TOWARDS CERTIFIED ADVERSARIAL ROBUSTNESS FOR AUDIO CLASSIFICATION

Patrick O'Reilly

Northwestern University

ABSTRACT

In discriminative tasks such as image classification, deep neural networks have been shown to be vulnerable to *adversarial examples*, artificially-generated perturbations of natural instances that cause a network to make incorrect predictions. *Certified adversarial robustness* methods provide a defense against adversarial examples in the form of mathematical guarantees of the behavior of classifiers under adversarial input. One such method is *randomized smoothing*, which produces a "smoothed" version of an existing classifier whose output can be proven constant within a certain distance of each input. While randomized smoothing has seen success in image-classification tasks, there have been few attempts to apply it to the audio domain. To better understand the viability of this technique in the audio domain, I apply randomized smoothing to a simple speaker-identification task and evaluate the practical quality of the resulting guarantees.

1. INTRODUCTION

In discriminative tasks such as image classification, deep neural networks have been shown to be vulnerable to *adversarial examples* - artificially-generated perturbations of natural instances that cause the network to make incorrect predictions. Audio-domain adversarial examples have recently gained attention due to the dangers they pose to voice-command systems; for example, a malicious actor may seek to attack the automatic transcription models in Apple's Siri or Amazon Alexa, gaining control of a victim's personal device by injecting imperceptible commands as background noise. Similarly, a malicious actor may wish to fool a speaker-detection model within a biometric verification system.

Certified adversarial robustness methods seek to provide mathematical guarantees of the behavior of classifiers under adversarial input. One such method is *randomized smoothing*, which produces a "smoothed" version of an existing classifier whose output can be proven constant within a certain distance of each input. For base classifier f and label set $\mathcal{Y}$, the smoothed classifier g is defined by

$$g(x) = \text{argmax}_{y \in \mathcal{Y}} P_{x' \sim \mu(x)}[f(x')]$$

where $\mu(x)$ is the smoothing distribution for the input x. In [4], the authors show that for $\mu(x) = \mathcal{N}(x, \sigma^2 I)$, the smoothed classifier's output is provably constant within an L_2 radius of $\sigma \cdot \phi^{-1}(\underline{p_A})$ of an input x; here, $\underline{p_A}$ is a lower bound on the probability of the most likely prediction of the base classifier over $\mu(x)$ and ϕ^{-1} is the inverse of the Gaussian CDF. This means that no adversarial attack within the L_2 bound will be successful in changing the smoothed classifier's prediction.

To compute this robust radius for an input x, the authors first generate n_0 perturbations $\{x + \delta_i\}_{i=1\dots n_0}$ by sampling noise from an isotropic Gaussian $\delta_i \sim \mathcal{N}(0, \sigma^2 I)$ and retain the most likely prediction A of the base classifier over the perturbed inputs. A much larger number of samples n is then drawn to estimate $\underline{p_A}$. If $\underline{p_A} \leq 0.5$, the smoothed classifier abstains from producing a radius; otherwise, the smoothed classifier returns the radius $\sigma \cdot \phi^{-1}(\underline{p_A})$.

In the study of adversarial examples, it is common to constrain the size of an attacker's perturbation in terms of an L_p norm: an adversarial example is successful if it forces a classifier to produce the desired prediction by perturbing a natural instance within the norm bound. Common norm constraints include $p = 0$, in which an attacker is allowed to arbitrarily perturb at most a fixed number of coordinates of the input; $p = \infty$, in which the attacker is allowed to perturb any coordinate by a at most a fixed amount; and $p = 2$, in which the attacker must perturb within a fixed Euclidean distance. The randomized smoothing method proposed in [4] provides meaningful guarantees within this threat model of norm-bounded adversaries.

Of course, real-world adversaries are not limited to norm-bounded perturbations of natural instances; for example, an attacker can use a generative model to produce novel high-likelihood instances of a target class for which a classifier will produce incorrect predictions [11]. However, if we restrict our attention to perturbation-based attacks, norm-based robustness guarantees can be of practical use insofar as they increase the likelihood of any successful attack requiring a perceptible perturbation of the input. This perceptual framing of L_p norm bounds dates back to the introduction of adversarial examples, and both L_2 and L_∞ norms have been used as proxy measures of perturbation perceptibility in the image domain [13] [2]. Underlying the use of these measures is the assumption that uniform low-magnitude perturbations of an image are harder for humans to perceive [6].

