INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Adversarial Separation Network for Speaker Recognition

Hanyi Zhang[1,2], *Longbiao Wang*[2,*], *Yunchun Zhang*[1,3,*], *Meng Liu*[2], *Kong Aik Lee*[4], *Jianguo Wei*[2]

[1]School of Software, Yunnan University, Yunnan, China
[2]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]Engineering Research Center of Cross-border Network Security, Ministry of Education, China
[4]Biometrics Research Laboratories, NEC Corporation, Japan

hanyizhang@mail.ynu.edu.cn, longbiao_wang@tju.edu.cn, yczhang@ynu.edu.cn

Abstract

Deep neural networks (DNN) have achieved great success in speaker recognition systems. However, it is observed that DNN based systems are easily deceived by adversarial examples leading to wrong predictions. Adversarial examples, which are generated by adding purposeful perturbations on natural examples, pose a serious security threat. In this study, we propose the adversarial separation network (*AS-Net*) to protect the speaker recognition system against adversarial attacks. Our proposed *AS-Net* is featured by its ability to separate adversarial perturbation from the test speech to restore the natural clean speech. As a standalone component, each input speech is pre-processed by *AS-Net* first. Furthermore, we incorporate the compression structure and the speaker quality loss to enhance the capacity of the *AS-Net*. Experimental results on the VCTK dataset demonstrated that the *AS-Net* effectively enhanced the robustness of speaker recognition systems against adversarial examples. It also significantly outperformed other state-of-the-art adversarial-detection mechanisms, including adversarial perturbation elimination network (APE-GAN), feature squeezing, and adversarial training.

Index Terms: speaker recognition, deep neural network, adversarial example, separation network

1. Introduction

The goal of speaker recognition is to determine the identity of a person through speech. Both the safety and robustness of speaker recognition systems have attracted much attention. It has been shown that speaker recognition models are vulnerable to many attacks [1, 2, 3], such as replay, speech synthesis, etc. Besides, with the widespread use of deep neural networks (DNN) in speaker recognition tasks, the robustness of DNNs also dramatically affects the security of the speaker recognition systems. Recent studies [4, 5] have shown that by adding imperceptible perturbations to the inputs can cause DNNs to produce incorrect results. We refer to the examples added with perturbations as the *adversarial examples*. Some well-designed adversarial algorithms were presented to search for subtle but effective adversarial perturbations when input with targeting speech.

Adversarial attacks have been studied previously in speech processing systems. In [6, 7], the existing speaker verification networks were found to be vulnerable against adversarial examples. Carlini et al. [8] verified the effectiveness of adversarial attacks on automatic speech recognition. It is widely acknowledged that adversarial examples have posed a serious security threat to speech processing systems.

*Corresponding author

To secure the deep learning models from adversarial attacks, many countermeasures have been introduced. Defensive distillation [9] worked by converting class labels into soft targets to mask the gradient information from attackers. Lu et al. [10] trained DNN-based binary classifiers to detect whether the inputs are adversarial or bona fide. However, only a few solutions have been proposed to improve the robustness of speaker recognition systems against adversarial examples. Moreover, as mentioned in [11, 12], the adaptive capacity of the countermeasure is limited by the transferability of the targeted adversarial algorithm. Therefore, it is a challenging problem to maintain the existing speaker recognition system with satisfactory performance against different adversarial attacks.

