International Research Journal of Multidisciplinary Sciences VOL-1 ISSUE-10 Oct 2025 page:7-12 ISSN:3107-930X(ONLINE)

Cross-Modal Mutual Search for Attacking Vision-Language Models

[1]BhawnaKaushik, [2]PriyaGupta

1. bhawna.kaushik@niu.edu.in,NoidaInternationalUniversity **2**. priya.gupta@niu.edu.in,NoidaInternationalUniversity

Abstract: Vision Language Pre Training (VLP) models like CLIP have demonstrated remarkable success in numerous downstream tasks by learning a unified representation space for images and text. However, their robustness against adversarial attacks remains a critical concern. Existing attacks often target a single modality (vision or language) or use simple, one directional fusion strategies, failing to fully exploit the intricate cross modal interactions that VLP models rely on. This paper introduces a novel Feedback Based Modal Mutual Search (FB MMS) framework for generating highly potent adversarial examples. Unlike conventional methods, FB MMS treats the attack on the image and text modalities as a cooperative search process. It leverages the model's own feedback—specifically, the gradient signals and loss computations from both modalities—to iteratively and mutually guide the perturbation of each. The image attacker uses text gradient information to refine its perturbations, and vice versa, creating a synergistic cycle that efficiently uncovers adversarial pairs in the joint embedding space. Our experiments on tasks such as image text retrieval and visual question answering demonstrate that FB MMS significantly outperforms strong unimodal and fused modal baselines, achieving higher attack success rates with smaller perceptual perturbations. This work highlights a fundamental vulnerability in VLP models and paves the way for more robust multimodal learning.

Keywords: Adversarial Attacks, Vision Language Models, Multimodal Learning, CLIP, Robustness, Gradient Feedback.

1. Introduction

The advent of large scale Vision Language Pre Training (VLP) models, such as CLIP [1] and ALBEF [2], has been a paradigm shift in artificial intelligence. By aligning visual and linguistic representations in a shared embedding space, these models achieve state of the art performance on a wide range of downstream tasks, including zero shot classification, image text retrieval, and visual question answering [3, 4].

Despite their impressive capabilities, the security and robustness of VLP models are under increasing scrutiny. Adversarial attacks, which involve adding small, imperceptible perturbations to input data to mislead a model, have been extensively studied in unimodal contexts [5, 6]. However, the multimodal nature of VLP models presents a unique and complex attack surface [7]. An adversary can target the image input, the text input, or both simultaneously.

Current adversarial strategies against VLP models can be broadly categorized as:

1. Unimodal Attacks: Applying established methods like PGD [6] solely to the image branch, ignoring the text modality [8].
2. Early Fusion Attacks: Concatenating image and text features and then applying a single attack [9].
3. Ensemble style Attacks: Averaging gradients from both modalities or performing separate attacks and combining them [10].

A key limitation of these approaches is their failure to dynamically model the bidirectional, synergistic relationship between the two modalities during the attack generation process. An effective perturbation in one modality is

highly dependent on the state of the other. For instance, a perturbation that slightly alters the perception of a "dog" in an image becomes far more potent if the text is simultaneously perturbed from "a photo of a cat" to "a photo of a wolf."

In this paper, we propose that attacking a VLP model is not two separate problems but a single, coupled optimization problem. We introduce the Feedback Based Modal Mutual Search (FB MMS) framework. The core idea is to perform a mutual search in the adversarial space of both modalities, where the attack on each modality is iteratively refined based on the feedback (gradients) from the current state of the attack on the other modality. This creates a cooperative "dialogue" between the image and text attackers, allowing them to collaboratively discover adversarial pairs that are much more effective than those found by independent or statically fused strategies.

Our contributions are threefold:

1. We formulate the problem of adversarial attack on VLP models as a coupled optimization problem and propose the FB MMS framework to solve it.
2. We design an iterative algorithm where image and text perturbations are updated mutually, using gradient feedback from the other modality to guide the search.
3. We conduct extensive experiments on image text retrieval and VQA tasks, demonstrating that FB MMS consistently and significantly outperforms existing attack methods, revealing a greater vulnerability in VLP models.

2. Methodology

Let a VLP model be denoted by $F(I, T)$, which takes an image I and a text token sequence T as input and produces a similarity score or an answer. The goal of a targeted adversarial attack is to find a perturbed image $I' = I + \delta_I$ and a perturbed text $T' = T + \delta_T$ (where $+\delta_T$ operates in the token embedding space) such that $F(I', T')$ produces a specific, incorrect target output (e.g., high similarity with a mismatched caption), while the perturbations δ_I and δ_T are small: $\|\delta_I\|_p \leq \epsilon_I$ and $\|\delta_T\|_p \leq \epsilon_T$.

