Coordinated Robustness Evalaution Framework for Vision-Language Models

Ashwin Ramesh Babu, Sajad Mousavi, Desik Rengarajan, Vineet Gundecha, Sahand Ghorbanpour, Avisek Naug, Antonio Guillen, Ricardo Luna Gutierrez, Soumyendu Sarkar*

Hewlett Packard Enterprise (Hewlett Packard Labs)

{ashwin.ramesh-babu, sajad.mousavi, desik.rengarajan, vineet.gundecha, sahand.ghorbanpour, avisek.naug, antonio.guillen, rluna, soumyendu.sarkar}@hpe.com

Abstract

Vision-language models, which integrate computer vision and natural language processing capabilities, have demonstrated significant advancements in tasks such as image captioning and visual question and answering. However, similar to traditional models, they are susceptible to small perturbations, posing a challenge to their robustness, particularly in deployment scenarios. Evaluating the robustness of these models requires perturbations in both the vision and language modalities to learn their inter-modal dependencies. In this work, we train a generic surrogate model that can take both image and text as input and generate joint representation which is further used to generate adversarial perturbations for both the text and image modalities. This coordinated attack strategy is evaluated on the visual question and answering and visual reasoning datasets using various state-of-the-art vision-language models. Our results indicate that the proposed strategy outperforms other multi-modal attacks and single-modality attacks from the recent literature. Our results demonstrate their effectiveness in compromising the robustness of several state-of-the-art pre-trained multi-modal models such as instruct-BLIP, ViLT and others.

1 Introduction

Evaluating the robustness of computer vision models and architectures has existed for a long time now. The success of vision-language models in bridging the gap between visual and textual representations has enabled a wide range of applications, from image captioning, visual question and answering, multi-modal information retrieval and generation tasks Dai et al. (2024); Liu et al. (2024); Kim et al. (2021). As these models are becoming more ubiquitous in real-world deployments, their vulnerability to adversarial attacks poses a significant liability concern. These attacks can deceive models into misinterpreting or misclassifying visual and textual inputs, leading to erroneous outputs and potentially harmful consequences. Understanding and mitigating these threats is crucial to ensuring the reliability and security of vision-language models as they become more integrated into critical applications.

Adversarial examples, carefully crafted by adding imperceptible perturbations to input data, can cause these models to make incorrect and potentially dangerous predictions. While adversarial

*Corresponding author.

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

attacks have been extensively studied in computer vision and natural language processing domains, the unique challenges posed by the multimodal nature of the vision-language model necessitate more detailed investigation. These models must contend with adversarial perturbations that can manifest in both the visual and textual components simultaneously, potentially exploiting intricate cross-modal interactions and challenging the model's ability to reason coherently across modalities. Furthermore, most of the studies in the computer vision domain and natural language processing (NLP) are designed only for classification tasks, Vision and Language models mainly involve different types of downstream tasks such as visual question and answering, and cross-modality retrieval.

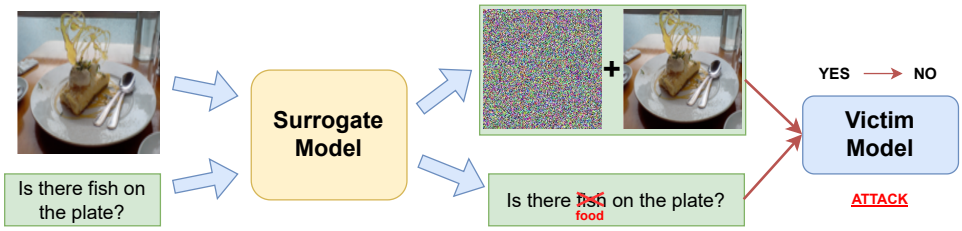

Figure 1: Sample outcome with the proposed method on VQA dataset model with ViLT as victim model

In this work, we propose a novel coordinated attack strategy that introduces perturbations to both the vision and language modules, aiming to cause the model to alter its outcomes. Our approach involves several key steps:
First, we train a multi-modal encoder (Multi-modal surrogate in figure 2 left) to align the embeddings generated from a combination of an image and the corresponding question with those generated by an answer encoder (text encoder in figure 2 left). Specifically, the multi-modal encoder processes an image and a corresponding question, generating a joint representation. This joint representation is encouraged to closely match an answer representation produced by a transformer-based text encoder. The trained multi-modal encoder then serves as a surrogate for our attack method. During the attack phase, both the surrogate and the text encoder generate feature representations. The goal is to craft adversarial perturbations that push these representations apart in the embedding space, achieved through a gradient-based update for both image and text modalities. This process is formulated as an optimization problem. Figure 1 shows a sample output from our proposed method. Our approach differs from existing methods in several significant ways:

