Benchmarking Robustness of Adaptation Methods on Pre-trained Vision-Language Models

Shuo Chen[1,3*], Jindong Gu[2*], Zhen Han[1^], Yunpu Ma[1], Philip Torr[2], Volker Tresp[1,4]

LMU Munich[1], University of Oxford[2], Siemens[3],
Munich Center for Machine Learning (MCML)[4]
*equal contribution. ^corresponding author.

Project Page

1. Introduction

Project Page

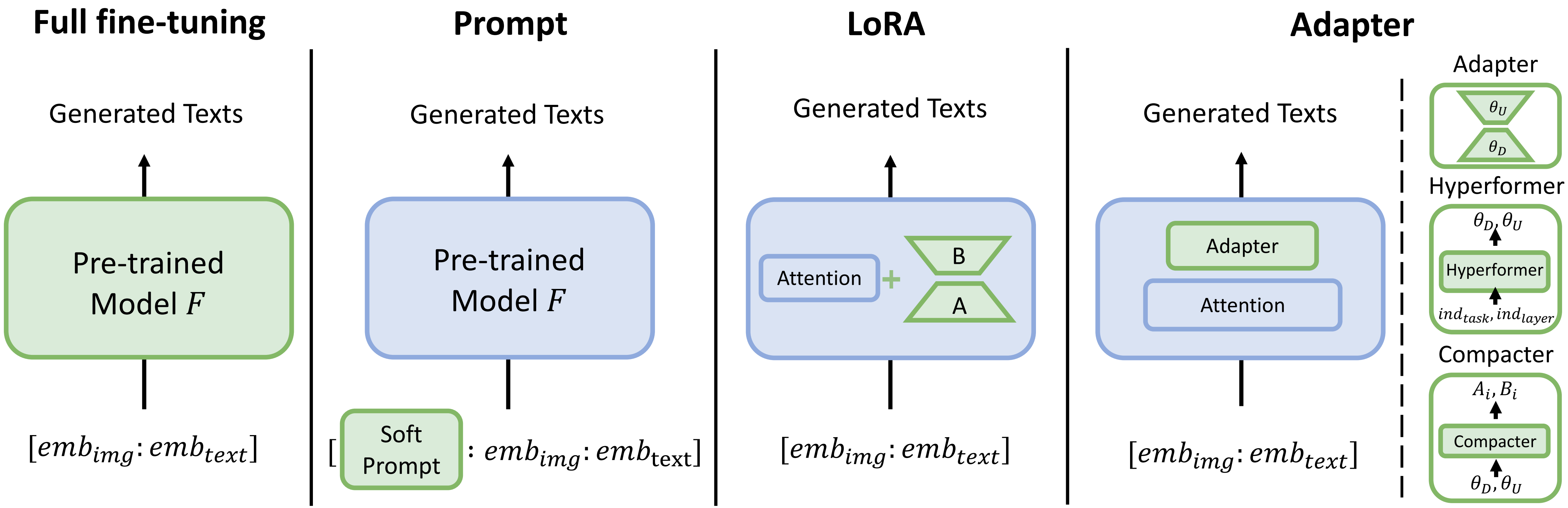


Figure 1. Various adaptation methods have been proposed to enhance the performance of pre-trained vision-language models in specific domains.

1. Introduction

Project Page

Test samples in real-world applications often
differ from the data used during pre-training and adaptation.
Model robustness is essential.

VQA during adaptation

What is this animal?

Source: https://www.pinterest.com/

VQA on test sample

What is this animal?

Source: https://www.pinterest.com/

1. Introduction

Project Page

Figure 2. Multimodal adaptation methods are sensitive to image and text corruptions.

