ADVERSARIALLY TRAINING FOR AUDIO CLASSIFIERS

Raymel Alfonso Sallo (M.Sc Student), Mohammad Esmaeilpour, Prof. Patrick Cardinal

École de Technologie Supérieure (ÉTS), Montréal, Québec, Canada

raymel.alfonso-sallo.1@ens.etsmtl.ca

Paper ID: 2639

Problem Statement

- Investigating the effect of adversarially training as a gradient obfuscation-free defense approach

Contributions

- Characterizing the adversarially training impact on six advanced deep neural network architectures for diverse audio representations
- Demonstrating that deep neural networks specially those with residual blocks have higher recognition performance on tonnetz features concatenated with DWT spectrograms compared to STFT representations
- Showing the adversarially trained AlexNet model outperforms ResNets with limiting the perturbation magnitude
- Experimentally proving that although adversarially training reduces recognition accuracy of the victim model, it makes the attack more costly for the adversary in terms of required perturbation.

Taxonomy of the Attacks

Attack	Adversary Knowledge	Type of misclassification
FGSM [1]	Whitebox	Targeted
BIM [2]	Whitebox	Targeted
JSMA [3]	Whitebox	Targeted
DeepFool [4]	Whitebox	Untargeted
PIA [5]	Blackbox	Targeted
CWA [6]	Whitebox	Targeted

Fast Gradient Sign Method (FGSM)

- Successful adversarial examples can be crafted due to limitation in precision of input features
- Analytical perturbations can be crafted by following the direction of the gradient of the cost function used to train the model

$$\eta = \epsilon sign(\nabla_x J(\theta, x, y))$$

Basic Iterative Method (BIM)

- It just iterates the FGSM algorithm using a small step size
- Intermediate features values are clipped to assure that features remain in the ϵ-neighborhood of the original input sample

$$X_{N+1}^{adv} = clip_{X,\epsilon}\left\{X_N^{adv} + \alpha sign\left(\nabla_X J(X_N^{adv}, y_{true})\right)\right\}$$

Jacobian Saliency Map Attack (JSMA)

- Construct an *adversarial saliency map* ***S*** by evaluating the forward derivative by means of the Jacobian matrix of the function learned by the classifier
- A set conditions are applied to the saliency map to narrow the search direction for crafting successful perturbations in the input space leading to wrong classification

$$\nabla \mathbf{F}(\mathbf{X}) = \frac{\partial \mathbf{F}(\mathbf{X})}{\partial \mathbf{X}} = \left[\frac{\partial \mathbf{F}_j(\mathbf{X})}{\partial \mathbf{X}_i} \right]_{i \in 1 \ldots M, j \in 1 \ldots N}$$

$$S(\mathbf{X}, t)[i] = \begin{cases} 0 & \text{if } \frac{\partial \mathbf{F}_t(\mathbf{X})}{\partial \mathbf{X}_i} < 0 \text{ or } \sum_{j \neq t} \frac{\partial \mathbf{F}_j(\mathbf{X})}{\partial \mathbf{X}_i} > 0 \\ \left(\frac{\partial \mathbf{F}(X)}{\partial \mathbf{X}_i} \right) \left| \sum_{j \neq t} \frac{\partial \mathbf{F}(\mathbf{X})}{\partial \mathbf{X}_i} \right| & \text{otherwise} \end{cases}$$

Carlini Wagner Attack (CWA)

- Assumes that not all features need to be perturbed during the attack without shattering the gradient information
- The algorithm generalize well its adversarial goal on three known distance metrics L_0, L_2 and L_∞
- Finding the constant ***c*** is done by binary search and it is a difficult hyperparameter to tune

$$\min \|\mathbf{x}' - \mathbf{x}\|_2^2 + c \cdot \mathcal{L}(\mathbf{x}')$$

$$\min_{\rho} \left\| \frac{1}{2}(\tanh(\rho) + 1) - \mathbf{x} \right\|_2^2 + c \cdot \mathcal{L}\left(\frac{1}{2}\tanh(\rho) + 1\right)$$

$$\delta^* = \frac{1}{2}(\tanh(\rho + 1)) - \mathbf{x}$$

Audio Representations

- The generation of the 2D representations is done by STFT and DWT with and without Tonnetz features
- In the case of STFT a discrete signal **a[n]** is combined over time with a Hann function and the Fourier transformation is computed as follows:

$$\mathrm{STFT}\left\{a[n]\right\}(m,\omega) = \sum_{n=-\infty}^{\infty} a[n]H[n-m]e^{-j\omega n}$$

