A E_8 LATTICE CODEBOOK

Other than the product quantization codebook, we also use the $E8$ lattice as a codebook. The $E8$ lattice is a special lattice in $\mathbb{R}^8$ that is not data-dependent and does not need training. Due to its definition, the $E8$ lattice only deals with 8-dimensional vectors. This is not convenient for input images, but we do use it on features, flattening each patch tensor and decomposing it into 8-dimensional subvectors. The parameter s controlling the scale of the lattice determines the density of codewords. We let $s \in \{0.1, 0.2, \ldots, 0.9\}$.

Results As shown in Figure 11, the behavior is similar to PQ codebooks, in the sense that earlier layers work better and a fine codebook (small s) preserves original accuracy but loses in adversarial accuracy, as if patch replacement missing (bottom right). Since the $E8$ lattice does not need training, it can be used to quickly explore the properties of patch replacement. However, the best $E8$ codebook for layer 1 is worse (original 68.1%, adversarial 50.2%) than the best PQ codebook for layer 1 (original 65.7%, adversarial 53.7%) in Figure 5. Replacement strategies can partially recover original accuracy but not adversarial accuracy. Hence, after quick exploration, we switch to PQ codebooks for performance.

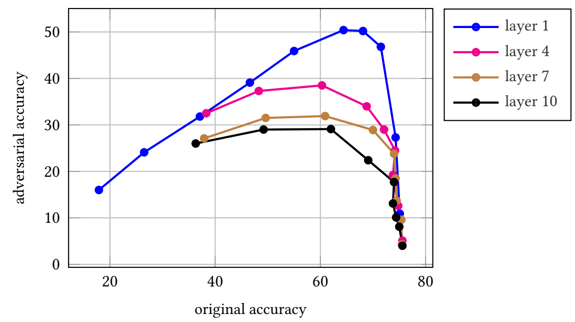

Figure 11: Effect of $E8$-lattice codebook quality, controlled by scale parameter s, applying patch replacement on different layers independently, using the plain strategy. We plot accuracy for varying s in the same curve per layer.

Method	Ori	PGD [19]			DDN [28]			BPDA [2]		
	Acc	Acc	P_{suc}	$\overline{D}$	Acc	P_{suc}	$\overline{D}$	Acc	P_{suc}	$\overline{D}$
Baseline	75.7	3.80	1.00	3.12	0.10	1.00	0.53	3.80	1.00	3.12
Patch replacement (ours)	71.8	**46.4**	0.83	**7.21**	**66.0**	**0.08**	**0.57**	**48.3**	**0.89**	**12.89**
Adv. training [19]	45.9	44.3	**0.65**	3.81	19.0	0.58	0.32	–	–	–
Bit3 [8]	64.7	32.9	0.98	6.30	55.1	0.15	0.49	0.9	1.00	1.00
Bit5 [8]	**74.9**	6.50	1.00	3.81	18.9	0.75	0.53	1.9	1.00	1.00
Ms2 [36]	74.2	26.5	0.92	5.62	47.9	0.33	0.51	3.3	1.00	1.76
Ms3 [36]	71.8	34.2	0.84	5.45	55.6	0.23	0.53	1.6	1.00	1.25
Pixel deflection [23]	73.2	31.0	1.00	6.70	58.6	0.20	0.53	1.3	1.00	0.89

Table 3: Original accuracy, adversarial accuracy, success rate (P_{suc}) and average distortion ($\overline{D}$) for combinations of defenses (adversarial training and transformation-based) and attacks, including gray-box (FGSM, BIM, PGD, DDN) and white-box (BPDA).