To alleviate the adversarial risks of speaker recognition systems and mitigate the drawbacks of existing solutions, this paper proposes a defense mechanism referred to as the *adversarial separation network (AS-Net)*. The key principle here is to separate the adversarial perturbations from the adversarial examples so as to restore the natural examples. Specifically, an independent filtering module is designed and applied before forwarding the input examples to the speaker recognition system. In this filtering module, the restored speeches are preserved, while adversarial perturbations are removed as noise. Then, the restored speeches are directly fed into the recognition system to finally output the recognition results. As the core functional module, *AS-Net* is designed to be responsible for filtering tasks. As part of the *AS-Net*, two optimized components, including compression structure and speaker quality loss, are introduced. The compression structure helps to facilitate adversarial perturbation reconstruction. The speaker quality loss is responsible for supervising whether the restored speeches generated by *AS-Net* are correctly labeled by the recognition system. The experimental results on the VCTK [13] corpus show that *AS-Net* significantly improves the robustness of the speaker recognition system against various adversarial attacks. It also outperforms other state-of-the-art strategies. Besides, *AS-Net* has no direct impact on the original speaker recognition system. Therefore, *AS-Net* is applicable and can be easily integrated with other countermeasures.

2. Baseline speaker recognition system

The d-vector [14] based speaker recognition system is used as our baseline model. First, the DNN is trained by using speaker IDs as classification labels. Then, in the enrollment phase, the enrolled speeches are input to the trained DNN, and the embeddings of the last layer are extracted as the speakers' identity vectors. If a speaker has multiple enrolled speeches, all embeddings are averaged. In particular, the identity vector c_k of

http://dx.doi.org/10.21437/Interspeech.2020-1966

speaker k is defined as

$$c_k = \frac{1}{M}\sum_{m=1}^{M} e_{k,m} \tag{1}$$

where $e_{k,m}$ represents the embedding of the m-th enrolled speech of speaker k. Afterwards, the cosine distance between the test speech's embedding and the identity vector is calculated in the testing phase. The final decision is made by comparing the cosine distance with the predefined threshold.

Both SE-Resnet and SE-Resnext [15] are applied as DNN models to evaluate the effectiveness of our proposed *AS-Net* on different speaker recognition systems. SE-Resnet and SE-Resnext are derived from Resnet [16] and Resnext [17] with Squeeze-Excitation network (SENet) [15], respectively. By introducing learning concepts, SENet selectively emphasizes important parts in features and ignoring useless parts. It is observed that SENet exhibit a certain adversarial regularization [7] effect that is similar to attention [18], and thus effective in resisting adversarial examples.

3. Generating adversarial examples

Adversarial examples are created by adding perturbations on natural examples to fool the target model. Therefore, the ideal adversarial attacks can mislead the system effectively while perturbations are imperceptible to humans. Assuming a natural example x and its correct label y_{true}, while $P(y|x)$ represents the output label of x from the speaker recognition system. After adding perturbation γ to the natural example x, an adversarial example $\hat{x}$ is generated ($\hat{x} = x + \gamma$). The adversarial perturbation is constructed by solving the following equation.

$$\min_{\gamma} ||\gamma||_2 \; s.t. \; P(y|x+\gamma) \neq y_{true} \tag{2}$$

To address the above optimization problem defined in Eq. (2), several effective adversarial algorithms are designed.

3.1. Fast gradient sign method (FGSM)

FGSM [19] generates adversarial perturbation by calculating the gradient that maximizes the loss. The perturbations computed in FGSM are as shown in Eq. (3).

$$\gamma = \alpha \cdot sign(\nabla_x J(P(y|x), y_{true})) \tag{3}$$

where $J(P(y|x), y_{true})$ is the loss of neural network and α is the intensity of the perturbation.

3.2. Projected gradient descent (PGD)

PGD [11] also relies on the gradient to calculate perturbations. Unlike FGSM [19], PGD [11] uses a multi-step solution. Through several iterations, finer perturbations are generated. PGD [11] can be expressed as

$$\gamma^{t+1} = \gamma^{t} + \alpha \cdot sign(\nabla_x J(P(y_t|x_t), y_{true})) \tag{4}$$

where γ^t represents the perturbation at t-th iteration.

3.3. Adversarial attacks with momentum (MT)

Momentum attack (MT) [20] uses the momentum gradient to iteratively change the perturbation. This method accumulates velocity vectors along the gradient direction of the loss function to stabilize the update direction. By replacing the current gradient with the cumulative gradient, MT [20] can be easily generalized to other algorithms.

