D4: Detection of Adversarial Diffusion Deepfakes Using Disjoint Ensembles

Ashish Hooda[†*], **Neal Mangaokar**[‡*], Ryan Feng[‡] , Kassem Fawaz[†] , Somesh Jha[†] , and Atul Prakash[‡]
† University of Wisconsin–Madison ‡ University of Michigan
{ahooda,kfawaz,sjha}@wisc.edu {nealmgkr,rtfeng,aprakash}@umich.edu

1. Appendix

1.1. Proofs

1.1.1 Orthogonal Gradients

Prior to providing proofs for the adversarial dimensions, we demonstrate that gradients for disjoint classifiers are always orthogonal to each other. We will use this for our later results. Given input space $\mathcal{X}$ and class-label space $\mathcal{Y}$, we have n disjoint classifiers $\mathcal{F}_1, ..., \mathcal{F}_n$. If T is the DCT transformation matrix, we can define T_i to be the transformation matrix for the classifier $\mathcal{F}_i$. Each matrix T_i has a lot of zeros. Only the rows corresponding to the unmasked frequencies of classifier $\mathcal{F}_i$ have non-zero entries. Moreover, since no frequency is shared by any two classifiers, the j^{th} row will have non-zero entries in exactly one of the n disjoint transformation matrices, i.e. $T_i T_j^\intercal = O \; \forall i \neq j$.

Next, the n disjoint classifiers $\mathcal{F}_1, ..., \mathcal{F}_n$, where $\mathcal{F}_i : T_i\mathbf{x} \rightarrow y$, are trained using loss functions $\mathcal{L}_{\mathcal{F}_1}, ..., \mathcal{L}_{\mathcal{F}_n}$ respectively. Now, the dot product between the gradients of classifiers $\mathcal{F}_i$ and $\mathcal{F}_j$ is given by

$$\begin{aligned}(\nabla_{\mathbf{x}}\mathcal{L}_{\mathcal{F}_i})^\intercal \left(\nabla_{\mathbf{x}}\mathcal{L}_{\mathcal{F}_j}\right) &= (T_i^\intercal \nabla_{T_i\mathbf{x}}\mathcal{L}_{\mathcal{F}_i})^\intercal \left(T_j^\intercal \nabla_{T_j\mathbf{x}}\mathcal{L}_{\mathcal{F}_j}\right) \\ &= (\nabla_{T_i\mathbf{x}}\mathcal{L}_{\mathcal{F}_i})^\intercal \, T_i T_j^\intercal \left(\nabla_{T_j\mathbf{x}}\mathcal{L}_{\mathcal{F}_j}\right) \\ &= 0\end{aligned} \tag{1}$$

1.1.2 Proof of Lemma 3.1

From [2], we know that for a classifier $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ where $\mathcal{X} \in \mathbb{R}^d$ is the input space and $\mathcal{Y}$ is the finite class label space, the dimension of the adversarial subspace around input-label pair $(\mathbf{x}, y)$ where $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, is approximated by the maximal number of orthogonal perturbations $\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_k}$ such that $||\mathbf{r_i}||_2 \leq \epsilon$ and $\mathbf{g}^\intercal \mathbf{r_i} \geq \gamma \; \forall \; 1 \leq i \leq k$. Here, $\mathbf{g} = \nabla_{\mathbf{x}} L(\mathcal{F}(\mathbf{x}), y)$ and γ is the increase in loss function L sufficient to cause a mis-classification. [2] provide a tight bound for k:

$$k = \min\left(d, \left\lfloor \frac{\epsilon^2 ||\mathbf{g}||_2^2}{\gamma^2} \right\rfloor\right) \tag{2}$$

We now extend this result for n disjoint classifiers. Let $\mathbf{g}' = \frac{\sum_{j=1}^{n} \mathbf{g_j}}{n}$.
Now, for *at-least-one voting*,

$$\mathbf{g}'^\intercal \mathbf{r_i} = \frac{\sum_{j=1}^{n} \mathbf{g_j}^\intercal \mathbf{r_i}}{n} \geq \frac{\sum_{j=1}^{n} \gamma_j}{n} \quad \forall \; 1 \leq i \leq k \tag{3}$$