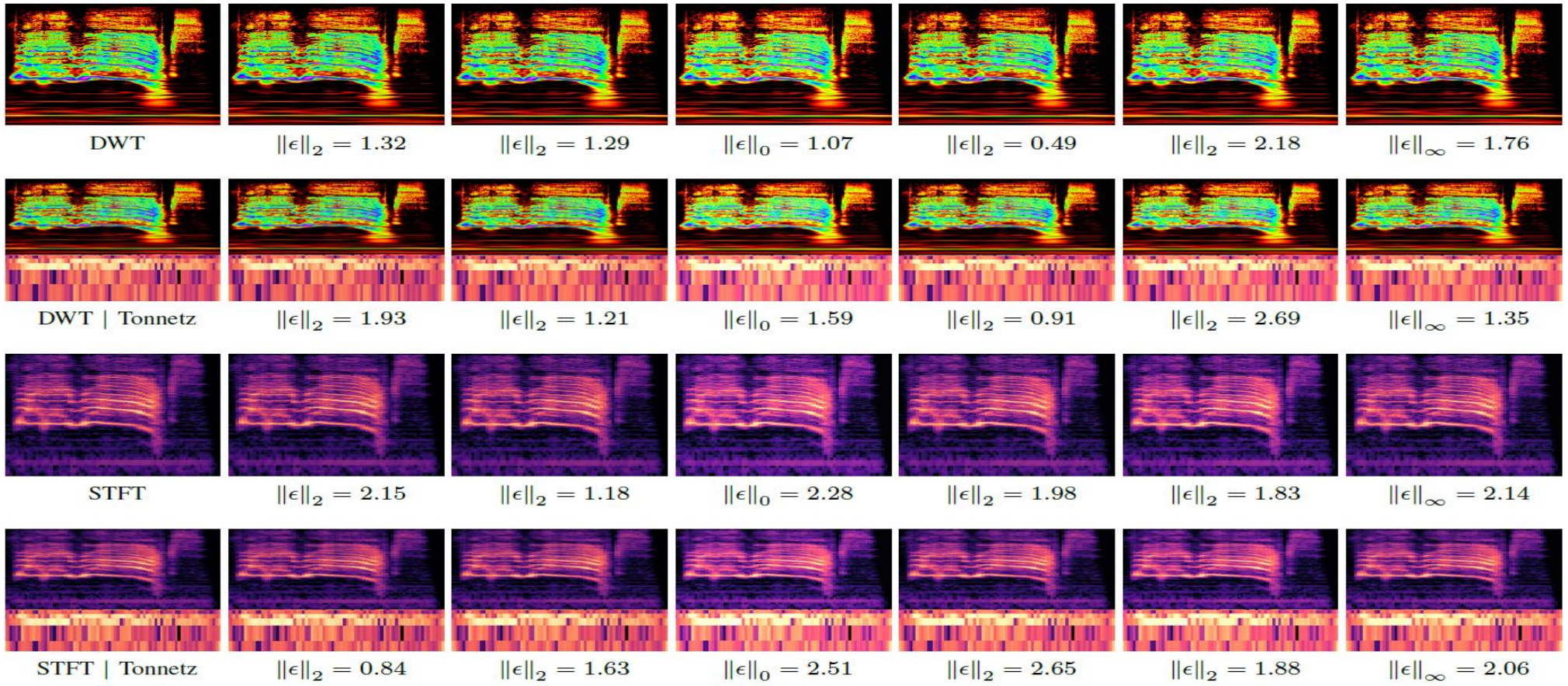
- In the case DWT a complex Morlet wavelet was used because of its nonlinear characteristics

$$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-j\omega t} e^{-t^2/2}$$

- Once the basis function is selected the Discrete Wavelet Transform is

$$\mathrm{DWT}\left\{a(t)\right\} = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} a(t)\psi\left(\tfrac{t-\tau}{s}\right) dt$$

Audio Representations



Adversarially Training

- Can be considered a sort of active learning, where the model plays the game of trying to minimize worst case error against corrupted data
- To include the adversarial component, the objective function must be modified in order to reflect the nature of the new type of crafted perturbations

$$J'(\mathbf{x}, l, \mathbf{w}) = \alpha J(\mathbf{x}, l, \mathbf{w}) + (1 - \alpha) J(\mathbf{x}', l, \mathbf{w})$$

- Use of one-shot FGSM adversarial examples to avoid shattered gradients
- This adversarial training setup runs as a fast non-iterative procedure

$$\tilde{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha) J(\theta, x + \epsilon sign(\nabla_x J(\theta, x, y)))$$

Adversarial Attack Setup

- We bind the fooling rate of all attacks algorithms to a threshold of *AUC > 0.9* associated to the area under the curve of the attack success
- Fine-tuning hyperparameters of the different attacks to meet the previous baseline performance

Dataset

- UrbanSound8K with 8732 short recording for 10 classes and ESC-50 containing 2K audio signals of equal length of 5s organized in 50 classes
- Preprocessing of samples by doing pitch-shifting operation using 1D filtration
- Resulting spectrograms of $1568\ x\ 768$ for both STFT and DWT representations, used standalone or in combination with $1568\ x\ 540$ chromagrams

