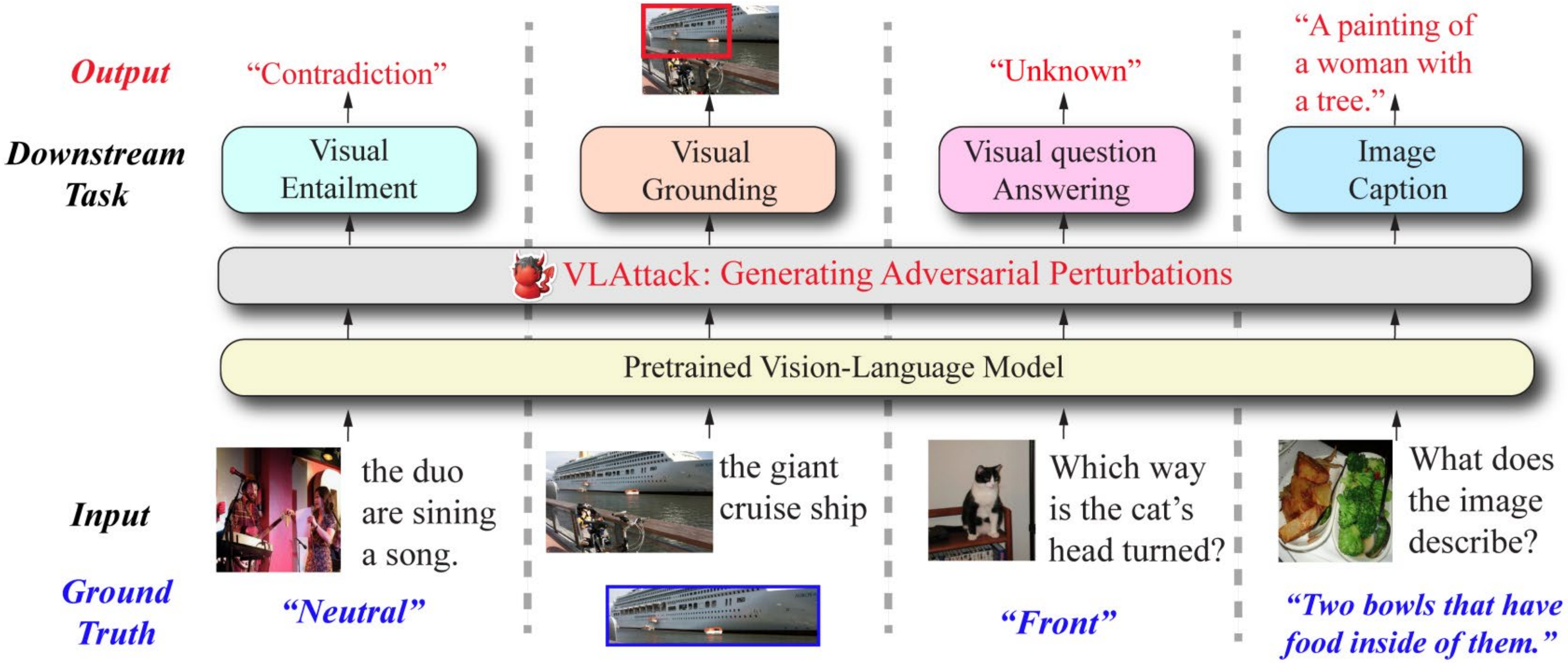
VLATTACK: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Han Liu, Jinghui Cheng, Ting Wang, Fenglong Ma

