A. Saliency attacks

Such 'salient' pixels are often identified using the magnitudes of gradients. This idea, while not particularly new [61], lends itself naturally to constructing adversarial attacks. Specifically, the idea is to only perturb a subset of the salient pixels thus implicitly satisfying the sparsity constraint. JSMA [54] and Maximal-JSMA [55] leverage this observation to construct k-sparse attacks by maximally perturbing k salient pixels. In maximal-JSMA, the authors calculate saliency of each pixel usign the following equation;

$$S^{+}(x_{i,c}) = \begin{cases} 0 \text{ if } \frac{\partial f(\mathbf{x})_c}{\partial x_i} < 0 \text{ or } \sum_{c' \neq c} \frac{\partial f(\mathbf{x})'_c}{\partial x_i} \\ -\frac{\partial f(\mathbf{x})_c}{\partial x_i} \cdot \sum_{c' \neq c} \frac{\partial f(\mathbf{x})'_c}{\partial x_i} \text{ otherwise,} \end{cases} \tag{1}$$

where x_i is the pixel in question, c is the true class, and f_i is a logit value specific to class i.

In this paper, we use a patch based block sparse attack [10], where the attack budget is defined by the number of patches (blocks) the attacker is allowed to perturb. This approach builds on JSMA [54] Maximal-JSMA [55], wherein the attacker identifies top salient pixels using gradients and perturb them to create attacks. A similar idea is extended to block sparsity. The main differences between JSMA and this approach lie in two places: (1) This uses a simplified construction for the saliency map that relies on the magnitude of the gradients with respect to each pixel, (2) instead of considering salient pixels, this instead identifies the most informative pixel blocks and further rely on gradient updates to generate an attack.

B. Experiments

For all experiments, we use SGD for optimization with a learning rate of 0.1 for a maximum of 100 steps.

C. Detailed Results

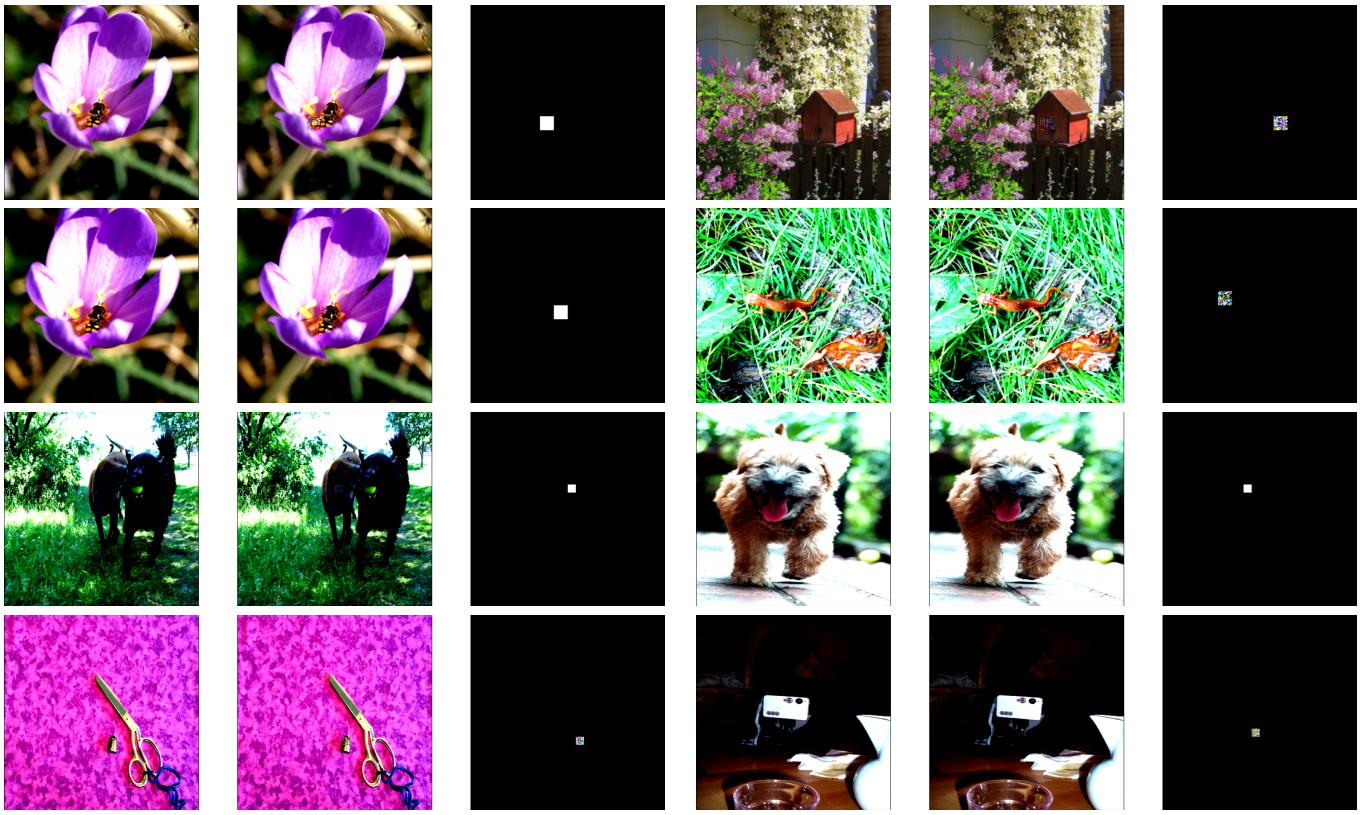

Figure 4. ***Patch attacks on Transformers***: *The attack images are generated with a fixed budget of* 1 *patch.*

Table 3. *Robust Accuracy vs Token Budget.* The models are attacked with a $p = 16$. Note that for smaller token budgets, the models perform nearly the same. However, as the token budget increases, Convnets are more robust than Vision Transformers.