1. Introduction

Project Page

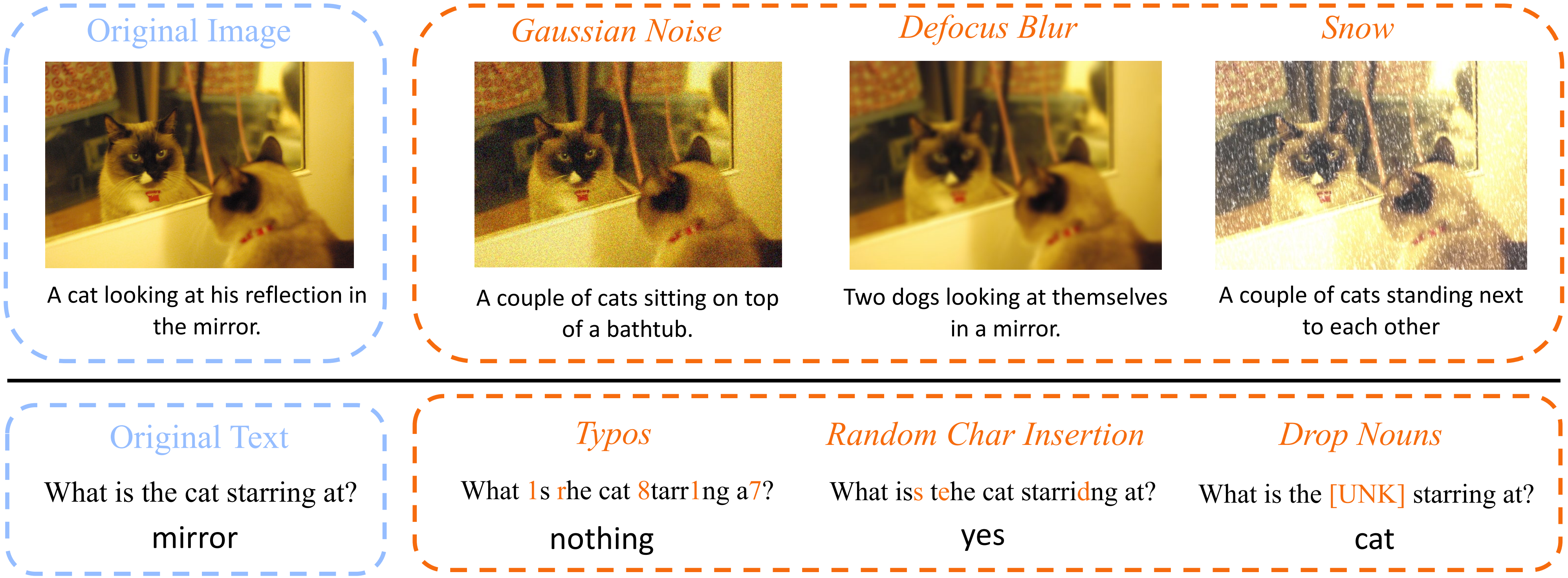


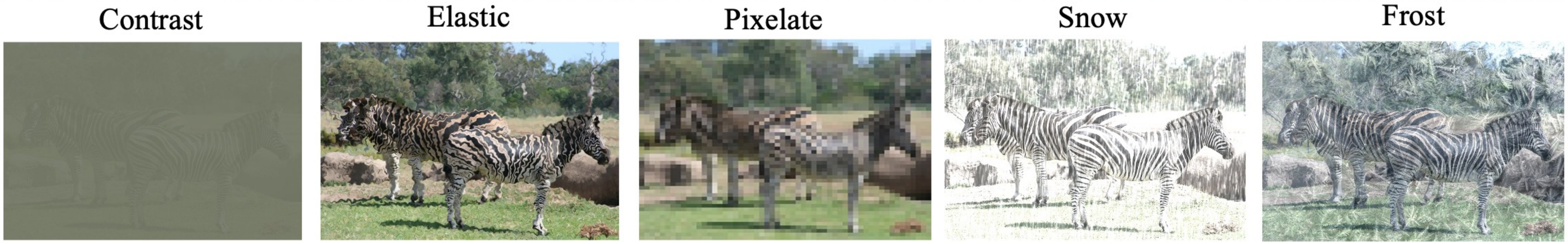
Figure 2. Multimodal adaptation methods are sensitive to image and text corruptions.

We want to know

- Which adaptation performs better on which tasks, w.r.t robustness and performance.
- Whether these methods are robust against multimodal corruptions.
- Whether more examples or more trainable parameters assure better robustness

Some examples of image and text corruption

Project Page



Some examples of image and text corruption

Project Page

Contrast | Elastic | Pixelate | Snow | Frost

Original Text

What is the cat starring at?