PennState

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

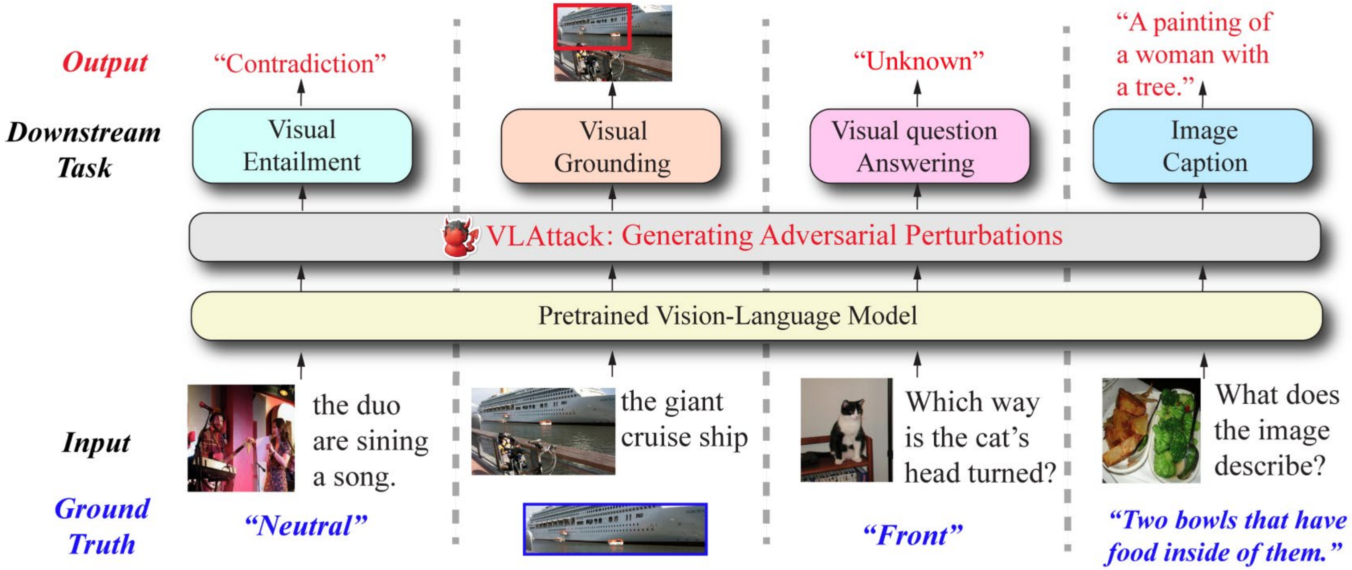


- The recent success of vision-language (VL) pre-trained models on multimodal tasks have attracted broad attention from both academics and industry. However, the adversarial robustness is still relatively unexplored.

- Therefore, we ask the following question: *Can we generate adversarial perturbations on a pre-trained VL model to attack various black-box downstream tasks fine-tuned on the pre-trained one ?*