As a first step towards certifiably robust defenses against audio adversarial attacks, it seems natural to explore (1) what audio-domain guarantees can be obtained using existing image-domain randomized smoothing methods, and (2) the usefulness of these guarantees as a form of ”perceptual lower bound” on adversarial perturbations.

2. AUDIO DOMAIN TASK

I apply the randomized smoothing method of [4] to a speaker-recognition task in which a deep neural network is given a recorded utterance and predicts the identity of the speaker. Much recent work in audio adversarial examples has focused on the comparatively challenging task of automatic speech recognition (ASR), in which a deep neural network produces a transcription of an utterance [10] [3]. However, the speaker-recognition task serves as a more feasible first step and provides a neater analogue of the experiments of [4]. Moreover, it aligns with the focus of recent work in audio-domain certified robustness guarantees [5].

Taking cues from [5], I use an existing third-party implementation of Baidu’s Deep Speaker architecture and the LibriSpeech dataset, which contains 1000 hours of short utterances sampled from public-domain audiobook recordings [1] [9] [7]. Deep Speaker is a neural network model that maps utterances to a unit hypersphere embedding in which speaker similarity is measured by cosine similarity. By excising the final length-normalization layer and appending a linear layer and softmax activation function, the architecture can be made to perform classification. I train models on the *train-clean-360* subset, which contains 360 hours of audio and 921 speakers, as well as a custom 10-speaker subset containing 4 hours of audio. In Table 1, both LibriSpeech subsets and associated models are compared with the datasets and models used in [4].

Deep Speaker accepts normalized Mel-filterbank energies as input rather than “raw” waveform audio, and thus waveforms must be processed through a series of representations before classification. Smoothing could hypothetically be performed at any of these intermediate representations, and it has been shown that the choice of smoothing representation can significantly affect the quality of guarantees. For example, in [2] the authors obtain improved image-domain guarantees by performing smoothing at a “robust” linear projection of the input. To investigate the effects of smoothing at intermediate representations, I perform smoothing experiments at the magnitude spectrogram representation in addition to the waveform and normalized Mel-filterbank energies.

3. EXPERIMENTS

I certify the 10-speaker model on 300 test-set points (30 per speaker) at the magnitude spectrogram and normalized Mel-filterbank energy representations with n_0 = 500, n = 10000, and confidence parameter α = 0.001. Due to time and resource constraints, experiments at the waveform and with the 921-speaker model were not completed by the assignment submission date.

Randomized smoothing guarantees rely on the accuracy of the base classifier over perturbed inputs, and therefore the noise level must be scaled accordingly. Images in the CIFAR-10 and ImageNet datasets have pixel values in the range [0, 1], per-channel means between 0.4 and 0.5, and per-channel standard deviations around 0.2. However, data at the waveform, magnitude spectrogram, and normalized Mel-filterbank energy representations of the LibriSpeech data follow very different distributions. To account for the various scales of the data, I perform smoothing with $\sigma = 0.01, 0.05, 0.10$ in addition to the values of $0.25, 0.50$ used in [4]; initial experiments with a noise level of 1.00 produced poor guarantees, and further experiments were put on hold due to time constraints.

The certified accuracy of a smoothed classifier at a given radius is the proportion of test inputs for which the smoothed classifier produces a correct prediction and a guarantee of at least the specified radius. Tables 2 and 3 show the certified accuracy of the 10-speaker model for various noise levels and radii at the magnitude spectrogram and normalized Mel-filterbank energy representations, respectively. Interestingly, while Cohen et al. found it necessary to train base classifiers on Gaussian-augmented data matched to the smoothing noise level at which certification was performed, Deep Speaker models trained on augmented data produced worse guarantees across the board; this may indicate a flaw in the data augmentation process.