Typos

What 1s rhe cat 8tarr1ng a7?

Random Char Insertion

What iss tehe cat starriding at?

Drop Nouns

What is the [UNK] starring at?

2. Corruption Methods

Project Page

Some examples of image and text corruption

Contrast | Elastic | Pixelate | Snow | Frost

Original Text

What is the cat starring at?

Typos

What 1s rhe cat 8tarr1ng a7?

Random Char Insertion

What iss tehe cat starridng at?

Drop Nouns

What is the [UNK] starring at?

Original Image | Severity 1 | Zoom Blur | Severity 5

2. Corruption Methods

Project Page

96 different levels of image corruption

87 different levels of text corruption

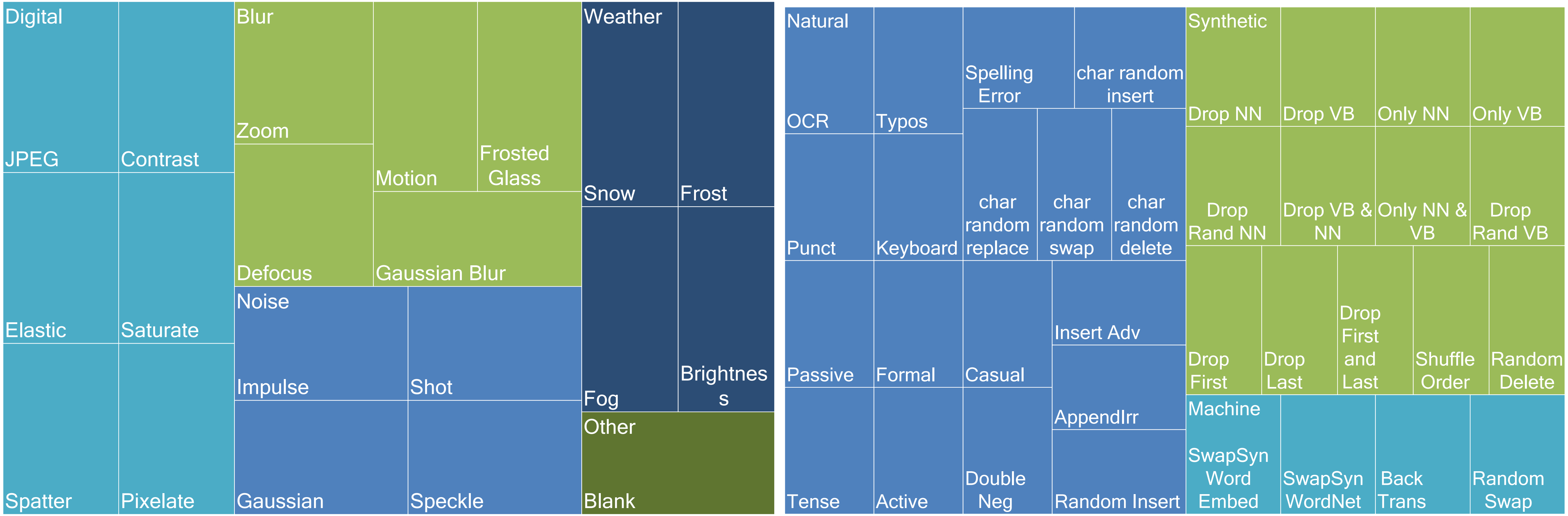


Figure 3. Corruption methods used in this study.

3. Benchmark

Project Page

	VQAv2		GQA		NLVR^2		MSCOCO Caption	
The Number of	Images	QA pairs	Images	QA pairs	Images	QA pairs	Images	Captions
Training set	113.2K	605.1K	72.1K	943.0K	103.2K	86.4K	113.2K	566.8K
Validation set	5.0K	26.7K	10.2K	132.1K	8.1K	7.0K	5.0K	5.0K
Test set	5.0K	26.3K	398	12.6K	8.1K	7.0K	5.0K	5.0K

Table 1. Dataset Statistics

https://adarobustness.github.io

We have built

11 widely used adaptation methods
20 different image corruption methods
96 different levels of image corruption
35 different text corruption methods
87 different levels of text corruption
7 out-of-distribution benchmark datasets