Model	**Clean**	**Token Budget**				
		1	2	5	10	20
BEiT-Base-224	84.69	54.46	35.09	10.30	2.11	0.12
BEiT-Large-224	87.34	67.58	52.61	24.76	7.67	0.85
BEiTv2-Base-224	86.27	52.05	37.91	17.71	7.12	1.61
BEiTv2-Large-224	88.34	61.23	44.23	20.32	7.18	0.95
Conv2-Base	86.72	64.47	47.81	20.10	6.54	0.98
Conv2-Huge	88.48	61.76	44.49	18.32	4.20	0.49
Conv2-Large	86.89	65.45	48.29	24.15	9.63	2.19
ConvViT-Base	82.18	26.44	7.72	0.21	0.02	0.01
ConvViT-Small	81.28	10.62	1.83	0.05	0.01	0.01
ConvViT-Tiny	73.48	4.03	0.24	0.01	0.01	0.01
DeiT224-Distill	83.16	44.03	19.54	0.98	0.01	0.01
DeiT3-Base-224	83.61	44.87	11.88	0.20	0.01	0.01
DeiT3-Huge-224-14	85.07	21.87	2.48	0.05	0.01	0.01
DeiT3-Large-224	84.62	55.27	16.07	0.33	0.02	0.01
DeiT3-Medium-224	82.86	46.59	17.61	0.67	0.01	0.01
DeiT3-Small-224	81.46	21.34	4.88	0.25	0.01	0.01
ResNet101-D	82.10	54.53	39.62	19.28	3.17	0.37
ResNet50	80.10	38.33	20.78	4.65	0.37	0.04
Wide Resnet	78.33	34.59	17.05	3.22	0.19	0.02
SWIN-224	82.90	69.11	58.36	30.73	5.42	0.05
SWIN-224-Base	85.11	73.11	62.33	34.48	7.81	0.06
SWIN-224-Large	86.24	73.43	61.52	27.07	3.04	0.03
ViT-224	85.03	59.46	39.95	10.46	0.70	0.01
CLIP-ViT/B-16	66.83	1.48	0.14	0.02	0.01	0.01
CLIP-ViT/L-14	73.54	3.61	1.27	0.28	0.11	0.04
CLIP-RN50	58.36	6.55	1.97	0.23	0.01	0.01
CLIP-RN101	61.18	7.24	2.53	0.33	0.03	0.01
CLIP-Finetuned	80.16	37.10	20.26	3.53	0.18	0.01

Table 4. *Robust Accuracy vs Patch Size.* The models are attacked with a $K = 5$. Note that for smaller patch sizes, the models perform nearly the same. However, as the patch size increases, Convnets are more robust than Vision Transformers.

Model	**Clean**	**Patch size**			
		1	4	8	16
BEiT-Base-224	84.69	81.63	66.66	45.07	10.30
BEiT-Large-224	87.34	84.96	74.17	57.21	24.76
BEiTv2-Base-224	86.27	83.80	71.06	48.22	17.71
BEiTv2-Large-224	88.34	86.71	76.97	54.77	20.32
Conv2-Base	86.72	84.31	73.36	52.71	20.10
Conv2-Huge	88.48	87.52	74.53	52.15	18.32
Conv2-Large	86.89	84.93	75.72	55.84	24.15
ConvViT-Base	82.18	79.02	65.53	17.74	0.21
ConvViT-Small	81.28	77.26	62.92	9.27	0.05
ConvViT-Tiny	73.48	68.43	52.32	6.55	0.01
DeiT224-Distill	83.16	79.84	65.73	36.70	0.98
DeiT3-Base-224	83.61	80.68	64.86	25.42	0.20
DeiT3-Huge-224-14	85.07	83.24	64.31	8.02	0.05
DeiT3-Large-224	84.62	82.49	68.56	26.84	0.33
DeiT3-Medium-224	82.86	79.60	68.02	35.38	0.67
DeiT3-Small-224	81.46	77.00	46.47	12.62	0.25
ResNet101-D	82.10	79.17	63.44	44.65	19.28
ResNet50	80.10	74.54	46.88	22.89	4.65
Wide Resnet	78.33	70.06	37.60	17.71	3.22
SWIN-224	82.9	80.83	74.02	61.64	30.73
SWIN-224-Base	85.11	83.19	77.85	65.07	34.48
SWIN-224-Large	86.24	84.67	78.33	63.73	27.07
ViT-224	85.03	82.42	68.23	51.07	10.46
CLIP-ViT/B-16	66.83	55.50	7.62	0.32	0.02
CLIP-ViT/L-14	73.54	64.28	15.17	2.47	0.28
CLIP-RN50	58.36	45.22	10.87	1.84	0.23
CLIP-RN101	61.18	46.82	11.99	2.24	0.33
CLIP-Finetuned	80.16	72.64	44.48	19.82	3.53