tasks.

Summary. We concluded some shared observations as guidance on designing multimodal adversarial attacks in VLP models. (1) Regarding the observations on different attack settings, perturbing bi-modal inputs is stronger than perturbing any single-modal input. This suggests that stronger adversarial attacks are expected if multiple modal inputs are allowed to be perturbed. In addition, for perturbing the single-modal input, [CLS] of embedding has less impact on image modality compared to the text modality. For perturbing the bi-modal inputs, attacking [CLS] of embedding is limited and even harmful compared to attacking the full embedding. (2) Regarding the observations on different downstream V+L tasks, for the tasks like image-text retrieval that require intermediate outputs (unimodal embedding) involved in the inference, the result of attacking the full embedding is consistent with attacking [CLS] of embedding. But overall, attacking the multimodal embedding has a superior or comparable performance than the unimodal embedding along all tasks.
Table 6: Visual grounding results of ALBEF on RefCOCO+ dataset. The reported value is attack success rate.

| Subset | Text@Uni CLS | Image@Uni CLS | Bi@Uni CLS | Text@Multi CLS | Image@Multi CLS | Bi@Multi CLS | Text@Uni full | Image@Uni full | Bi@Uni full | Text@Multi full | Image@Multi full | Bi@Multi full |
|--------|-------------|---------------|-----------|----------------|----------------|-------------|--------------|---------------|-------------|----------------|----------------|-------------|
| Val    | 10.45       | 5.313         | 14.85     | 9.27           | 12.81          | 16.77       | 8.44         | 12.03         | 16.50       | 9.82           | 15.97          | 19.21       |
| TestA  | 12.08       | 6.00          | 17.35     | 11.84          | 14.30          | 19.96       | 10.18        | 12.55         | 18.87       | 10.79          | 16.59          | 22.33       |
| TestB  | 7.18        | 3.50          | 10.41     | 6.97           | 10.04          | 10.96       | 5.17         | 10.04         | 11.70       | 6.81           | 11.94          | 13.99       |

\[
\delta_t \quad \delta_i
\]

Vanilla Co-Attack \((x_i, x_t)\) \((x_i', x_t')\) \((x_i', x_t)\) \((x_i', x_t')\) \((x_i', x_t')\)

\[
\beta_v \quad \beta_c \quad \delta_i \quad \delta_t \quad \delta_i \quad \delta_t
\]

(a) multimodal embedding space (for fused VLP model) (b) unimodal embedding space (for aligned VLP model)

Figure 3: Illustration of Co-Attack. \(\delta_{i,t}\) represents the resultant perturbation from image-modal perturbation \(\delta_i\) and text-modal perturbation \(\delta_t\) in embedding space. (a) In multimodal embedding space, \(\beta\) represents the angle between \(\delta_i\) and \(\delta_t\). Co-Attack aims to reduce \(\beta\) and enlarge \(\delta_{i,t}\). (b) In unimodal embedding space, \(\gamma\) represents the angle between \(\delta_i\) and \(\delta_t\). Co-Attack aims to enlarge both \(\gamma\) and \(\delta_{i,t}\).

Figure 4: Statistical results on SNLI-VE of the angles between the image-modal attack and text-modal attack in the embedding space. (a) and (b) denote vanilla attack in multimodal embedding space and unimodal embedding space, respectively. (c) and (d) denote Co-Attack in multimodal embedding space and unimodal embedding space, respectively.

In addition, we derived some insights on constructing robust VLP models. Regarding the observations on different VLP models, the pre-training objectives of ALBEF (fused VLP model) and CLIP\_ViT (aligned VLP model) contribute no apparently different impact on the adversarial robustness. For the models with alternative image encoders, such as CLIP, ViT is a superior choice compared to CNN for image encoder in terms of accuracy and robustness.