Relative Robustness: $RR = 1 - \frac{\Delta P}{P_I}, \Delta \mathrm{P} = (\mathrm{P_I} - \mathrm{P_o})$

P_I: performance on in-distribution dataset

P_O: performance on out-of-distribution dataset

Equation 1. Evaluation Protocol

4. Results and Analysis

Project Page

Adaptation method *Image Corruptions*	Updated Params	VQAv2 Acc (%)	VQAv2 RR (%)	GQA Acc (%)	GQA RR (%)	$NLVR^2$ Acc (%)	$NLVR^2$ RR (%)	MSCOCO Caption CIDEr	MSCOCO Caption RR (%)
Full Fine-tuning	100%	66.75	84.86$\pm$5.17	55.04	89.20$\pm$0.04	73.01	90.34$\pm$0.04	115.03	68.40$\pm$0.14
Single Adapter	4.18%	65.35	**85.76**$\pm$5.32	54.14	82.49$\pm$0.04	73.89	90.04$\pm$0.05	115.04	68.68$\pm$0.14

Adaptation method *Text Corruptions*	Updated Params	VQAv2 Acc (%)	VQAv2 RR (%)	GQA Acc (%)	GQA RR (%)	$NLVR^2$ Acc (%)	$NLVR^2$ RR (%)
Full Fine-tuning	100%	66.75	73.65$\pm$22.38	55.04	66.92$\pm$24.14	73.01	87.06$\pm$11.00
Single Adapter	4.18%	65.35	**77.64**$\pm$21.09	54.14	67.47$\pm$20.03	73.89	88.49$\pm$10.87

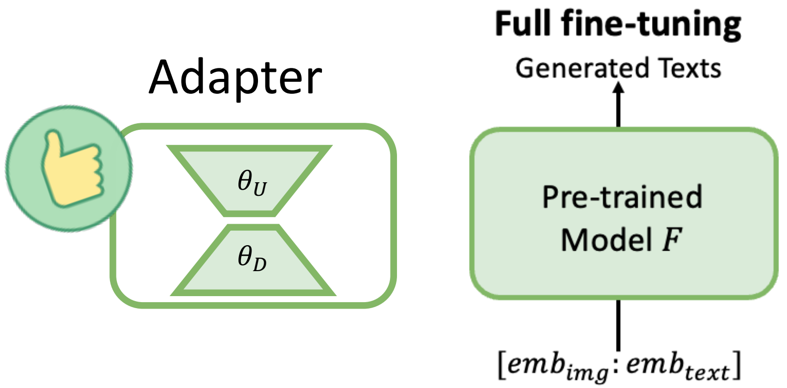
A higher sensitivity towards ***text corruptions, especially to character-level corruptions***

4. Results and Analysis

Project Page

Adaptation method *Image Corruptions*	Updated Params	**VQAv2** Acc (%)	RR (%)	**GQA** Acc (%)	RR (%)	**NLVR^2** Acc (%)	RR (%)	**MSCOCO Caption** CIDEr	RR (%)
Full Fine-tuning	100%	66.75	84.86±5.17	55.04	89.20±0.04	73.01	90.34±0.04	115.03	68.40±0.14
Single Adapter	4.18%	65.35	**85.76**±5.32	54.14	82.49±0.04	73.89	90.04±0.05	115.04	68.68±0.14

Adaptation method *Text Corruptions*	Updated Params	**VQAv2** Acc (%)	RR (%)	**GQA** Acc (%)	RR (%)	**NLVR^2** Acc (%)	RR (%)
Full Fine-tuning	100%	66.75	73.65±22.38	55.04	66.92±24.14	73.01	87.06±11.00
Single Adapter	4.18%	65.35	**77.64**±21.09	54.14	67.47±20.03	73.89	88.49±10.87