2.1 Preliminaries: Adversarial Attacks on VLP

A straightforward baseline is the Projected Gradient Descent (PGD) attack applied to both modalities independently or in a simple fused manner. For example, a fused PGD might update perturbations as:

$$\delta_I^{t+1} = \Pi_{\epsilon_I} (\delta_I^t + \alpha \cdot \text{sign}(\nabla_{\delta_I} \mathcal{L}(I+\delta_I^t, T+\delta_T^t)))$$

$$\delta_T^{t+1} = \Pi_{\epsilon_T} (\delta_T^t + \alpha \cdot \text{sign}(\nabla_{\delta_T} \mathcal{L}(I+\delta_I^t, T+\delta_T^t)))$$

where $\mathcal{L}$ is the adversarial loss (e.g., targeted cross entropy), and Π denotes projection onto the ϵ ball. This approach, while better than unimodal attacks, treats the gradients as separate signals that are computed at the same point but do not actively inform each other's search direction.

2.2 Feedback Based Modal Mutual Search (FB MMS)

The FB MMS framework is designed to create a feedback loop between the two modalities. The intuition is that the optimal perturbation for the image depends on what the text is being perturbed towards, and this relationship is dynamic. We implement this via a two step, iterative process.

ISSN:3107-930X(ONLINE)

Let θ_I^t and θ_T^t be the "search states" for the image and text attackers at iteration t. These states contain the current perturbation estimates and any necessary momentum terms.

Step 1: Text Guided Image Attack Update.

First, we update the image perturbation using the current state of the text attack . We compute the gradient of the loss with respect to the image, but we condition it on the latest text perturbation δ_T^t.

$$g_I^t = \nabla_{\delta_I} \mathcal{L}(I + \delta_I^t, T + \delta_T^t)$$

However, the key to FB MMS is to then use this gradient to update a momentum term [11] for the image attack, which accumulates the search direction. The update for the image perturbation is then:

$$m_I^{t+1} = \mu \cdot m_I^t + \frac{g_I^t}{\|g_I^t\|_1}$$

$$\delta_I^{t+1} = \Pi_{\epsilon_I} (\delta_I^t + \alpha_I \cdot \text{sign}(m_I^{t+1}))$$

This step ensures the image attack is moving in a direction that is effective given the current text adversary .

Step 2: Image Guided Text Attack Update.

Next, we update the text perturbation using the newly updated image $I + \delta_I^{t+1}$. This creates the feedback loop. We compute the gradient with respect to the text, conditioned on the new image.

$$g_T^t = \nabla_{\delta_T} \mathcal{L}(I + \delta_I^{t+1}, T + \delta_T^t)$$

We then update the text perturbation's momentum and state:

$$m_T^{t+1} = \mu \cdot m_T^t + \frac{g_T^t}{\|g_T^t\|_1}$$

$$\delta_T^{t+1} = \Pi_{\epsilon_T} (\delta_T^t + \alpha_T \cdot \text{sign}(m_T^{t+1}))$$

Crucially, the text attack now reacts to the changes made in the image attack in the same iteration. This allows the text attacker to adjust its strategy based on the new "position" of the image in the joint embedding space.

This two step process is repeated for a fixed number of iterations. The mutual guidance ensures that the two attackers are not working at cross purposes but are cooperatively navigating the loss landscape towards a highly effective adversarial configuration. The algorithm is summarized below.

Algorithm 1: Feedback Based Modal Mutual Search (FB MMS)

Input: Clean image I, clean text T, target y_{target}, VLP model F, loss function $\mathcal{L}$, steps N, step sizes α_I, α_T, bounds ϵ_I, ϵ_T.

1. Initialize $\delta_I^0, \delta_T^0, m_I^0, m_T^0$ to zeros.
2. for $t = 0$ to N 1 do
3. // Text Guided Image Update
4. Compute $g_I^t = \nabla_{\delta_I} \mathcal{L}(F(I+\delta_I^t, T+\delta_T^t), y_{target})$

ISSN:3107-930X(ONLINE)

5. Update momentum: \(m_I^{t+1} = \mu \cdot m_I^t + g_I^t / \|g_I^t\|_1 \)
6. Update perturbation: \(\delta_I^{t+1} = \Pi_{\epsilon_I} (\delta_I^t + \alpha_I \cdot \text{sign}(m_I^{t+1})) \) 7. // Image Guided Text Update
8. Compute \(g_T^t = \nabla_{\delta_T} \mathcal{L}(F(I+\delta_I^{t+1}, T+\delta_T^t), y_{target}) \)
9. Update momentum: \(m_T^{t+1} = \mu \cdot m_T^t + g_T^t / \|g_T^t\|_1 \)
10. Update perturbation: \(\delta_T^{t+1} = \Pi_{\epsilon_T} (\delta_T^t + \alpha_T \cdot \text{sign}(m_T^{t+1})) \)
11. end for

Output: Adversarial pair \((I + \delta_I^N, T + \delta_T^N) \).

3. Experiments and Results

3.1 Experimental Setup

Model: We use the publicly available CLIP (ViT B/32) model [1] as our primary target VLP model.

Datasets: We evaluate on the Flickr30k dataset for image text retrieval and a subset of the VQAv2 dataset [12] for visual question answering.

Tasks:

Retrieval Attack: For a given ground truth image text pair, the goal is to reduce their cosine similarity in the CLIP embedding space while maximizing the similarity of the adversarial image with a target negative caption and the adversarial text with a target negative image .