4 COLLABORATIVE MULTIMODAL ADVERSARIAL ATTACK IN VLP MODEL

4.1 Methodology

Although the above analysis found that perturbing the textual and visual modalities simultaneously is more effective than perturbing one modality singly. However, as discussed in Figure 1, there is a risk of leading to a countervailing \(1+1<1\) effect from attacking both modalities independently. We address this issue by developing a collaborative multimodal adversarial attack solution called collaborative multimodal adversarial attack (Co-Attack). This allows us to collectively carry out the attack on the image modality and the text modality. The aim of Co-Attack is to encourage the perturbed multimodal embedding away from the original multimodal embedding, or the perturbed image-modal embedding from the perturbed text-modal embedding. Since Co-Attack can be adapted for attacking both multimodal and unimodal embeddings, it is applicable to both fused VLP and aligned VLP models.

4.1.1 Attacking Multimodal Embedding. For attacking multimodal embedding, Co-Attack attempts to collaboratively perturb the input text and input image, which encourages the perturbed multimodal embedding away from the original multimodal embedding. Figure 3(a) shows that without considering the consistency between
the two attacks, the vanilla attack yields smaller resultant perturbation $\delta_{\text{Img}}$ and greater angle $\beta_0$ between text-modal perturbation $\delta_t$ and image-modal perturbation $\delta_i$. In contrast, Co-Attack collectively carries out the text-modal perturbation $\delta_t$ and image-modal perturbation $\delta_i$, achieving stronger resultant perturbation $\delta_{\text{Img}}$ and smaller angle $\beta_c$. As shown in Figure 4(a) and Figure 4(c), $\beta_c$ yielded by vanilla attack is distributed over $[0, \pi]$, and Co-Attack narrows the distribution and makes $\beta_c$ mainly distributions around $[0, \frac{\pi}{2}]$.

Next, we elaborate on how to realize Co-Attack. The main challenge of implementing a collaborative adversarial attack lies in the input representation gap between the continuous image modality and the discrete text modality. To solve this problem, we use a step-wise scheme that first perturbs the discrete inputs (text) and then perturbs the continuous inputs (image) given the text perturbation. The reason we first perturb text-modal input is that it is difficult to optimize the designed objective in discrete space. So we start with a text-modal perturbation and use it as a criterion, and then proceed to the image-modal perturbation. The adversarial text $x'_t$ can be derived by Equation (4). The adversarial attack on image modality is conducted by:

$$\max \mathcal{L}(E_t(x'_t), E_t(x_t)) + \alpha_1 \cdot \mathcal{L}(E_t(x'_t), E_t(x_t)), \quad (6)$$

where the second term corresponds to $\delta_{\text{Img}}$ in Figure 3(a), and $\alpha_1$ is a hyper-parameter that controls the contributions of the second term. The sensitivity to $\alpha_1$ is discussed in Section 4.2.3. The above optimization problem can be easily solved by PGD-like procedures.

4.1.2 Attacking Unimodal Embedding. For attacking unimodal embedding, Co-Attack attempts to encourage the perturbed image-modal embedding away from the perturbed text-modal embedding. It should be noted that the unimodal embedding space is slightly different from the multimodal embedding space. As shown in Figure 3(b), the image-text sample pair corresponds to two different representations in the unimodal embedding space that are close together rather than one shared representation in the multimodal embedding space. It is shown that without considering the consistency between the two attacks, the vanilla attack yields both smaller resultant perturbation $\delta_{\text{Img}}$ and angle $\gamma_v$ between the text-modal perturbation $\delta_t$ and image-modal perturbation $\delta_i$. In contrast, Co-Attack collectively carries out the text-modal perturbation $\delta_t$ and image-modal perturbation $\delta_i$, achieving both greater resultant perturbation $\delta_{\text{Img}}$ and angle $\gamma_v$. As shown in Figure 4(b) and Figure 4(d), compared to the the $\gamma_v$ yielded by vanilla attack, Co-Attack shifts the $\gamma_v$ to a greater distribution.

