JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Improving Robustness of Vision Transformers by Reducing Sensitivity to Patch Corruptions

Yong Guo, David Stutz, Bernt Schiele

max planck institut informatik

Background & Comparison of perturbations

Vision transformers (ViTs) are often more robust than CNNs but still remain very vulnerable against corruptions and perturbations.

- Since ViTs are inherently patch-based, we explicitly study the sensitivity to patch corruptions/perturbations.
- We randomly sample a small number of patches to be perturbed/corrupted (10%, keeping the mask fixed) and introduce different perturbations and corruptions into the selected patches.

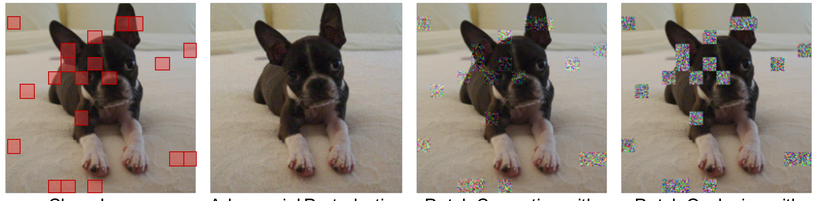


Figure 1: Impact of different corruptions and perturbations on ViTs.

Observations:

- Transformers are very sensitive to adversarial patch perturbations.
- Directly introducing corruptions yields marginal degradation but is much more efficient.
- Occluding patches with noise can significantly hamper the prediction.

Idea: Occluding patches with noise is a good proxy of adversarial patch perturbations.

Sensitivity of ViTs to Patch-based Corruptions

We seek to understand the vulnerability of ViTs by investigating the stability of the self-attention mechanism against patch-based corruptions.

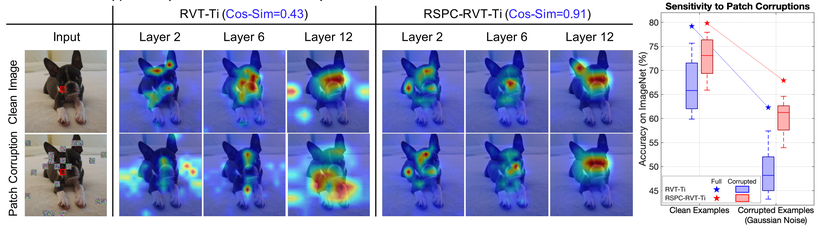


Figure 2: Sensitivity to patch corruptions in terms of attention stability (left) and accuracy (right).

Observations: The self-attention mechanism is very sensitive to patch-based corruptions, which could be a major reason for the lack of robustness.

Contributions

- We propose a new training method that **improves robustness by reducing sensitivity to patch corruptions (RSPC)**. To this end, we first construct effective **patch-based corruptions** and then reduce the sensitivity to them by aligning the intermediate features.
- When constructing patch corruptions, we develop a patch corruption model to **find particularly vulnerable patches that severely distract intermediate attention layers**. In practice, the corruption model is trained adversarially to the classification model, which, however, is essentially different from adversarial training methods
- In experiments, we demonstrate that the robustness improvement against patch corruptions can **generalize well to diverse architectures on various robustness benchmarks**. More critically, we can show, both qualitatively and quantitatively, that these improvements stem from the **more stable attention mechanism** across layers.

RSPC: Reducing Sensitivity to Patch Corruptionss

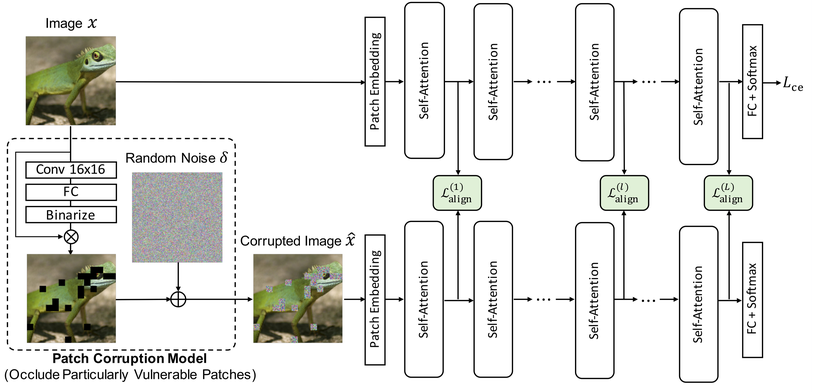


Figure 2: Overview of our reducing sensitivity to patch corruptions (RSPC) training procedure.