Project Page

***Language information** plays a more significant role than visual information*

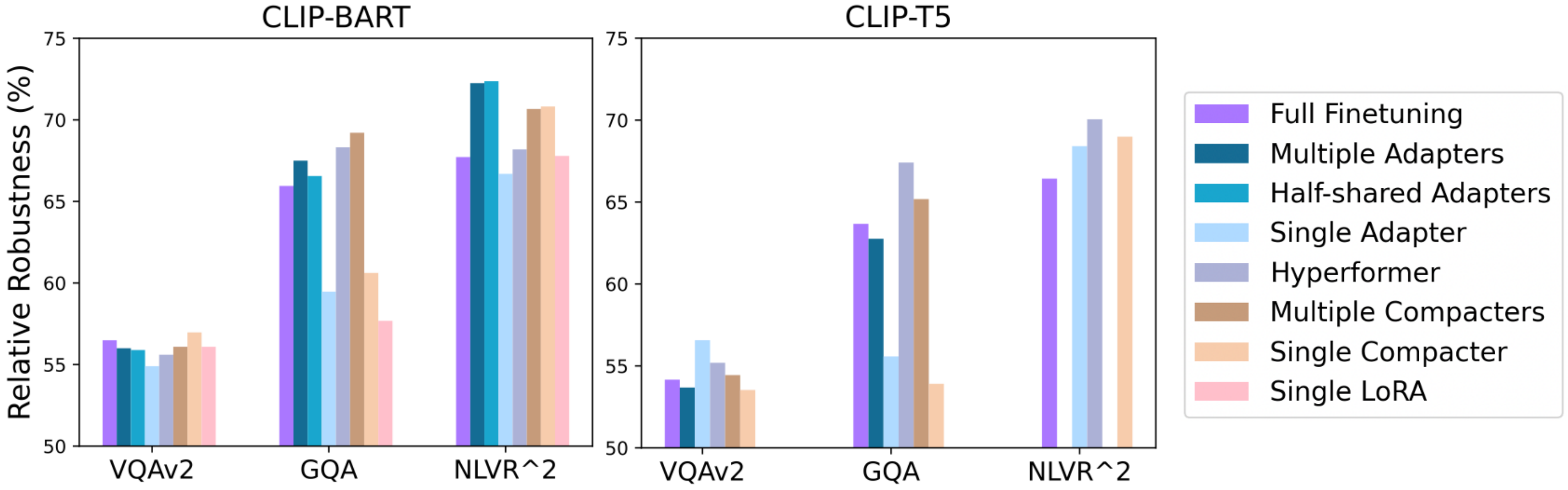


Figure 4. RR against blank-image corruption.