PennState

Introduction

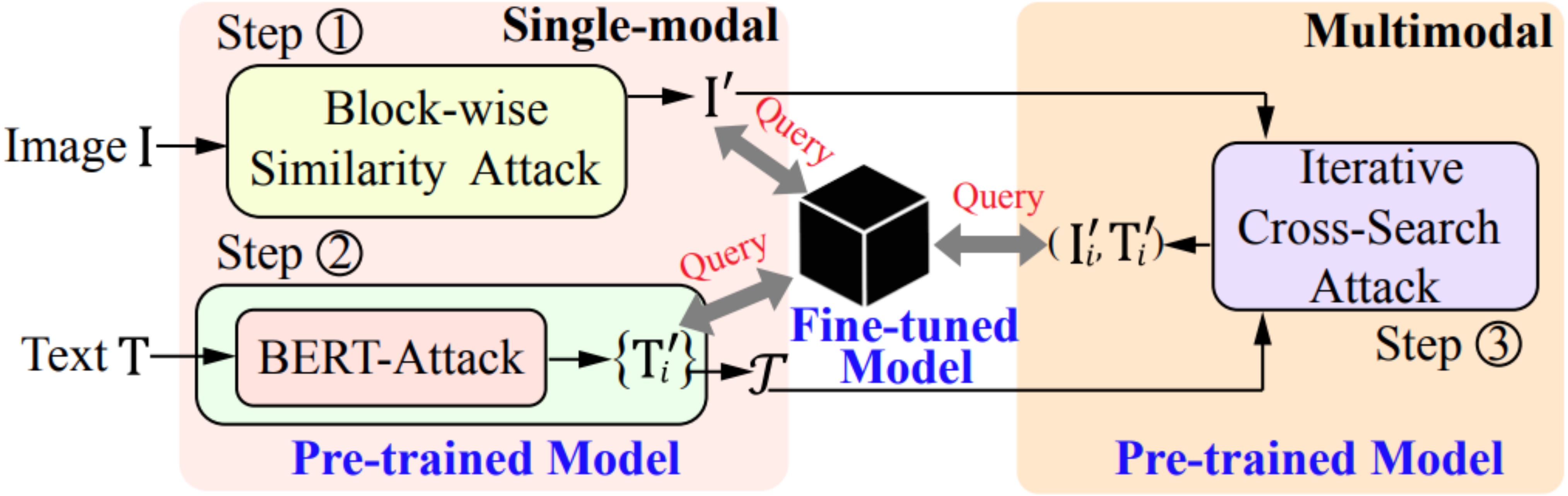


$$\max_{\mathbf{I}',\mathbf{T}'} \mathbb{1}\{S(\mathbf{I}',\mathbf{T}') \neq \mathbf{y}\}, \quad s.t. \ \|\mathbf{I}' - \mathbf{I}\|_\infty < \sigma_i, \ \ Cos(U_s(\mathbf{T}'), U_s(\mathbf{T})) > \sigma_s,$$

- ***Task-specific challenge***: The attack mechanism needs to be general and work for attacking multiple tasks.

- ***Model-specific challenge***: The attack method needs to automatically learn the transferability between pre-trained and fine-tuned models on different modalities

VLATTACK



➢ ***Single-modal Level Attack:*** Attacking using a “from image to text” order as the former can be perturbed on a continuous space. Image Attack: BSA. Text Attack: BERT-Attack[1].

➢ **Multi-modal Level Attack:** Cross-updating image and text perturbations at the multimodal level based on previous outputs.

[1] Li, Linyang, et al. "BERT-ATTACK: Adversarial Attack Against BERT Using BERT." *EMNLP* 2020.

PennState

Block-wise Similarity Attack (BSA)

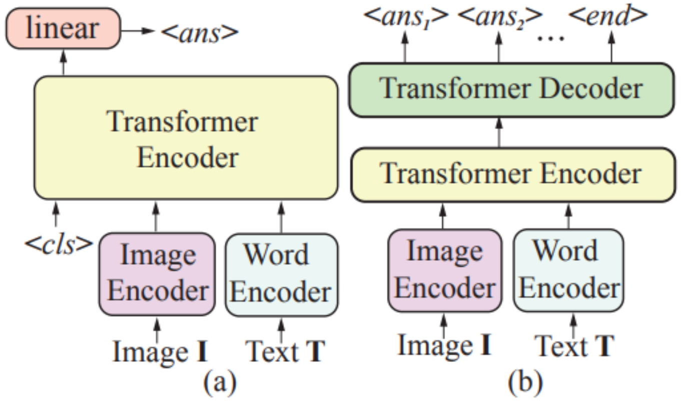


Figure 3: A brief illustration of the encoder-only (a) and encoder-decoder (b) structures.

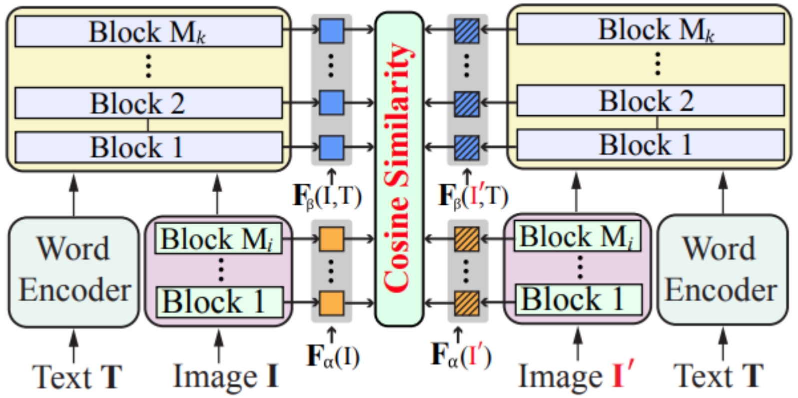


Figure 4: Block-wise similarity attack. $\mathbf{F}_\alpha$ is the image encoder, and $\mathbf{F}_\beta$ is the Transformer encoder.

$$\mathcal{L} = \underbrace{\sum_{i=1}^{M_i}\sum_{j=1}^{M_j^i} Cos(\mathbf{F}_\alpha^{i,j}(\mathbf{I}),\ \mathbf{F}_\alpha^{i,j}(\mathbf{I}^{'}))}_{\text{Image Encoder}} + \underbrace{\sum_{k=1}^{M_k}\sum_{t=1}^{M_t^k} Cos(\mathbf{F}_\beta^{k,t}(\mathbf{I},\mathbf{T}),\ \mathbf{F}_\beta^{k,t}(\mathbf{I}^{'},\mathbf{T}))}_{\text{Transformer Encoder}},$$