Recognition Accuracy with and without Adversarially Training

Dataset	Representations	GoogLeNet	AlexNet	ResNet-18	ResNet-34	ResNet-56	VGG-16
ESC-50	STFT	67.83, (06.89)	64.32, (10.91)	66.85, (12.13)	67.21, (14.43)	**69.77**, (09.29)	68.94, (08.32)
	DWT	70.42, (08.42)	65.39, (11.23)	67.06, (15.71)	67.55, (18.76)	**71.56**, (11.09)	71.43, (16.28)
	STFT | Tonnetz	70.11, (24.09)	64.21, (23.76)	67.62, (19.48)	66.75, (23.31)	**70.22**, (25.19)	70.18, (23.68)
	DWT | Tonnetz	68.76, (19.07)	68.31, (18.53)	68.49, (24.27)	67.15, (21.56)	**71.79**, (18.21)	68.37, (18.73)
UrbanSound8K	STFT	88.32, (10.35)	86.07, (21.43)	88.24, (14.94)	88.61, (09.19)	**88.77**, (23.06)	87.93, (14.66)
	DWT	90.10, (16.35)	87.51, (19.59)	88.07, (15.08)	88.38, (19.04)	**90.14**, (15.49)	90.11, (16.35)
	STFT | Tonnetz	88.44, (25.77)	86.81, (22.05)	88.13, (17.64)	88.38, (26.42)	89.41, (20.73)	**89.42**, (21.38)
	DWT | Tonnetz	89.32, (16.83)	87.34, (20.41)	88.76, (29.12)	89.80, (27.45)	**91.36**, (26.08)	89.97, (24.56)

Robustness of Adversarially Trained Models

Dataset	Representations	GoogLeNet	AlexNet	ResNet-18	ResNet-34	ResNet-56	VGG-16
ESC-50	STFT	53.12	**50.97**	61.13	65.31	65.87	63.05
	DWT	55.68	**51.03**	62.56	64.18	67.26	64.23
	STFT | Tonnetz	56.18	**50.46**	53.10	68.29	68.19	64.82
	DWT | Tonnetz	55.74	**49.33**	58.87	69.77	70.42	66.37
UrbanSound8K	STFT	56.09	**53.24**	62.06	65.91	74.30	75.35
	DWT	58.98	**51.92**	63.59	63.40	73.86	74.66
	STFT | Tonnetz	65.80	**50.71**	62.75	64.02	75.11	73.39
	DWT | Tonnetz	68.46	**52.23**	60.13	67.81	76.38	75.26

Average Perturbation Ration for Legitimate and Adversarially Trained Examples

Dataset	Representations	GoogLeNet	AlexNet	ResNet-18	ResNet-34	ResNet-56	VGG-16
ESC-50	STFT	1.412	1.231	1.897	2.154	**2.312**	2.107
	DWT	1.562	1.309	1.741	1.982	1.976	**2.307**
	STFT | Tonnetz	1.804	1.918	2.003	**2.161**	2.095	1.674
	DWT | Tonnetz	2.014	2.336	1.788	1.903	**2.609**	2.230
UrbanSound8K	STFT	1.562	1.903	**2.439**	1.372	1.991	1.703
	DWT	2.154	2.287	2.764	1.644	**2.892**	1.789
	STFT | Tonnetz	2.231	2.108	1.981	2.003	1.401	**2.308**
	DWT | Tonnetz	1.606	2.199	2.405	1.604	**2.501**	1.602

$$\epsilon_r = \left| \frac{\epsilon_a}{\epsilon_o} \right|$$

ÉTS
Le génie pour l'industrie

ICPR 2020

Conclusions

- We trained six advanced deep learning classifiers on four different 2D representations of environmental audio signals
- We run five white-box and one black-box attack algorithms against these victim models
- We demonstrated that adversarially training considerably reduces the recognition accuracy of the classifier but improves the robustness against six types of targeted and non-targeted adversarial examples
- We demonstrated that adversarially training is not a remedy for the threat of adversarial attacks, however it escalates the cost of attack for the adversary with demanding larger adversarial perturbations compared to the non-adversarially trained models

References

[1] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.
[2] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” arXiv preprint arXiv:1607.02533, 2016.
[3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, andA. Swami, “The limitations of deep learning in adversarial settings,” in 2016 IEEE European Symposium on Security and Privacy (EuroS&P).
IEEE, 2016, pp. 372–387.
[4] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in IEEE Conf Comp Vis Patt Recog, 2016, pp. 2574–2582.
[5] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” arXiv preprint arXiv:1804.08598, 2018.
[6] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in IEEE Symp Secur Priv, 2017, pp. 39–57.