4. Results and Analysis

Project Page

More adaptation data ***does not consistently enhance*** *robustness.*

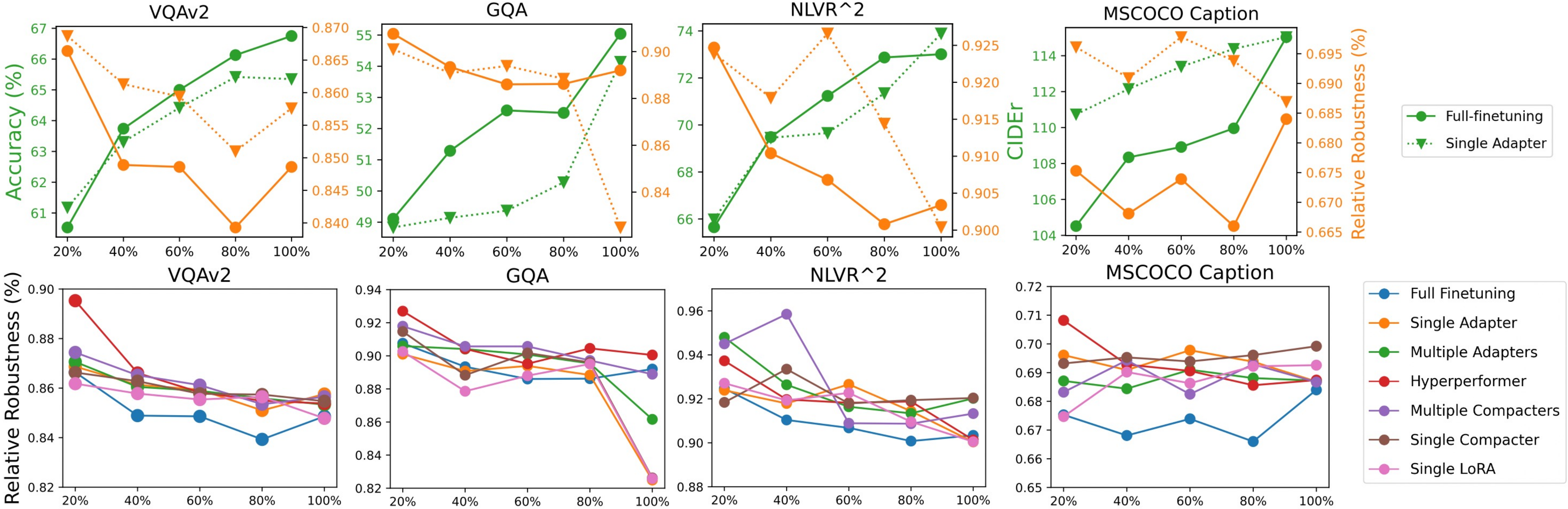


Figure 5. RR given different size of adaptation dataset.

Project Page

More parameters ***do not ensure enhanced robustness*** *and some even* ***reduce*** *it*

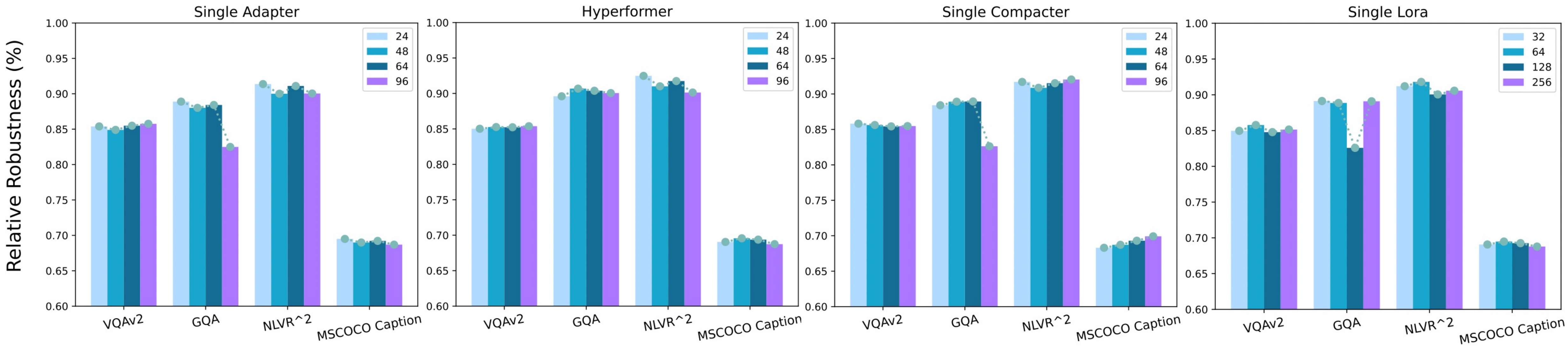


Figure 6. RR given different size of adaptation modules.