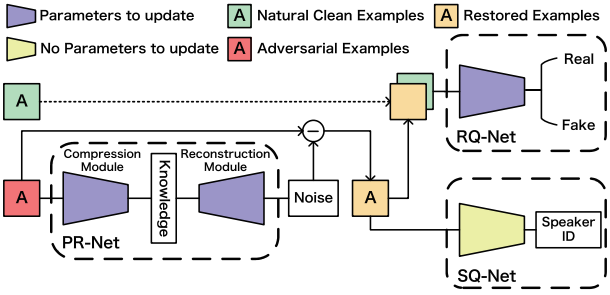

Figure 1: *The architecture of the proposed adversarial separation network (AS-Net).*

3.4. Decoupled direction and norm (DDN)

DDN [21] that decouples the direction and norm of the adversarial perturbation can achieve satisfactory misclassification with only a few changes. Instead of punishing adversary with low norm values, this method optimizes the cross-entropy loss.

4. Adversarial separation network

The key to a successful attack is the perturbation added to the natural example. If there is a mechanism to separate perturbations from natural examples, the threat posed by adversarial examples would be largely alleviated. Based on the above idea, we propose a separation mechanism. However, in practice, two vulnerabilities are observed. One is that information redundancy is common in the natural speech data, and the other one is that the restored data are difficult to be correctly discriminated by the speaker recognition system. These two vulnerabilities undoubtedly affect the embeddings of test speeches. To this end, we design the compression structure and the speaker quality loss.

Finally, we propose the adversarial separation network (*AS-Net*) that can reconstruct adversarial perturbations from adversarial examples and filter them off. Thereby, *AS-Net* can convert adversarial speeches into natural speeches. As shown in Fig. 1, the proposed *AS-Net* consists of three parts, namely, a perturbation reconstruction network (PR-Net), a reality quality network (RQ-Net), and a speaker quality network (SQ-Net).

4.1. Algorithm overview

To recover the perturbed adversarial speech into the natural clean speech, the compression module of PR-Net first extracts the relevant knowledge K about the adversarial perturbation from an adversarial speech A ($K = f_{Com}(A)$). Then, PR-Net reconstructs the perturbation $\hat{N}$ based on the knowledge K ($\hat{N} = f_{\mathrm{Re}c}(K)$). Afterwards, the result of the adversarial speech A minus the reconstructed perturbation $\hat{N}$ is the restored natural speech $\hat{A}$, as defined in Eq. (5).

$$\hat{A} = A - f_{\mathrm{Re}c}(f_{Com}(A)) \tag{5}$$

The restored speech should not only comply with the same data distribution as the real clean speeches but should also be correctly recognized by the target recognition models. The above two tasks are accomplished with the RQ-Net and the SQ-Net, respectively.

4.2. Architecture of AS-Net

The PR-Net is designed as the compression structure to generate the adversarial perturbations for the input examples. The

compression structure consists of the compression and the reconstruction module. The compression module is responsible for ensuring that the extracted knowledge efficiently retains relevant information on perturbations. The reconstruction module aims to improve the quality of reconstructed perturbations.

The RQ-Net encourages the restored examples, which are perturbations free, to be similar to real clean examples. For supervision, RQ-Net is capable of distinguishing between generated examples and real clean examples.

The SQ-Net is the DNN to implement speaker recognition task. During the training phase of *AS-Net*, SQ-Net supervises the *AS-Net* to generate the restored examples that can be recognized by DNN as the correct speaker IDs. Thus, in the testing phase, the DNN can generate correct embeddings for the input examples which are processed by *AS-Net*. The parameters of SQ-Net are fixed during the training process of *AS-Net*.