While most literature on multi-modal attacks deals with classification problems where output logits are readily accessible, our method targets complex downstream tasks such as visual question and answering (VQA) rather than simple classifications. Additionally, majority of surrogate-based approaches use feedback from victim models to generate adversarial perturbations. our method employs a generic surrogate model that generates adversarial samples that are not victim model dependent and can effectively mislead several state-of-the-art victim models. Finally, most multi-modal adversarial attack approaches in the literature use word-swapping techniques Li et al. (2020) for text attacks, which result in uncoordinated changes between text and image modalities. In contrast, our method ensures that perturbations in both modalities are coordinated, enhancing the effectiveness of the attack. The contribution of our approach can be summarized as;

1. We propose a coordinated attack strategy that has been designed for both vision and language components to highlight the unique vulnerabilities in the multi-modal context.
2. The proposed method acts as a surrogate model that can craft generic adversarial samples that can be used against several victim models without model specific feedback.
3. Results on visual question and answering task and visual reasoning task demonstrate the superiority of the proposed method when compared to our competitors.

2 Related Works

Adversarial attacks were first introduced in computer vision, which demonstrates the vulnerability of neural networks. Unlike black-box attacks, where the attacker has no access to the model's

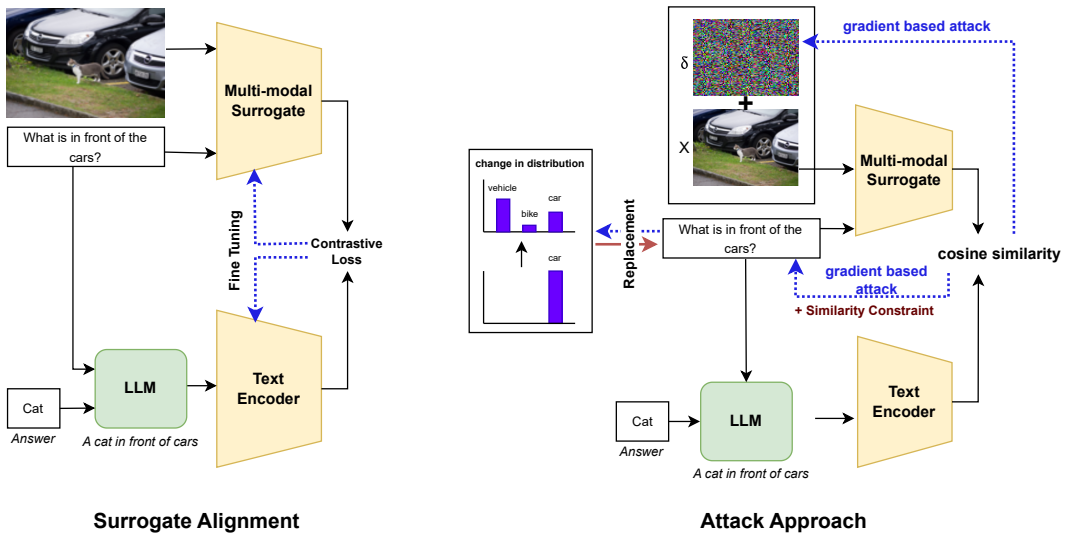

Figure 2: Overview of the workflow. The left figure represents the alignment step to align a custom architecture comprising of an image and question module with the answer module. An additional LLM component is used in the alignment step to convert a single-worded ground-truth to a sentence based on the question. The dotted line represents the flow of gradients. The figure in the right represents the attack approach. The embeddings generated by the multi-modal surrogate and the text encoder are compared using cosine similarity and are used to generate perturbations for both the text and the image modality. The figure represents one iteration of the attack and the training process. The surrogate once aligned is used across all victim models which proves the generalizability of the proposed method.

internal parameters, white-box attacks assume full knowledge of the model, allowing for more precise and effective perturbations. One of the pioneering works in this area is by Szegedy et al. (2013), who demonstrated the vulnerability of neural networks to adversarial examples in the context of image classification. This concept was later extended to text models by Papernot et al. (2016), who introduced a technique for generating adversarial sequences by leveraging the gradients of the model.

2.1 Attacks for text models

Subsequent works have built upon these foundations, exploring various methods for crafting adversarial text to mislead text based models Xu et al. (2020); Wang et al. (2021, 2019); Zhang et al. (2020,?). Adversarial attack on text are broadly divided into two approaches, where one deals with sentence level and the other at word level. The primary difference between the sentence-level attacks and the word-level attacks is in the granularity and the nature of the modification made to the input text. The sentence level attack involve modifying entire sentences or adding new sentences to the text Wang et al. (2020); Lin et al. (2021); Huang and Chang (2021); Han et al. (2020)

2.1.1 Word-level attacks

Word level adversarial attack focus on modifying individual word within a sentence. This attack deals with smaller and granular changes. Some of the most impactful works in this area include Bert-attack Li et al. (2020). Jin et al. in the work Jin et al. (2020) propose a textfooler which uses a powerful pretrained BERT to generate adversarial adversarial attack. Some of the popular attacks in the word level He et al. (2021) Yang et al. in their work Yang et al. (2022) proposes adversarial attack that is capable of adversarially transforming inputs to make victim models produce wrong outputs. Similarly, there have been deveral other derivatives of the listed attacks such as Garg and Ramakrishnan (2020); Sun et al. (2020); Xu et al. (2021); Li et al. (2020); Ye et al. (2022); Chang et al. (2023).