4. Results and Analysis

Project Page

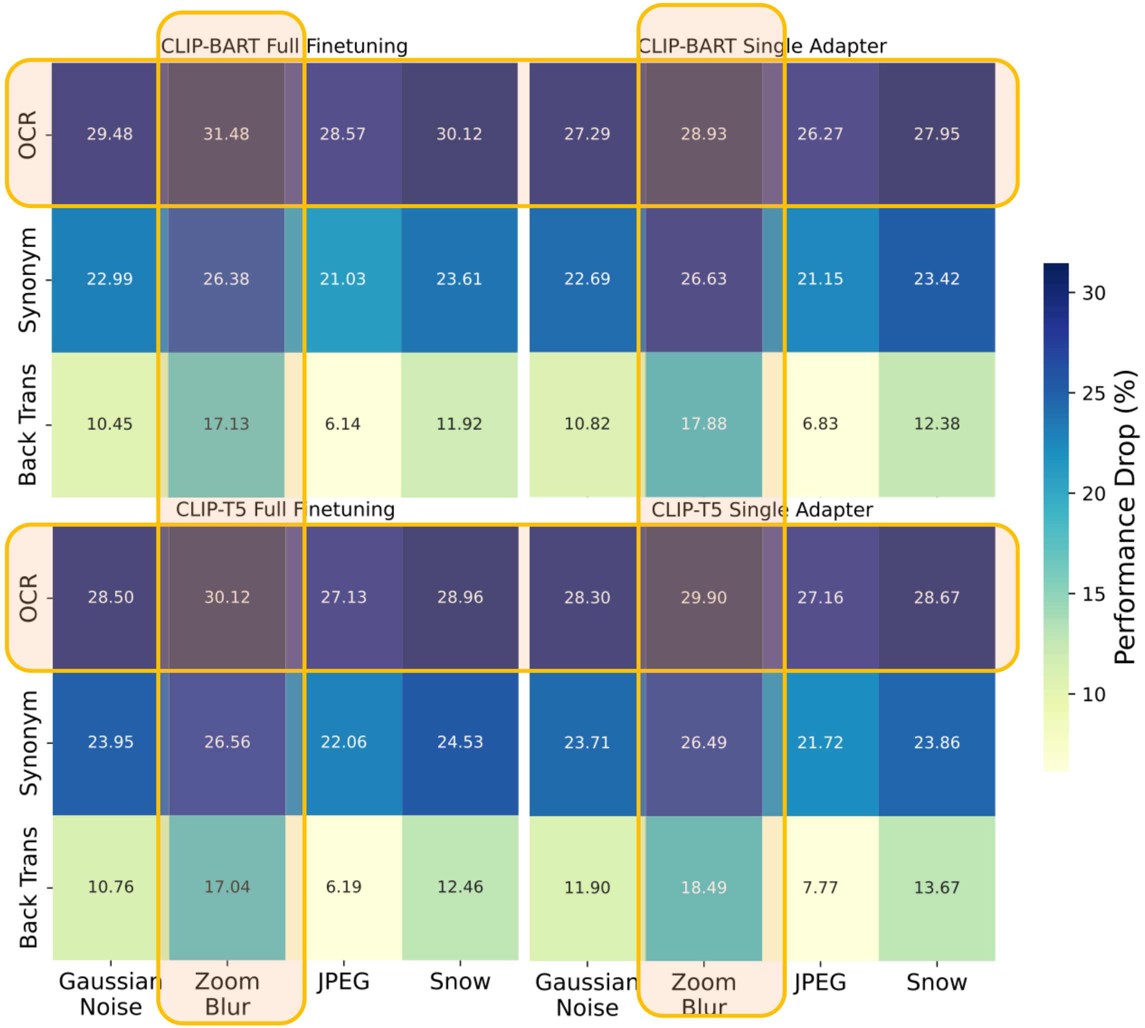


Figure 7. RR given both visual and text corruptions.

Combining corruptions from two modalities can lead to a greater drop in robustness

Project Page

Robustness against natural dataset distribution shift follows the similar conclusions.