Key Components of the proposed RSPC:

- **Finding Vulnerable Patches to be Corrupted**:

$$\max_{\mathcal{C}} \mathbb{E}_{x\sim\mathcal{D}}\, \mathcal{L}_{\text{align}}(x,\hat{x}),$$
$$\text{where } \mathcal{L}_{\text{align}}(x,\hat{x}) = \frac{1}{L}\sum_{l=1}^{L} \|\mathcal{F}_l(x) - \mathcal{F}_l(\hat{x})\|^2. \tag{1}$$

- **Reducing Sensitivity via Feature Alignment**:

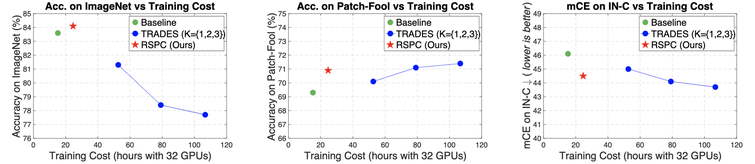
$$\min_{\mathcal{F}} \max_{\mathcal{C}} \mathbb{E}_{x\sim\mathcal{D}}[\mathcal{L}_{\text{ce}}(x) + \lambda\mathcal{L}_{\text{align}}(x,\hat{x})], \tag{2}$$

Results: Improving Corruption Robustness

- **Comparisons on ImageNet**

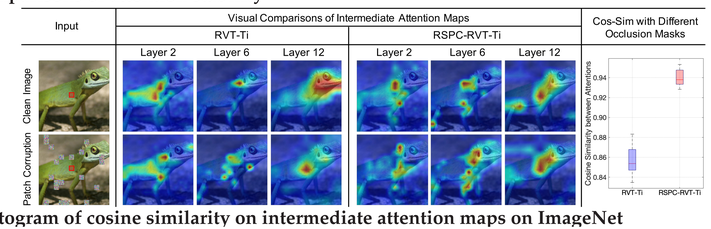
	Model	#FLOPs (G)	#Params (M)	ImageNet	Robustness Benchmarks IN-A	IN-C ↓	IN-C w/o Noise ↓	IN-P ↓
ViT-Tiny	DeiT-Ti [48]	1.3	5.7	72.2	7.3	71.1	72.9	56.7
	ConViT-Ti [11]	1.4	5.7	73.3	8.9	68.4	70.4	53.7
	RVT-Ti [34]	1.3	10.9	79.2	14.6 (+0.0)	57.0 (-0.0)	58.9 (-0.0)	39.1 (-0.0)
	+ RSPC (Ours)	1.3	10.9	79.5	16.5 (+1.9)	**55.7 (-1.3)**	**57.5 (-1.4)**	38.0 (-1.1)
	FAN-T-Hybrid [61]	3.5	7.5	80.1	21.9 (+0.0)	58.3 (-0.0)	59.8 (-0.0)	38.3 (-0.0)
	+ RSPC (Ours)	3.5	7.5	**80.3**	**23.6 (+1.7)**	57.2 (-1.1)	58.4 (-1.4)	**37.3 (-1.0)**
ViT-Small	DeiT-S [48]	4.6	22.1	79.9	6.3	54.6	56.6	36.9
	ConViT-S [11]	5.4	27.8	81.5	18.9	49.8	52.1	35.8
	RVT-S [34]	4.7	23.3	81.9	25.7 (+0.0)	49.4 (-0.0)	51.6 (-0.0)	35.2 (-0.0)
	+ RSPC (Ours)	4.7	23.3	82.2	27.9 (+2.2)	48.4 (-1.0)	50.4 (-1.2)	34.3 (-0.9)
	FAN-S-Hybrid [61]	6.7	25.7	83.5	33.9 (+0.0)	48.5 (-0.0)	50.7 (-0.0)	34.5 (-0.0)
	+ RSPC (Ours)	6.7	25.7	**83.6**	**36.8 (+2.9)**	**47.5 (-1.0)**	**49.4 (-1.3)**	**33.5 (-1.0)**
ViT-Base	MAE (ViT-B) [21]	17.6	86.6	83.6	35.9	51.7	-	-
	DeiT-B [48]	17.6	86.6	82.0	27.4	48.5	50.9	32.1
	ConViT-B [11]	17.7	86.5	82.4	29.0	46.9	49.3	32.2
	RVT-B [34]	17.7	91.8	82.6	28.5 (+0.0)	46.8 (-0.0)	49.8 (-0.0)	31.9 (-0.0)
	+ RSPC (Ours)	17.7	91.8	82.8	32.1 (+3.6)	45.7 (-1.1)	48.5 (-1.3)	31.0 (-0.8)
	FAN-B-Hybrid [61]	11.3	50.5	83.9	39.6 (+0.0)	46.1 (-0.0)	48.1 (-0.0)	31.3 (-0.0)
	+ RSPC (Ours)	11.3	50.5	**84.2**	**41.1 (+1.5)**	**44.5 (-1.6)**	**46.8 (-1.3)**	**30.0 (-1.2)**

- **Comparisons with Adversarial Training**



Visualization Results & More Discussions

- **Comparisons of attention stability**



- **Histogram of cosine similarity on intermediate attention maps on ImageNet**