2.2 Attacks for vision models

Adversarial attacks on vision models have been studied for over a decade. Some of the significant reviews on adversarial attack for vision models are Akhtar et al. (2021); Xu et al. (2020); Khamaiseh et al. (2022); Chakraborty et al. (2021). Among adversarial attacks, square attack has created significant impact in the adversarial attack literature Andriushchenko et al. (2020). Similarly, Moosavi-Dezfooli et al. (2016) proposes an efficient approach to fool deep networks using gradient information. The authors use a natural saddle point formulation to capture the notion of security against adversarial attacks in a principled manner

2.3 Multi-modal Adversarial attacks

With recent developments in the multi-modal foundational models, they have similar vulnerabilities as the text models or computer vision models. Zhou et al. Zhou et al. (2024) in their work introduce a multimodal attack to align clean and adversarial text embeddings with clean and adversarial visual features. They evaluate their method on image classification tasks to prove their superiority. Zhang et al. Zhang et al. (2022) in their work co-attack add perturbations on multi-modality settings. The work uses CLIP-based architecture which are aligned to generate adversarial perturbations for the image and text modality. The authors evaluate their work on image text retrieval and visual entailment tasks to demonstrate the effectiveness of their work. Zhao et al. Zhao et al. (2024) in their work craft adversarial examples to fool VLMs for image captioning tasks by alternating between text-to-image and image-to-text models with having a specific target text to drive the perturbation towards a specific direction. VLAttack proposes an attack strategy which involves querying a black-box model exhaustively to learn the mutual connections between the perturbed image and text to cause misclassification in their work Yin et al. (2023).

3 Problem Formulation

In this work, we aim to generate adversarial perturbations for an image and a text question to make their combined representation and the corresponding answer representation move apart in the embedding space. These adversarial samples to cause different responses from a variety of victim models compared to the original inputs. Let x_q, x_i, x_a be a sample from dataset D.

$$(x_q, x_i, x_a) \sim \mathcal{D}$$

$E_{iq}(x_i, x_q) = \mathbf{r}_{iq}$ Encodes an image i_m and a question i_q into a joint representation r_{iq}. $E_a(x_a) = \mathbf{r}_a$ encodes an answer x_a into a representation r_a. Generate perturbations δx_i for the image and δx_q for the question such that the similarity between $\mathbf{r}_{iq}$ and $\mathbf{r}_a$ decreases, pushing $\mathbf{r}_{iq}$ and $\mathbf{r}_a$ farther apart in the embedding space. The perturbed image and question should have different responses from a victim model compared to the original inputs. We formulate the problem as an optimization task where we minimize the similarity between the perturbed joint representation and the answer representation while also causing different outputs from a victim model V.

Let $V(x_i, x_q)$ be the response of the victim model to the original image and question and $V(x_i + \Delta x_i, x_q + \Delta x_q)$ be the response of the victim model to the perturbed image and question.We want $V(x_i + \Delta x_i, x_q + \Delta x_q) \neq V(x_i, x_q)$.

So, the overall problem can be formulated as,

$$\min_{\Delta x_i, \Delta x_q} \text{similarity}(E_{iq}(x_i + \Delta x_i, x_q + \Delta x_q), E_a(x_a))$$

$$\text{subject to} \quad V(x_i + \Delta x_i, x_q + \Delta x_q) \neq V(x_i, x_q)$$

4 Proposed Method

In the proposed method, we first explain the architecture of the surrogate model used and how the surrogate model was aligned for the purpose of the adversarial example generation. Next, we explain the attack strategies for the individual modalities, and how they are combined to generate adversarial perturbation. The overall flow of the proposed method is represented in the figure 2.

4.1 Surrogate Model to generate perturbations

4.1.1 Surrogate architecture

Current approaches in vision-language models heavily rely on image feature extraction processes that involve regional supervision such as object detection which is computationally more expensive and has limitations in applications. Hence at the core of the surrogate model lies a transformer-based architecture to capture the interaction between the text and the image to generate a joint representation. The surrogate architecture handles two modalities in a single unified manner consisting of stacked blocks that include a multithreaded self-attention layer and MLP layer.

The input text $x_q \in \mathbb{R}^{L\times|V|}$ is embedded to $\tilde{x}_q \in \mathbb{R}^{L\times H}$ with a word embedding matrix $T \in \mathbb{R}^{|V|\times H}$ and a position embedding matrix $T^{\text{pos}} \in \mathbb{R}^{(L+1)\times H}$. Here, L represents the input sequence length, V represents the vocabulary size, and H represents the embedding dimension size.