Algorithm Details

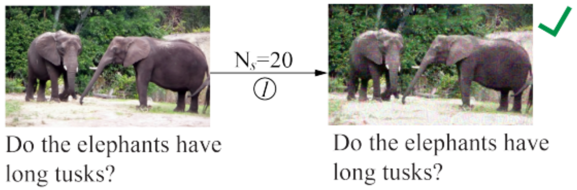


Figure 12: An adversarial image from BSA.

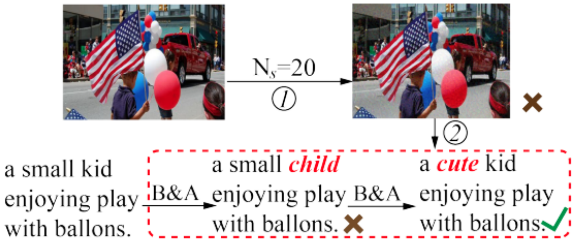


Figure 13: An adversarial sentence from text attack.

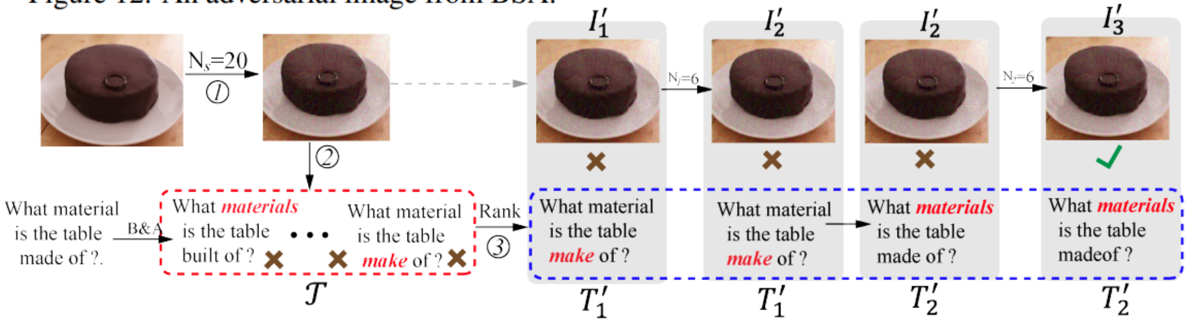


Figure 14: An adversarial image-text pair from multimodal attack.

Algorithm 1 VLATTACK

Input: A pre-trained model F, a fine-tuned model S, a clean image-text pair $(\mathbf{I}, \mathbf{T})$ and its prediction y on the S , and the Gaussian distribution $\mathcal{U}$;

Parameters: Perturbation budget σ_i on $\mathbf{I}$, σ_s on $\mathbf{T}$. Iteration number N and N_s.

//Single-modal Attacks: From Image to Text (Section 4.1)
Initialize $\mathbf{I}' = \mathbf{I} + \delta,\ \delta \in \mathcal{U}(0,1), \mathcal{T} =$
// Image attack by updating $\mathbf{I}'$ using Eq. (2) for N_s steps
$\mathbf{I}' = \mathrm{BSA}(\mathcal{L}, \mathbf{I}', \mathbf{T}, N_s, \sigma_i, F)$
if $S(\mathbf{I}', \mathbf{T}) \neq y$ **then return** $(\mathbf{I}^{'}, \mathbf{T})$
else
    // Text attack by applying BERT-attack
    for pertubed text $\mathbf{T}_i^{'}$ in BERT-attack **do**
        if $\gamma_i = Cos(U_s(\mathbf{T}_i'), U_s(\mathbf{T})) > \sigma_s$ **then**
            Add the pair $(\mathbf{T}_i', \gamma_i)$ into $\mathcal{T}$;
            if $S(\mathbf{I}, \mathbf{T}_i^{'}) \neq y$ **then return** $(\mathbf{I}, \mathbf{T}_i^{'})$
            end if
        end if
    end for
end if
// Multimodal Attack (Section 4.2)
Rank $\mathcal{T}$ according to similarity scores $\{\gamma_i\}$ and get top-K samples $\{\hat{\mathbf{T}}_1', \cdots, \hat{\mathbf{T}}_K'\}$ according to Eq. (3);
for $k = 1, \cdots, K$ **do**
    if $S(\mathbf{I}_k', \mathbf{T}_k') \neq y$ **then return** $(\mathbf{I}_k', \mathbf{T}_k')$
    end if
    Replace $(\mathbf{I}_k', \hat{\mathbf{T}}_k')$ with $(\mathbf{I}^{'}, \mathbf{T})$ in Eq. (2);
    $\mathbf{I}_{k+1}^{'} = \mathrm{BSA}(\mathcal{L}, \mathbf{I}_k^{'}, \hat{\mathbf{T}}_k', N_k, \sigma_i, F)$
    if $S(\mathbf{I}_{k+1}', \mathbf{T}_k') \neq y$ **then return** $(\mathbf{I}_{k+1}', \mathbf{T}_k')$
    end if
end for
return None