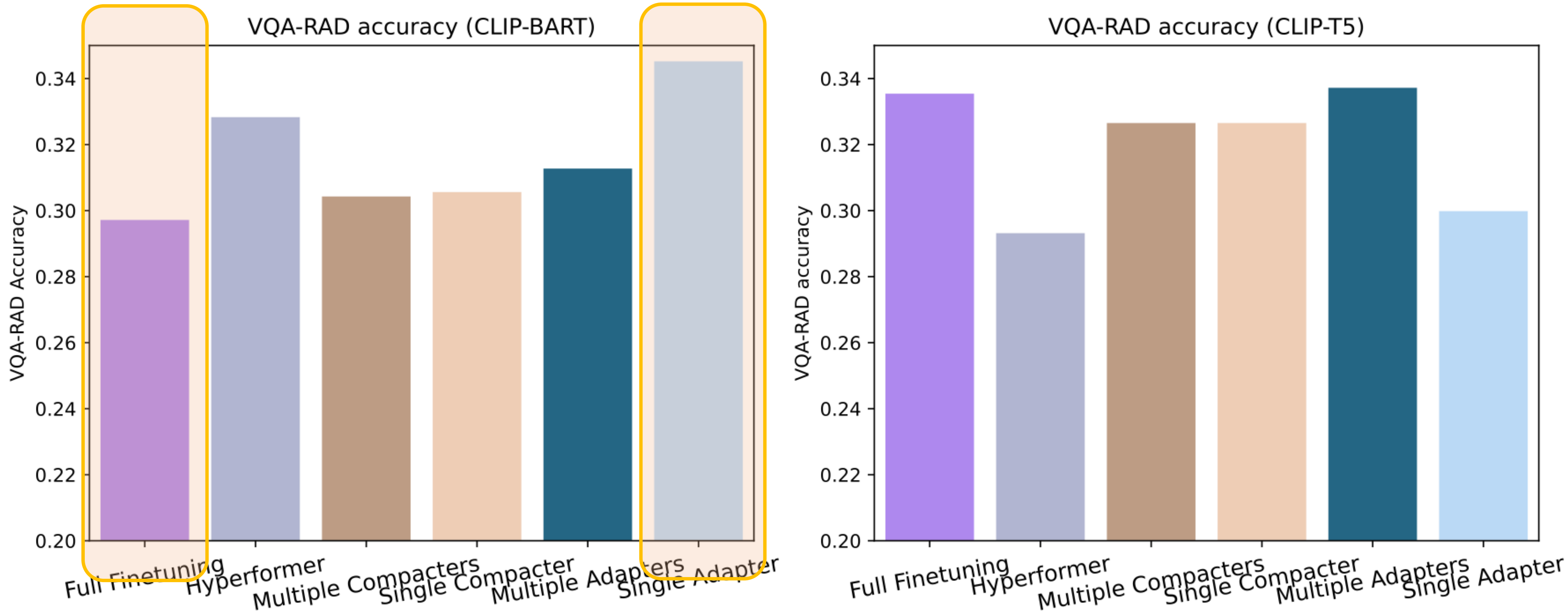


Figure 8. Performance on natural distribution shift dataset (VQA-RAD).

5. Conclusion

Project Page

We have built

11 widely used adaptation methods

20 different image corruption methods

96 different levels of image corruption

35 different text corruption methods

87 different levels of text corruption

7 out-of-distribution benchmark datasets

https://adarobustness.github.io

5. Conclusion

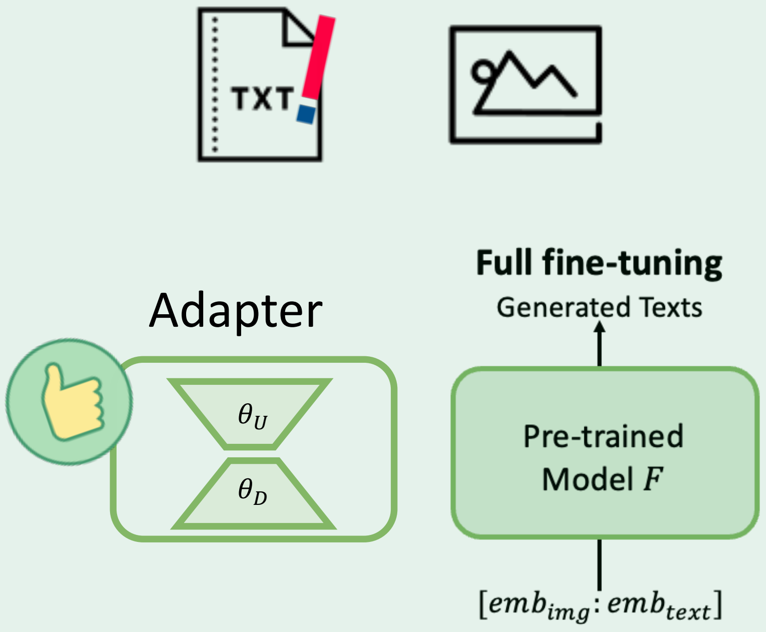
Project Page

We have built

11 widely used adaptation methods
20 different image corruption methods
96 different levels of image corruption
35 different text corruption methods
87 different levels of text corruption
7 out-of-distribution benchmark datasets

https://adarobustness.github.io

We find out



......
Check our paper for more !

Project Page

Benchmarking Robustness of Adaptation Methods on Pre-trained Vision-Language Models

Thank you!

adarobustness.github.io
chenshuo.cs@outlook.com