The input image $x_i \in \mathbb{R}^{C\times H_t\times W}$ with C being the channel, Ht and W being the height and width is sliced into patches and flattened to $v \in \mathbb{R}^{N\times(P^2\cdot C)}$ where (P, P) is the patch resolution and $N = Ht \times W/P^2$ with N signifying total number of patches. Followed by linear projection $V \in \mathbb{R}^{(P^2\cdot C)\times H}$ and position embedding $V_{\text{pos}} \in \mathbb{R}^{(N+1)\times H}$, v is embedded into $\tilde{v} \in \mathbb{R}^{N\times H}$.

The text and image embeddings are summed with their corresponding modal-type embedding vectors $t^{\text{vtx}}, v^{\text{vtx}} \in \mathbb{R}^H$, then are concatenated into a combined sequence z^0. The contextualized vector is iteratively updated through D-depth transformer layers up until the final contextualized sequence. A pooled representation of the whole multimodal input, and is obtained by applying linear projection. Figure 3 represents the surrogate architecture. In the positional embedding, the first element represents the modal-type embedding. For text, the second position represents the token position embedding, and for image, the second position represents the patch position. $*$ represents the extra learnable embedding.

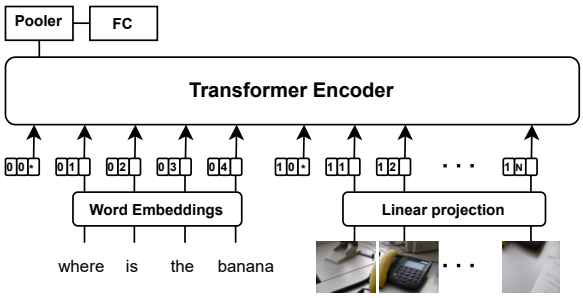

Figure 3: Surrogate Architecure inspired from Dosovitskiy et al. (2020)

4.1.2 Surrogate Alignment

The goal is to align the surrogate architecture that takes in an image and text and generates a joint representation. To align the representations r_{iq} and r_a, we adopt a contrastive loss to encourage the representations to be closer. Specifically, we aim to minimize the cosine similarity between the positive pair (r_{iq}, r_a) while maximizing the cosine similarity between the negative pairs (r_{iq}, r'_a) and (r'_{iq}, r_a), where r'_a and r'_{iq} are negative samples from the batch. The contrastive loss is defined as:

$$\mathcal{L} = -\log\left(\frac{ps(r_{iq}, r_a)}{ps(r_{iq}, r_a) + \sum_i ps(r_{iq}, r'_a i) + \sum_j ps(r'_{iq} j, r_a)}\right) \tag{1}$$

where ps represents $\exp(\cos(x, y))$

Where $\cos(x, y)$ is the cosine similarity between vectors x and y. By minimizing this loss, we encourage the multimodal and text representations of the correct (x_i, x_q, x_a) (image, question, answer) triplets to be pulled closer together in the embedding space. We fine-tuned the surrogate architecture on two training sets, MSCOCO and Flicker30k dataset.

4.2 Coordinated Attack

In this section, we will discuss the attack strategy for the individual modules. Algorithm 1 represents the overall flow of generating the adversarial samples. Once the adversarial samples are generated, they are evaluated on victim models to see the change in the output. We use the same gradients from a defined objective to generate adversarial perturbations for both the image module and the text module making attack more unified.

4.2.1 Image perturbations

Our system employs a Surrogate encoder that has been aligned to generate adversarial perturbations, to reduce the similarity between the embeddings from the surrogate model and the answer model. This section elucidates the methodology to craft adversarial samples for the image input, thereby fooling a victim model V.

Given a question x_q, an image x_i, and an answer x_a, the encoders generate embeddings $E_{iq}(x_q, x_i)$ and $E_a(x_a)$, respectively. The similarity between these embeddings is measured using cosine similarity. The objective of the adversarial attack is to reduce this similarity between the embeddings by adding adversarial perturbations thereby deceiving the model. Equation 2 represents the cosine similarity between the two embeddings.

$$\mathcal{L}(x_q, x_i, x_a) = \frac{E_{iq}(x_q, x_i) \cdot E_a(x_a)}{\|E_{iq}(x_q, x_i)\| \|E_a(x_a)\|} \tag{2}$$

The attack strategy involves iteratively perturbing the image to minimize the cosine similarity between $E_{iq}(x_q, x_i)$ and $E_a(x_a)$ to generate x_i'.

where $x_i' = x_i + \delta$ represents the perturbed image and δ is the adversarial perturbation applied to the original image x_i. We apply an iterative method used to optimize the adversarial perturbation δ such that it minimizes the objective function $\mathcal{L}(x_i', q, a)$. The process starts with an initial perturbation $\delta_0 = random\_bounded\_initialization$.