Similar to attacking multimodal embedding, we first perturb input text to derive the adversarial text $x'_t$ according to Equation (2). Then similar to Equation (6), we perform the adversarial attack on image modality as follows:

$$\max \mathcal{L}(E_t(x'_t), E_t(x_t)) + \alpha_2 \cdot \mathcal{L}(E_t(x'_t), E_t(x_t)), \quad (7)$$

where the second term corresponds to $\delta_{\text{Img}}$ in Figure 3(b), and $\alpha_2$ is a hyper-parameter that controls the contributions of the second term. The sensitivity to $\alpha_2$ is discussed in Section 4.2.3.

4.2 Experiments

In this subsection, we used the same experimental settings as in Section 3.1. The momentum item for MIM is set to 0.9, and the number of scale copies for SI is set to 5. We set both $\alpha_1$ and $\alpha_2$ to 3.

4.2.1 Comparison Results. To demonstrate the performance improvement of our methods, we compared the proposed Co-Attack with 5 baseline attacks as follows.

- **Fooling VQA** is an attack method for the classification problem (VQA model), which uses ADAM optimizer to solve the cross-entropy loss to add image noises [26].
- **Single-Source Adversarial Perturbations (SSAP)** is used by Yang et al. [28] to evaluate adversarial robustness of VLP models, which is implemented by PGD to solve the cross-entropy loss to add image noises.
- **SSAP-MIM and SSAP-SI** are two baseline methods that replace PGD by introducing the more advanced MIM [2] and SI [11] respectively, since the optimization algorithms in SSAP are replaceable.
- **Vanilla** is the strongest attack analyzed in Section 3, which is also introduced as a baseline method, such as B1@Multi.full for the ALBEF model on the VE task.

Similarly, Co-Attack can also replace the PGD in the optimization algorithm with the SI, denoted as Co-Attack-SI. For a fair comparison, for the task that does not use cross-entropy, i.e., image-text retrieval, our method and all baseline methods attack embedding-wise representation as the target. For tasks that use cross-entropy, i.e. VE and VG, our method and all baseline methods attack logit-wise representation as the target.

Table 7 shows the comparison results for the image-text retrieval task. Table 8 presents the comparison results for the VE task. We can observe that: (1) Since Vanilla employs the strongest attack settings observed from our analysis in Section 3, Vanilla achieves basically superior attack performance over other baselines. This validates the reliability of the analysis results. (2) Co-Attack outperforms all
Table 7: Comparison results on image-text retrieval. The reported value is attack success rate.

| Model       | Attack       | Flickr30K (1K test set) | MSCOCO (5K test set) |
|-------------|--------------|-------------------------|----------------------|
|             |              | TR | IR | TR | IR |
|             |              | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| ALBEF       | Fooling VQA  | 12.80 | 4.70 | 3.10 | 12.72 | 6.62 | 4.40 | 39.68 | 31.50 | 24.82 | 34.23 | 34.73 | 29.70 |
|             | SSAP-MIM     | 66.10 | 55.30 | 49.60 | 63.54 | 64.82 | 60.62 | 58.84 | 64.18 | 61.40 | 41.44 | 52.80 | 53.34 |
|             | SSAP-SI      | 61.20 | 51.50 | 45.70 | 60.52 | 60.94 | 57.42 | 48.72 | 51.64 | 48.56 | 42.95 | 53.95 | 54.29 |
|             | Vanilla      | 70.20 | 61.50 | 54.70 | 66.70 | 67.90 | 64.32 | 58.08 | 62.64 | 59.16 | 49.05 | 62.62 | 63.57 |
|             | Co-Attack    | 65.70 | 55.40 | 48.10 | 62.92 | 62.80 | 58.28 | 60.54 | 65.78 | 62.72 | 51.08 | 65.86 | 67.09 |
|             | Co-Attack-SI | 70.60 | 60.50 | 53.50 | 67.22 | 67.10 | 62.76 | 58.84 | 64.18 | 61.40 | 51.13 | 66.17 | 67.73 |
| CLIPViT     | Fooling VQA  | 9.90 | 5.60 | 3.40 | 8.26 | 5.54 | 4.20 | 14.52 | 13.68 | 12.10 | 6.58 | 8.55 | 8.02 |
|             | SSAP-MIM     | 56.20 | 49.90 | 42.20 | 43.68 | 47.00 | 42.58 | 41.44 | 52.80 | 53.34 | 26.70 | 41.51 | 45.43 |
|             | SSAP-SI      | 48.10 | 42.20 | 35.30 | 38.40 | 39.42 | 35.52 | 38.40 | 48.10 | 47.70 | 24.37 | 37.47 | 40.74 |
|             | Vanilla      | 58.90 | 53.90 | 46.70 | 46.72 | 53.54 | 51.20 | 43.44 | 57.06 | 58.50 | 27.22 | 43.38 | 48.33 |
|             | Co-Attack    | 64.00 | 60.90 | 52.90 | 51.08 | 60.34 | 58.54 | 46.34 | 62.10 | 64.34 | 29.75 | 48.78 | 54.88 |
|             | Co-Attack-SI | 73.80 | 79.50 | 74.90 | 58.14 | 75.52 | 77.38 | 50.98 | 72.34 | 78.12 | 32.31 | 55.66 | 64.67 |