Experimets

Table 1: Comparison of VLATTACK with baselines on ViLT, Unitab, and OFA for different tasks, respectively. All results are displayed by ASR (%). B&A means the BERT-Attack approach.

Pre-trained Model	Task	Dataset	Image Only				Text Only		multimodality	
			DR	SSP	FDA	BSA	B&A	R&R	Co-Attack	VLATTACK
ViLT	VQA	VQAv2	23.89	50.36	29.27	65.20	17.24	8.69	35.13	**78.05**
	VR	NLVR2	21.58	35.13	22.60	52.17	32.18	24.82	42.04	**66.65**
BLIP	VQA	VQAv2	7.04	11.84	7.12	26.36	21.04	2.94	14.24	**49.26**
	VR	NLVR2	6.66	6.88	10.22	27.16	33.08	16.92	8.70	**52.66**
Unitab	VQA	VQAv2	22.88	33.67	41.80	48.40	14.20	5.48	33.87	**62.20**
	REC	RefCOCO	21.32	64.56	75.24	89.70	13.68	8.75	56.48	**93.52**
	REC	RefCOCO+	26.30	69.60	76.21	90.96	6.40	2.46	68.69	**93.40**
	REC	RefCOCOg	26.39	69.26	78.64	91.31	22.03	18.52	65.50	**95.61**
OFA	VQA	VQAv2	25.06	33.88	40.02	54.05	10.22	2.34	51.16	**78.82**
	VE	SNLI-VE	13.71	15.11	20.90	29.19	10.51	4.92	18.66	**41.78**
	REC	RefCOCO	11.60	16.00	27.06	40.82	13.15	7.64	32.04	**56.62**
	REC	RefCOCO+	16.58	22.28	33.26	46.44	4.66	7.04	45.28	**58.14**
	REC	RefCOCOg	16.39	24.80	33.22	54.63	19.23	15.13	30.53	**73.30**

Table 2: Evaluation of the Uni-modal tasks on OFA. We highlight the prediction score reported by the original OFA paper with ∗.

Dataset	MSCOCO				ImageNet-1K
Metric	BLEU@4 (↓)	METEOR (↓)	CIDEr (↓)	SPICE (↓)	ASR(↑)
OFA∗	42.81	31.30	145.43	25.37	-
DR	30.26	24.47	95.52	17.89	10.43
SSP	10.99	12.52	23.54	5.67	19.44
FDA	17.77	17.92	55.75	11.36	12.31
BSA (Ours)	3.04	8.08	2.16	1.50	41.35

Table 3: CLIP model evaluation on SVHN.

Dataset	SVHN	
Model	CLIP-ViT/16	CLIP-RN50
DR	3.32	71.62
SSP	6.36	84.26
FDA	6.20	83.52
BSA (Ours)	15.74	84.98

PennState

Conclusion

- Explore the adversarial vulnerability across pre-trained and fine-tuned VL models.

- We propose VLATTACK to attack from different levels.

- Extensive experiments on five VL models and six tasks.

- Currently, our research problem is formulated by assuming the pre-trained and downstream models share similar structures. The adversarial transferability between different pre-trained and fine-tuned models is worth exploring, which we left to our future work.

PennState