4. RESULTS

As waveform experiments are still underway, only a sketch of a final analysis is presented here. For a fixed L_2 norm bound, one can imagine two simplified adversaries: one which produces uniform low-magnitude perturbations of the input, and one which produces sparse or localized high-magnitude perturbations. On the CIFAR-10 dataset, the image-domain radius guarantees obtained in [4] appear sufficient to force the second hypothetical adversary’s perturbations into the realm of perceptibility, and within an order of magnitude of the radius required to expose the first. In the audio domain, the experiments conducted so far suggest that waveform guarantees will be insufficient to ensure the perceptibility of perturbations by either adversary, and will fall shorter in the case of the first adversary. This is partially due to the fact that local high-magnitude attacks can be concealed in the audio domain through the use of perceptual masking [10]. However, an (as yet incomplete) analysis of the manifestation of low-magnitude uniform-bounded and local high-magnitude perturbations at the magnitude spectrogram and normalized Mel-filterbank energies may show that the smoothing guaran-

Table 1. Comparison of Datasets and Associated Models

Dataset	Classes	Dimension	Examples	Architecture	Parameters	Accuracy
LibriSpeech 10-Speaker Subset	10	25000 (waveform), 41120 (spectrogram), 10240 (filterbank)	30k / 10k	Deep Speaker ResCNN	24m	0.99
CIFAR-10	10	3072	50k / 10k	ResNet-110	1.7m	0.90
LibriSpeech *train-clean-360* subset	921	25000 (waveform), 41120 (spectrogram), 10240 (filterbank)	550k / 92.1k	Deep Speaker ResCNN	24m	0.96
ImageNet	1000	196608	1.2m / 50k	ResNet-50	25m	0.75

Table 2. Certified Accuracy, Magnitude Spectrogram

	$r = 0.01$	$r = 0.05$	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1.0$	$r = 1.25$	$r = 1.5$
$\sigma = 0.01$	**0.93**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\sigma = 0.05$	0.86	**0.79**	0.00	0.00	0.00	0.00	0.00	0.00
$\sigma = 0.10$	0.73	0.68	**0.39**	0.00	0.00	0.00	0.00	0.00
$\sigma = 0.25$	0.36	0.35	0.25	**0.11**	0.04	0.00	0.00	0.00
$\sigma = 0.50$	0.13	0.13	0.12	0.11	**0.09**	**0.08**	**0.06**	**0.04**

Table 3. Certified Accuracy, Normalized Mel-Filterbank Energies

	$r = 0.01$	$r = 0.05$	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1.0$	$r = 1.25$	$r = 1.5$
$\sigma = 0.01$	**0.90**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\sigma = 0.05$	0.71	**0.70**	0.00	0.00	0.00	0.00	0.00	0.00
$\sigma = 0.10$	0.52	0.51	**0.47**	0.00	0.00	0.00	0.00	0.00
$\sigma = 0.25$	0.20	0.19	0.18	**0.15**	**0.11**	0.00	0.00	0.00
$\sigma = 0.50$	0.10	0.10	0.10	0.10	0.10	**0.09**	**0.08**	**0.05**

tees obtained at these representations are of practical value.

5. FUTURE WORK

The above experiments constitute only a small step towards a full exploration of the viability of randomized smoothing as an audio defense, and suggest a number of follow-up tasks. These include: completing the analysis of the quality of smoothing guarantees obtained at the waveform, magnitude spectrum, and normalized Mel-filterbank energy representations; empirical evaluation of the robustness of classifiers smoothed at the waveform, magnitude spectrogram, and normalized Mel-filterbank energy representations against simple and perceptually-motivated audio adversarial attacks; comparisons against domain-specific heuristic defense baselines, such as those proposed in [12]; and the exploration of whether useful guarantees can be obtained from perceptually-motivated randomized smoothing approaches, such as the technique proposed in [8].

References

[1] Deep speaker: An end-to-end neural speaker embedding system. `https://github.com/philipperemy/deep-speaker`.

[2] Pranjal Awasthi, Himanshu Jain, Ankit Singh Rawat, and Aravindan Vijayaraghavan. Adversarial robustness via robust low rank representations. In *NeurIPS*, 2020.

[3] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE Security and Privacy Workshops*, 2018.

[4] Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.

[5] Krishnamurthy Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Sven Xiao, Kai andGowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *ICLR*, 2020.

[6] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*, 2019.

[7] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. In *arXiv*, 2017.

[8] Ethan Mendes and Kyle Hogan. Defending against imperceptible audio adversarial examples using proportional additive gaussian noise. 2020.

[9] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpu. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 2015.

[10] Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *ICML*, 2019.

[11] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *NeurIPS*, 2018.

[12] Vinod Subramanian, Emmanouil Benetos, Ning Xu, SKoT McDonald, and Mark Sandler. Robustness of adversarial attacks in sound event classification. In *arXiv*, 2019.

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.