Table 8: Comparison results on visual entailment. The reported value is attack success rate.

| Model       | Attack       | Fooling VQA | SSAP | SSAP-MIM | SSAP-SI | Vanilla | Co-Attack | Co-Attack-SI |
|-------------|--------------|-------------|------|----------|---------|---------|-----------|--------------|
| ALBEF       |              | 58.21       | 77.30 | 76.58    | 70.18   | 70.57   | 79.27     | 76.33        |
| TCL         |              | 51.88       | 72.47 | 71.80    | 68.89   | 66.46   | 76.43     | 74.94        |

Figure 6: The impact of $\alpha_1$ and $\alpha_2$. (a) The visual entailment result of ALBEF with different $\alpha_1$. (b) The image-text retrieval result of CLIPViT with different $\alpha_2$.

4.2.2 Visualization Results. To understand Co-Attack more intuitively, we provide the Grad-CAM visualizations for the VQ task from ALBEF on RefCOCO+ dataset in Figure 5. Grad-CAM visualizations reflect the heat map where model would look when making decisions [19]. We noticed that perturbing single-modal input hardly changes the heat map. Vanilla slightly shifts the heat map, yet remaining noticeable regions on the object of interest. Co-Attack makes the model focus on the regions that deviate from the ground-truth so as to mislead the inference result.

4.2.3 Ablation Study. We conducted the ablation experiments to study the impact of $\alpha_1$ in Equation (6) and $\alpha_2$ in Equation (7). We adjusted $\alpha_1$ and $\alpha_2$ within the range of $[0, 5]$ with a step of 1 and examined their influences on ALBEF and CLIPViT, respectively. The visual entailment result of ALBEF on SNLI-VE dataset is shown in Figure 6(a), and the image-text retrieval result of CLIPViT on COCO is shown in Figure 6(b). It is shown that as $\alpha_1 > 0$ and $\alpha_2 > 0$, the attack performance becomes stronger. This demonstrates the importance of the second term in Equation (6) and Equation (7). The results are comparable when $\alpha_1 \geq 1$ and $\alpha_2 \geq 1$, indicating that Co-Attack is not sensitive to hyper-parameters and does not require elaborate tuning for the hyper-parameters.

5 CONCLUSION

In this paper, we investigated the adversarial attack on VLP models. First, we analyzed the performance of the adversarial attacks with different attack settings. Regarding the derived observations, we conclude the insights on designing the multimodal adversarial attacks and improving robustness in VLP models. Second, we developed a novel multimodal adversarial attack for VLP models.

We hope this study can draw attention on the distinct properties of adversarial robustness of multimodal models.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (Grant No. 2018AAA0100604), the National Natural Science Foundation of China (Grant No. 61832002, 62172094), and Beijing Natural Science Foundation (No. JQ20023). The computing resources of Pengcheng Cloudbrain are used in this research.
