

Introduction

VFMs have become increasingly popular and most tasks (classification, segmentation, captioning, VQA, ...) today are tackled through pre-trained VFMs.

We introduce Task Agnostic Attacks (TAAs), which degrade performances across tasks by maximally perturbing feature representations of VFMs independently of the task. We find that feature space is easily manipulated and that our attacks are competitive with PGD task-specific attacks (TSAs).

Method

We adversarially modify images as to minimize the cosine similarity of the features extracted with the VFM

$$\mathcal{L}(\mathbf{x}_{ ext{adv}}) = rac{f(\mathbf{x})f(\mathbf{x}_{ ext{adv}})}{||f(\mathbf{x})|| \left||f(\mathbf{x}_{ ext{adv}})||
ight|}$$

Where f is the feature extraction backbone. For ViTs models, we compute the average cosine similarity across patch tokens.

Task-agnostic Attacks against Vision Foundation Models

¹University of Geneva ²University of Rennes, Inria, CNRS, IRISA

Experimental Results

Backbone	Attack	Туре	Classificat	ion abs↓ (rel↑) Segn	nentation a	abs↓ (rel↑)
ViT-S	No attack		96.	3 (0%)		81.4 (09	%)
	Class token	TAA	7.9	(92%)		19.2 (86	%)
	Patch tokens	TAA	0.1	(100%)		11.6 (97	%)
	Class+patch tokens	TAA	2.0	(98%)		13.3 (94	%)
	Classification	TSA	0.0	(100%)	19.8 (85%)		%)
	Segmentation	TSA	40.0	5 (58%)		9.2 (100	%)
ViT-B	No attack		97.	0 (0%)		80.8 (09	%)
	Class token	TAA	11.	8 (88%)		23.5 (80%)	
	Patch tokens	TAA	0.0	(100%)		7.5 (102	%)
	Class + patch tokens	TAA	2.1	(98%)	11.8 (96%)		%)
	Classification	TSA	0.0	(100%)		14.1 (93	%)
	Segmentation	TSA	43.9	9 (55%)		8.9 (100	%)
Attack	Captioning COCO					answering	•
PSNR	$ $ BLEU-4 \downarrow METEO	$\mathbf{DR} \downarrow \mathbf{F}$	ROUGE-L \downarrow	CIDEr $\downarrow \mid n$	umber↓	yes/no ↓	other \downarrow
No attack	29.6 30.3	3	59.0	131.4	72.5	95.9	76.9

Attack	Captioning COCO				Question answering VQAv2		
PSNR	BLEU-4↓	METEOR \downarrow	ROUGE-L \downarrow	$\text{CIDEr} \downarrow$	number ↓	yes/no ↓	other \downarrow
No attack	29.6	30.3	59.0	131.4	72.5	95.9	76.9
45 dB	5.8	17.3	32.2	33.0	50.6	83.4	53.4
40 dB	3.8	13.6	27.2	16.4	38.7	76.0	41.6
35 dB	1.9	9.7	22.3	3.6	25.6	67.5	28.5

Original



A cat is sleeping in front of a laptop

map

cat

yes

yes

What program is being utilized in the background on the computer? What kind of animal is this? Does this computer have Firefox installed? Is the cat sleeping?



A young boy is kneeling in the snow with a snowboard.

What is this child doing? What is the child doing? Is the person happy? Is the child happy?

snowboarding snowboarding yes yes

45 dB



a toshiba laptop with a keyboard that says home on it.

unanswerable
unanswerable
yes
no



A table with a football and a graduation cap on it.

unanswerable
walking
no
yes

Brian Pulfer^{*1}, Yury Belousov^{*1}, Vitaliy Kinakh¹, Teddy Furon², Slava Voloshynovskiy¹



Attack on segmentation

35 dB

a phone with the word this on it

adobe media encoder

unanswerable

class = horse

Table 7. Classification accuracy over ImageNette for fine-tuned models and transferable Task-Specific / Task-Agnostic Attacks.

Attack	Backbone	No FT	LoRA FT	Full FT
TAA	DiNOv2	0.2	10.4	92.4
TAA	MAE	6.2	15.6	87.3
TAA	MSN	9.9	14.0	84.5
TSA	DiNOv2	4.4	5.0	92.0
TSA	MAE	5.9	6.1	77.1
TSA	MSN	5.1	6.0	75.7



a laptop with a keyboard that says home is the new try.

adobe media encoder

unanswerable

no

no



A table with a football a graduation cap on it.



yes





jumping jumping no no



40 dB





Conclusion

In this work, we show that TAAs are possible and competitive with TSAs. Furthermore, TAAs show better transferability across tasks.