For each iteration $t = 0, 1, \ldots, T-1$, the perturbed image along with the question is used to compute the loss using the cosine similarity. The gradient of the loss function $\mathcal{L}$ with respect to the perturbed input of the previous step is calculated using backpropagation. The sign of this gradient is used to iteratively update the perturbation, ensuring that each step moves in the direction that maximally reduces the cosine similarity between the embeddings. The perturbation after each step is represented as,

$$\delta_{t+1} = \Pi_\epsilon \left(\delta_t + \alpha \cdot \text{sign} \left(\nabla_\delta \left(- \frac{E_{iq}(q, x_i + \delta_t) \cdot E_A(a)}{\|E_{iq}(q, x_i + \delta_t)\| \|E_A(a)\|} \right) \right) \right) \tag{3}$$

The perturbed image that has caused the lowest cosine similarity after T iterations is $x_i' = x_i + \delta_T$. α represents the step size,

$\nabla_\delta \mathcal{L}$ represents the gradient of the loss function with respect to the perturbation δ. - Π_ϵ is the projection operator ensuring the perturbation remains within the ϵ-ball around the original image x_i:

$$\Pi_\epsilon(\delta) = \text{clip}(\delta, -\epsilon, \epsilon)$$

The projection operator Π_ϵ ensures that the perturbation δ satisfies the constraint $\|\delta\|_\infty \leq \epsilon$, keeping the perturbation imperceptible. For our experiments, we maintained the $\epsilon = 8/255$.

4.2.2 Text Attack

To generate adversarial examples that minimize the cosine similarity between question and answer embeddings and instead of searching for a single adversarial example, we aim to find a distribution of adversarial questions $P_\Theta(x_q)$ parameterized by Θ, such that when sampling $\tilde{x}q \sim P_\Theta(x_q)$, the cosine similarity between the embeddings $E_{iq}(\tilde{x}_q, x_i)$ and $E(x_a)$ of the adversarial question and original answer is minimized.

To instantiate the adversarial distribution $P_\Theta(x_q)$, we leverage the Gumbel-Softmax technique Jang et al. (2016) which provides a simple way to sample from a categorical distribution while

Table 1: Comparison of the proposed method with State-of-the-art competitors for visual question and answering task on VQA dataset. All results are displayed by Average Success Rate (%).

Pre-trained Model	Image Only			Text only		Multi-modality		
	SSP	FDA	BSA	BA	RR	CO-Attack	VLAttack	Ours
ViLT	50.36	29.27	65.20	17.24	8.69	35.13	78.05	**94.3**
BLIP	11.84	7.12	25.04	21.04	2.94	14.24	48.78	**91.0**
GIT	-	-	-	-	-	51.16	78.82	**80.43**

Table 2: Comparison of the proposed method with competitors for visual reasoning dataset. All results are displayed by Average success rate

Pre-trained Model	Image Only			Text only		Multi-modality		
	SSP	FDA	BSA	BA	RR	CO-Attack	VLAttack	Ours
ViLT	21.58	35.13	52.17	32.18	24.82	40.04	66.65	**73.32**
BLIP	6.88	10.22	27.16	33.8	16.92	8.70	52.66	**58.45**
GIT	-	-	-	-	-	18.66	41.78	**54.54**

maintaining differentiability. Let $\tilde{\pi}_1, ..., \tilde{\pi}n$ be samples from $\tilde{P}\Theta$, the Gumbel-Softmax distribution with temperature τ parameterized by $\Theta \in \mathbb{R}^{n \times V}$, which draws samples π by independently sampling where V is the vocabulary size and n is the sequence length. Each $\tilde{\pi}_i \in \mathbb{R}^V$ is a vector representing a categorical distribution over vocabulary tokens at position i. We define the adversarial question $\tilde{x}_q = e(\tilde{\pi}_1) \oplus ... \oplus e(\tilde{\pi}_n)$, the sequence formed by looking up and concatenating the embeddings $e(\cdot)$ of the sampled token distributions. The objective is to minimize the cosine similarity between the joint question and image embedding $E_{iq}(\tilde{x}_q, x_i)$ and the given answer embedding $E(x_a)$. The cosine similarity which is the objective is same as equation 2 with both the image and the question perturbed as $\tilde{x_i}$ and $\tilde{x_q}$.