VQA Attack: The goal is to change the model's predicted answer for a given (image, question) pair to a specific target answer.

Baselines: We compare FB MMS against:

1. Image Only PGD [6]: Attacking only the image modality.
2. Text Only PGD [13]: Attacking only the text modality (in the word embedding space).
3. Fused PGD: The standard bimodal PGD described in Section 2.1.
4. Modal Ensemble Attack (MEA) [10]: Averaging the gradients from both modalities before taking a step.

Metrics: Attack Success Rate (ASR), Mean Perturbation Norm (L2 for image, L2 in embedding space for text), and the final adversarial similarity score.

3.2 Results and Analysis

Table 1: Image Text Retrieval Attack Results on Flickr30k (Targeted)

Attack Method	Image→Text ASR (%)	Text→Image ASR (%)	Avg. Image \(\ell_2 \)	Avg. Text \(\ell_2 \)
Image Only PGD	68.4		4.12	
Text Only PGD		55.1		1.85
Fused PGD	82.7	73.9	4.15	1.88
MEA	85.2	76.5	4.18	1.87
FB MMS (Ours)	94.8	88.3	4.21	1.89

The results in Table 1 clearly demonstrate the superiority of FB MMS. It achieves a significantly higher Attack Success Rate (ASR) in both retrieval directions compared to all baselines, with only a marginal increase in perturbation

size. This indicates that the mutual search strategy is far more efficient at finding successful adversarial examples within the constraint budget.

Table 2: VQA Attack Success Rate (%) on VQAv2

Attack Method	ASR (%)
Image Only PGD	45.6
Text Only PGD	52.3
Fused PGD	65.8
MEA	68.1
FB MMS (Ours)	79.5

On the more complex VQA task, FB MMS again shows a substantial performance gain (Table 2). Changing a model's answer requires a more precise corruption of the cross modal reasoning process, which FB MMS is uniquely equipped to achieve through its cooperative perturbation strategy.

Ablation Study: We ablated the core component of FB MMS—the feedback loop. A variant, "FB MMS (Parallel)," which updates both modalities using the gradients from \((I+\delta_I^t, T+\delta_T^t) \) (removing the sequential dependency), performed worse than the full FB MMS but better than Fused PGD. This confirms that the iterative, guided update (first image, then text with new image feedback) is critical to the method's success.

4. Conclusion and Future Work

In this work, we presented FB MMS, a novel adversarial attack framework for Vision Language Pre Training models. By reformulating the attack as a cooperative, feedback driven search between modalities, FB MMS uncovers vulnerabilities that are missed by traditional attacks. Our experiments validate that this mutual search strategy is significantly more effective and efficient.

This research underscores the need for a deeper investigation into the robustness of multimodal systems. The very mechanism that makes them powerful—tight cross modal alignment—can be exploited as a vulnerability. Future work will explore defensive strategies, such as adversarial training that specifically accounts for such coupled attacks [14]. Furthermore, we plan to extend the FB MMS principle to other multimodal tasks and model architectures, and to investigate black box attack scenarios using feedback from the model's output scores rather than gradients.

References

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning (ICML) .

[2] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems (NeurIPS)
.

[3] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). CoCa: Contrastive Captioners are Image Text Foundation Models. arXiv preprint arXiv:2205.01917 .

[4] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual Question Answering. International Conference on Computer Vision (ICCV) .

[5] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR) .

[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR) .

[7] Xu, X., Chen, X., Liu, C., Rohrbach, A., Darell, T., & Song, D. (2018). Fooling vision and language models despite localization and attention mechanism. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
.

[8] Zhao, Z., Liu, Z., & Larson, M. (2021). On Success and Simplicity: A Second Look at Transferable Targeted Attacks. Advances in Neural Information Processing Systems (NeurIPS) .

[9] Liu, Y., Zhang, W., & Wang, J. (2022). Multimodal Adversarial Attack and Defense in the Pixel and Text Embedding Space. IEEE Transactions on Image Processing .

[10] Yang, J., Jiang, Y., Huang, Z., Yang, B., & Zhao, Y. (2022). M^3A: Model Modal Adversarial Attack for Multimodal Learning. ACM International Conference on Multimedia .

[11] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) .

[12] Goyal, Y., Khot, T., Summers Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) .

[13] Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White Box Adversarial Examples for Text Classification. Association for Computational Linguistics (ACL) .

[14] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. ICLR .

[15] Chen, H., Zhang, H., Chen, P. Y., Yi, J., & Hsieh, C. J. (2017). Attacking visual language grounding with adversarial examples: A case study on neural image captioning. Association for Computational Linguistics (ACL)
.

[16] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. International Conference on Learning Representations (ICLR) .

[17] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy (S&P) .

[18] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. International Conference on Learning Representations (ICLR) .

[19] Papernot, N., McDaniel, P., Swami, A., & Harang, R. (2016). Crafting adversarial input sequences for recurrent neural networks. IEEE Military Communications Conference (MILCOM) .

[20] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics (NAACL) .