Applying the result from [2] (Equation 2) on the above inequality (Equation 3), we get:

$$\begin{aligned} k &= \min\left(d, \left\lfloor \frac{\epsilon^2 n^2 ||\mathbf{g}'||_2^2}{\left(\sum_{j=1}^{n} \gamma_j\right)^2} \right\rfloor\right) \\ &= \min\left(d, \left\lfloor \frac{\epsilon^2 \sum_{j=1}^{n} ||\mathbf{g_j}||_2^2}{\left(\sum_{j=1}^{n} \gamma_j\right)^2} \right\rfloor\right). \end{aligned} \tag{4}$$

(since $\mathbf{g_i}^\intercal \mathbf{g_j} = 0 \;\; \forall i \neq j$, using Equation 1)

Now, for *majority voting*, we again apply the results from [2] (Equation 2). However, the derivation now depends on the selection of $\lceil \frac{n}{2} \rceil$ models that the adversary chooses to target. To obtain the lower and upper bounds, we can select $\lceil \frac{n}{2} \rceil$ with the most and least adversarial dimensions respectively. Following a similar derivation as before, we get :

[*] Indicates equal contribution

$$k \geq \min\left(d, \left\lfloor \min_{|K|=\lceil \frac{n}{2} \rceil} \frac{\epsilon^2 \sum_{j=1}^{n} ||\mathbf{g_j}||_2^2}{\left(\sum_{j=1}^{n} \gamma_j\right)^2} \right\rfloor\right) \tag{5}$$

$$k \leq \min\left(d, \left\lfloor \max_{|K|=\lceil \frac{n}{2} \rceil} \frac{\epsilon^2 \sum_{j=1}^{n} ||\mathbf{g_j}||_2^2}{\left(\sum_{j=1}^{n} \gamma_j\right)^2} \right\rfloor\right) \tag{6}$$

1.1.3 Proof of Lemma 3.2

Follow up work from [1] also provides a tight bound for the adversarial dimension in the ℓ_∞ case. They provide a tight bound for the number of k orthogonal perturbations $\mathbf{r_1}, ..., \mathbf{r_k} \in \mathbb{R}^d$ such that $||\mathbf{r_i}||_\infty \leq \epsilon$, given by $sign(\mathbf{g})^\intercal \mathbf{r_i} = \frac{\epsilon d}{\sqrt{k}}$ $\forall 1 \leq i \leq k$ where $sign(\mathbf{g})$ is the signed gradient.

We now extend this result for n disjoint classifiers. For $\mathbf{g}' = \frac{\sum_{j=1}^{n} \mathbf{g_j}}{n}$, since $\mathbf{g_j}'s$ are non-zero only on non-overlapping dimensions, we can see that $sign(\mathbf{g}')^\intercal r = \sum_{j=1}^{n} sign(\mathbf{g_j})^\intercal r$ $\forall \mathbf{r} \in \mathbb{R}^d$. Applying the above results here, we get

$$\sum_{j=1}^{n} sign(\mathbf{g_j})^\intercal \mathbf{r_i} = \frac{\epsilon d}{\sqrt{k}} \;\forall 1 \leq i \leq k \tag{7}$$

Now, similar to [1], we compute the perturbation magnitude along a random permutation of the signed gradient. For each $1 \leq j \leq n$ and $1 \leq i \leq k$, we get :

$$\begin{aligned} \mathbb{E}[\mathbf{g_j}^\intercal \mathbf{r_i}] &= \mathbb{E}\left[\sum_{p=1}^{d} |g_j^{(p)}| \cdot sign(g_j^{(p)}) \cdot r_i^{(p)}\right] \\ &= \sum_{p=1}^{d} |g_j^{(p)}| \mathbb{E}\left[sign(g_j^{(p)}) \cdot r_i^{(p)}\right] \\ &= \frac{\epsilon ||\mathbf{g_j}||_1}{n\sqrt{k}} \end{aligned} \tag{8}$$