Additionally, to ensure the generated adversarial questions remain fluent and semantically preserving, we incorporate two additional constraints. The first promotes fluency by minimizing the negative log-likelihood $\text{NLL}_g(\tilde{x}_q)$ of the adversarial text under an external language model g. The second controls semantic divergence by minimizing $\rho_g(x_q, \tilde{x}_q)$ based on the BERTScore Zhang et al. (2019) which measures the semantic similarity between the original question x_q and adversarial $\tilde{x}_q$ using contextualized embeddings from g. The full objective is a weighted combination:

$$\mathcal{J}(\Theta) = \mathcal{L}(\Theta) + \lambda_{lm}\text{NLL}_g(\tilde{x_q}) + \lambda sim\rho_g(x_q, \tilde{x}_q)$$

Where $\lambda_{lm}, \lambda_{sim} > 0$ control the strengths of the language model and semantic similarity constraints respectively. We optimize Θ using gradient descent on $\mathcal{J}(\Theta)$ to find the parameters of the adversarial question distribution $P_\Theta(x_q)$. At inference time, we can efficiently sample $\tilde{x}q \sim P\Theta(x_q)$ and input it to the QA system. The sampled $\tilde{x}_q$ will have minimized cosine similarity to the given answer embedding $g(x_a)$, while being fluent and preserving semantic similarity to the original question x_q as guided by the constraints. This distributional adversarial attack framework provides a powerful and general approach compared to previous heuristic word replacement methods. By leveraging gradient-based optimization on a continuous distribution over the input space, along with differentiable constraints, it can navigate the landscape more effectively to find stronger and more natural adversarial examples that are not hand crafted or hard set.

5 Experiments

5.1 Exerimental Setup

Experiments are conducted on two different datasets VQA dataset and visual reasoning dataset. We sample 1000 samples randomly from the validation set of the above mentioned dataset. Each selected sample is correctly classified by all the target models to be considered for the evaluation. The generated adversarial samples were evaluated on 3 different models, ViLT, GIT, BLIP Kim et al. (2021); Liu et al. (2023); Dai et al. (2024). The experiments were conducted on 3 pre-trained VL models and the use of Attack Success Rate (ASR) to evaluate the performance. We evaluate on two different datasets, VQA dataset Goyal et al. (2017) and the visual reasoning dataset Suhr et al. (2018).

Algorithm 1 High-Level flow of adversarial perturbation generation for image and text modality

Input: Clean image x_i, question x_q, answer x_a, victim model V, encoders E_{iq} and E_A, perturbation bounds ϵ_i and , step size α, number of iterations T, language model g, BERTScore function ρ_g, weight parameters λ_{lm} and λ_{sim}
Initialize perturbations $\delta_i = \mathbf{0}$, $\Theta \in \mathbb{R}^{n \times V}$ with small random values
for $t = 0$ to $T - 1$ **do**
  Sample adversarial question $\tilde{x}_q \sim P_\Theta(x_q)$ using Gumbel-Softmax
  Compute joint embedding $\mathbf{r}_{iq} = E_{iq}(x_i + \delta_i, \tilde{x}_q)$
  Compute answer embedding $\mathbf{r}_a = E_A(x_a)$
  Compute cosine similarity loss:

$$\mathcal{L}_{iq} = \frac{\mathbf{r}_{iq} \cdot \mathbf{r}_a}{\|\mathbf{r}_{iq}\| \|\mathbf{r}_a\|}$$

  if $\mathcal{L}_i <$ best_similarity **then**
    Update best perturbation for image: $\delta_i^* = \delta_i$
    Update best perturbation for text: $\Theta^* = \Theta$
    Update best similarity: best_similarity $= \mathcal{L}_i$
  end if
  Compute full objective for text module:

$$\mathcal{J}(\Theta) = \mathcal{L}_{iq} + \lambda_{lm} \cdot \text{NLL}_g(\tilde{x}_q) + \lambda_{sim} \cdot \rho_g(x_q, \tilde{x}_q)$$

  Update Θ using gradient descent:

$$\Theta \leftarrow \Theta - \alpha \cdot \nabla_\Theta \mathcal{J}(\Theta)$$

  Compute gradient of the loss w.r.t. δ_i:

$$\nabla_{\delta_i} \mathcal{L}_{iq} = \nabla_{\delta_i} \left(-\frac{\mathbf{r}_{iq} \cdot \mathbf{r}_a}{\|\mathbf{r}_{iq}\| \|\mathbf{r}_a\|} \right)$$

  Update perturbation δ_i:

$$\delta_i = \Pi_{\epsilon_i} \left(\delta_i + \alpha \cdot \text{sign}(\nabla_{\delta_i} \mathcal{L}_{iq}) \right)$$

end for
Output: Adversarial image $x_i' = x_i + \delta_i$ and adversarial question $\tilde{x}_q \sim P_\Theta(x_q)$

For our baselines, we compare our performance with several uni-modal approaches and multi-modal approaches from the recent literature SSP Naseer et al. (2020), FDA Ganeshan et al. (2019), BA Li et al. (2020), RR Xu et al. (2021), Co-Attack Zhang et al. (2022), VLA-attack Yin et al. (2023).

Experimental details: For all experiments, our surrogate architecture is composed of weights from ViT-B/32 pre-trained on ImageNet, hence the name ViLT-B/32. Hidden size H is 768, layer depth D is 12, patch size P is 32, MLP size is 3072 and the number of attention heads is 12. For the answer encoder E_a, we use a bert based architecture with 12 transformer blocks, and hidden size as 768, 12 self-attention heads, and a feed-forward network size of 3072. During the attack we use adam optimizer to compute gradients with respect to our corresponding inputs with a learning rate set to 0.0005. Our training experiments were conducted on an Ubuntu machine with 8 Tesla V100S-PCIE-32GB GPUs and an Intel Xeon Gold 6246R CPU @ 3.40GHz with 16 cores. Training was distributed across multiple GPUs for surrogate alignment.

5.2 Details on victim models

ViLT has proven to have performed well in several downstream tasks. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$ and a sentence T, ViLT yields M image tokens using a linear transformation on the flattened image patches, where each token is a 1D vector and $M = \frac{HW}{P^2}$ for a given patch resolution (P, P). By attending visual and text tokens and a special token $\langle cls \rangle$ in a Transformer encoder with twelve layers, the output feature from the $\langle cls \rangle$ token is fed into a task-specific classification head for the final output.

Instruct-BLIP is a vision-language instruction tuning framework that enables general-purpose model to solve a wide variety of visual language tasks. Instruct-BLIP performs vision-language instruction tuning for zero-shot evaluation. The work uses instruction-aware visual feature extraction that enables flexible feature extraction according to given instructions by providing textual instruction to both the frozen language model and the Q-former, allowing it to extract instruction-aware features from the frozen image encoder.

GIT Wang et al. (2022) Generative Image-to-text Tranfromer unifies vision-language tasks such as image/video captioning and question answering. The method simplifies the architecture as one image encoder and one text decoder under a single language modeling task.

6 Results and Discussion

In this section, we compare the performance of our proposed method with competitors on the VQA and visual reasoning datasets evaluated on average success rate across several recent pretrained models. Table 1 shows our result on the visual question and answering task, where our method consistently outperforms other attack strategies in the multi-modality category. For the ViLT model, our proposed method achieves an ASR of 94.3 percent which is 21 percent more than the most recent state-of-the-art VLAttack and 59 percent more than Co-attack which was released in the year 2022. With the BLIP model, we reach a success rate of 91 percent, and for GIT with a success rate of 80.4 percent.

Table 2 presents similar trends for the visual reasoning dataset. For the ViLT model, we achieve an ASR of 73.32 percent which is 10.01 percent higher than the VLAttack and 33 percent higher than the Co-attack. With the BLIP we are 10.99 percent more than VLAttack and 49.75 percent more than co-attack. The results indicate the superiority of our proposed method in handling multi-modality attacks across both datasets. The performance trends highlight the effectiveness of leveraging both visual and textual information for robust VQA and visual reasoning downstream tasks.

7 Broader Impact and limitations

The development of multi-modal foundational models has transformed many sectors such as health-care, finance and many others. The unique ability of multi-modal models to accept more than one modality give more opportunities for effective and easy interaction. This work probes these models with small crafted perturbations which completely misleads models that have billions of parameters and are trained on humongous data. By building such attack strategies helps to expose the model vulnerabilities as well as generate more samples that can be efficiently utilized to fine-tune these models to improve their robustness. One of the limitation with surrogate based attacks is that, they often rely on synthetic perturbations that may not reflect the real-world perturbations, demanding more evaluation in that direction.

8 Conclusion

In conclusion, this research presents a novel coordinated attack strategy tailored for vision-language models, addressing the unique vulnerabilities posed by their multimodal nature. By aligning a surrogate model's responses with those of a text encoder, we establish a foundation for generating adversarial examples across both visual and textual modalities. Our approach, distinct from existing methods, leverages a generic surrogate model capable of crafting adversarial samples without specific victim model feedback, thereby demonstrating its versatility and applicability across various state-of-the-art vision-language models. Through experimentation on benchmark datasets and evaluation against multiple victim models, our method showcases superior performance in compromising the robustness of vision-language models compared to existing multi-modal and single-modality attack techniques. Our results underscore the effectiveness of our coordinated approach in generating adversarial perturbations that induce disparate model outputs while minimizing similarity between joint representations and corresponding answers.

References

W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, Advances in Neural Information Processing Systems 36 (2024).

H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, Y. J. Lee, Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International conference on machine learning, PMLR, 2021, pp. 5583–5594.

L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, Bert-attack: Adversarial attack against bert using bert, arXiv preprint arXiv:2004.09984 (2020).

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199 (2013).

N. Papernot, P. McDaniel, A. Swami, Crafting adversarial input sequences for recurrent neural networks, in: MILCOM 2016-2016 IEEE Military Communications Conference, IEEE, 2016, pp. 49–54.

H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, A. K. Jain, Adversarial attacks and defenses in images, graphs and text: A review, International journal of automation and computing 17 (2020) 151–178.

X. Wang, H. Wang, D. Yang, Measure and improve robustness in nlp models: A survey, arXiv preprint arXiv:2112.08313 (2021).