1.2. Saliency Distribution

Saliency of a feature may be viewed as a heuristic measure of its "robustness", as larger saliencies imply that the model is more sensitive to perturbations of that frequency. Figure 2 plots the distribution of absolute saliency values for D4 (SIZE=1) and D4 (SIZE=4) ensembles. We observe that saliencies for saliency-partitioning configurations D4 (SIZE=4) are of relatively lower values, and are *sharply concentrated around their mode*, implying higher feature robustness. This can be attributed to the round-robin, equal distribution of robust frequencies amongst the constituent models. Improved approaches to saliency partitioning could increase this separation, improving model robustness even further. We leave this exploration to future work.

Detector	LDM	DDIM	PNDM	ProGAN
CNNDet	66% (100%)	64% (100%)	67% (100%)	**97%** (100%)
D4 **(SIZE=4)**	**93% (28%)**	**79% (4%)**	**93% (33%)**	73% (68%)
Both	**93%** (29%)	**79% (4%)**	**93% (33%)**	83% (**64%**)

Table 1. Generalization of CNNDet (trained on ProGAN) and D4 (SIZE=4) (trained on LDM) to CelebaHQ diffusion and GAN deepfakes that were unseen during training. Results are presented in the following format: non-adversarial AP (ASR).

1.3. Generalization to Unseen Image Domains and Unseen Generative Models

Tab. 1 presents the generalization results, but for the CelebaHQ dataset.

1.4. Frequency Distribution

D4 partitions frequency features based on saliency. In Fig. 1, we plot how these partitions are distributed among low or high frequencies for a D4 (SIZE=4) ensemble. Since an image has a 2-dimensional frequency spectrum, we look at four quadrants of the spectrum corresponding to low and high frequency regions for each axis. We observe that saliency based partitioning also leads to a uniform distribution of low and high frequency features.

1.5. Applicability to domains other than deepfake detection.

We presented D4 as a framework for adversarially robust deepfake detection. However, we hypothesize that this approach may apply to other classification tasks that exhibit redundancy in a feature space. While we are unaware of such a space for the popular CIFAR10 and ImageNet classification tasks, there are several classification tasks in, say, the audio domain that exhibit redundancy in features, e.g., keyword spotting and fake speech detection. Exploring this hypothesis is an interesting future research direction.

References

[1] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 2

[2] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable ad-

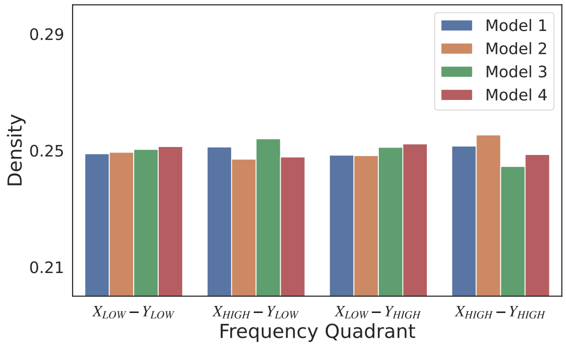

Figure 1. Distribution of partitioned frequencies for D4 (SIZE=4) among low and high frequency regions for each axis of the spectrum.

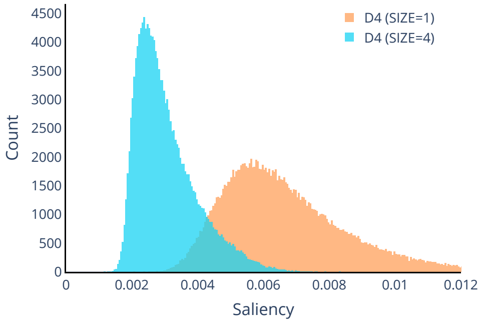

Figure 2. Distribution of frequency feature saliencies for D4 (SIZE=1) and D4 (SIZE=4) ensembles. Lower saliency implies higher feature robustness.

versarial examples. *arXiv preprint arXiv:1704.03453*, 2017. 1