W. Wang, R. Wang, L. Wang, Z. Wang, A. Ye, Towards a robust deep neural network in texts: A survey, arXiv preprint arXiv:1902.07285 (2019).

W. E. Zhang, Q. Z. Sheng, A. Alhazmi, C. Li, Adversarial attacks on deep-learning models in natural language processing: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (2020) 1–41.

T. Wang, X. Wang, Y. Qin, B. Packer, K. Li, J. Chen, A. Beutel, E. Chi, Cat-gen: Improving robustness in nlp models via controlled adversarial text generation, arXiv preprint arXiv:2010.02338 (2020).

J. Lin, J. Zou, N. Ding, Using adversarial attacks to reveal the statistical bias in machine reading comprehension models, arXiv preprint arXiv:2105.11136 (2021).

K.-H. Huang, K.-W. Chang, Generating syntactically controlled paraphrases without using annotated parallel pairs, arXiv preprint arXiv:2101.10579 (2021).

W. Han, L. Zhang, Y. Jiang, K. Tu, Adversarial attack and defense of structured prediction models, arXiv preprint arXiv:2010.01610 (2020).

D. Jin, Z. Jin, J. T. Zhou, P. Szolovits, Is bert really robust? a strong baseline for natural language attack on text classification and entailment, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 8018–8025.

X. He, L. Lyu, Q. Xu, L. Sun, Model extraction and adversarial transferability, your bert is vulnerable!, arXiv preprint arXiv:2103.10013 (2021).

Z. Yang, J. Shi, J. He, D. Lo, Natural attack for pre-trained models of code, in: Proceedings of the 44th International Conference on Software Engineering, 2022, pp. 1482–1493.

S. Garg, G. Ramakrishnan, Bae: Bert-based adversarial examples for text classification, arXiv preprint arXiv:2004.01970 (2020).

L. Sun, K. Hashimoto, W. Yin, A. Asai, J. Li, P. Yu, C. Xiong, Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert, arXiv preprint arXiv:2003.04985 (2020).

L. Xu, A. Cuesta-Infante, L. Berti-Equille, K. Veeramachaneni, R&r: Metric-guided adversarial sentence generation, arXiv preprint arXiv:2104.08453 (2021).

D. Li, Y. Zhang, H. Peng, L. Chen, C. Brockett, M.-T. Sun, B. Dolan, Contextualized perturbation for textual adversarial attack, arXiv preprint arXiv:2009.07502 (2020).

M. Ye, C. Miao, T. Wang, F. Ma, Texthoaxer: Budgeted hard-label adversarial attacks on text, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 3877–3884.

G. Chang, H. Gao, Z. Yao, H. Xiong, Textguise: Adaptive adversarial example attacks on text classification model, Neurocomputing 529 (2023) 190–203.

N. Akhtar, A. Mian, N. Kardan, M. Shah, Advances in adversarial attacks and defenses in computer vision: A survey, IEEE Access 9 (2021) 155161–155196.

S. Y. Khamaiseh, D. Bagagem, A. Al-Alaj, M. Mancino, H. W. Alomari, Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification, IEEE Access 10 (2022) 102266–102291.

A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, A survey on adversarial attacks and defences, CAAI Transactions on Intelligence Technology 6 (2021) 25–45.

M. Andriushchenko, F. Croce, N. Flammarion, M. Hein, Square attack: a query-efficient black-box adversarial attack via random search, in: European conference on computer vision, Springer, 2020, pp. 484–501.

S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.

W. Zhou, S. Bai, Q. Zhao, B. Chen, Revisiting the adversarial robustness of vision language models: a multimodal perspective, arXiv preprint arXiv:2404.19287 (2024).

J. Zhang, Q. Yi, J. Sang, Towards adversarial attack on vision-language pre-training models, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5005–5013.

Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, M. Lin, On evaluating adversarial robustness of large vision-language models, Advances in Neural Information Processing Systems 36 (2024).

Z. Yin, M. Ye, T. Zhang, T. Du, J. Zhu, H. Liu, J. Chen, T. Wang, F. Ma, Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models, arXiv preprint arXiv:2310.04655 (2023).

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

E. Jang, S. Gu, B. Poole, Categorical reparameterization with gumbel-softmax, arXiv preprint arXiv:1611.01144 (2016).

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, arXiv preprint arXiv:2310.03744 (2023).

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. 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, pp. 6904–6913.

A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, Y. Artzi, A corpus for reasoning about natural language grounded in photographs, arXiv preprint arXiv:1811.00491 (2018).

M. Naseer, S. Khan, M. Hayat, F. S. Khan, F. Porikli, A self-supervised approach for adversarial robustness, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 262–271.

A. Ganeshan, V. BS, R. V. Babu, Fda: Feature disruptive attack, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8069–8079.

J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, Git: A generative image-to-text transformer for vision and language, arXiv preprint arXiv:2205.14100 (2022).