4.3. Loss functions

Generally speaking, *AS-Net* can perfectly separate perturbations from adversarial examples and restore clean examples. To achieve this goal, some novel loss functions are introduced to constrain the parameter learning of the PR-Net. Three loss functions are designed in this paper, including perturbation loss, reality loss, and speaker quality loss.

Perturbation loss: The perturbation loss is defined as

$$l_{pertur} = \frac{1}{TF}\sum_{n=1}^{N}\sum_{i=1}^{T}\sum_{j=1}^{F}\left(Pertur_{i,j} - f_{PR}(A^{adv})_{i,j}\right)^2 \tag{6}$$

where N is the batch size, T is the number of frames, F is the dimension of each frame and $Pertur$ is the real perturbation. The perturbation loss drives the reconstructed perturbations that infinitely close to the real perturbations. As shown in Eq. (6), we use the l_2 norm to calculate the dissimilarity between the reconstructed perturbations and the real perturbations.

Reality loss: Since the restored examples should be consistent enough with the real clean examples, they should also comply with the same data distribution as the real clean examples. Reality loss encourages the generated examples to be regarded by RQ-Net as real clean examples, thereby improving the quality of reconstructed perturbations. The reality loss function is defined in Eq. (7).

$$l_{reality} = \sum_{n=1}^{N}\left[1 - \log f_{RQ}(A^{adv} - f_{PR}(A^{adv}))\right] \tag{7}$$

Speaker quality loss: The *AS-Net* is distinguished by its ability to produce the correct embeddings when input its restored examples into the recognition system's DNN. To this end, *AS-Net* should learn to generate examples that can make DNN output the correct speaker IDs during the training stage. Based on this analysis, we introduce the recognition system's DNN as SQ-Net and guide PR-Net's parameter learning through speaker quality loss. The effect of the speaker quality loss can be embodied as vectors pointing to the correct categories, which can greatly improve the embeddings' accuracy. The speaker quality loss function is defined in Eq. (8).

$$l_{\text{speaker}} = -\sum_{n=1}^{N} ID \cdot \log(f_{SQ}(A^{adv} - f_{PR}(A^{adv}))) \tag{8}$$

where ID is the correct speaker ID in one-hot form.

All in all, the loss function of PR-Net is the weighted average of the above loss functions.

5. Experiments and Analyses

5.1. Experimental setup

To evaluate the feasibility of our *AS-Net*, all models were tested on the VCTK corpus [13]. The VCTK corpus contained speech recordings of speakers with different ages and genders. Each audio was recorded in laboratory quality with 96 kHz sampling.

Two baseline models, including SE-Resnet and SE-Resnext [15] with 512-dimensional and 2048-dimensional embeddings, respectively, were used for performance comparison. These models were constructed based on the d-vector [14], while constant-Q transform (CQT) [22] was applied on extracting features. Similarities among the given examples were computed by the cosine distance. For all recognition systems, the equal error rate (EER) was applied as the evaluation indicator.

All adversarial attack algorithms, including FGSM, PGD, DDN, and MT, were implemented based on AdverTorch [23] library. Both the original clean examples and the adversarial examples were combined to train *AS-Net*. Meanwhile, MT simulated unknown adversarial algorithms that existed in real applications, and was only introduced during the testing stage.

5.2. Adversarial example experiment

In Table 1, the EERs of SE-Resnet and SE-Resnext under different settings were evaluated. It is observed that both systems achieve low EERs in recognizing natural examples while no adversarial attacks exist (Normal). However, EERs of those systems significantly increase when attacked by adversarial examples. This indicated that speaker recognition systems were vulnerable when attacked by adversarial examples. Therefore, the effective defense mechanisms were necessary to secure the speaker recognition systems against adversarial attacks.

Table 1: *Results of speaker recognition system for the normal examples and the adversarial examples.*

EER (%)	Normal	FGSM	PGD	DDN	MT
SE-Resnet	0.89	13.81	16.66	24.75	12.96
SE-Resnext	1.43	13.67	13.72	26.65	12.24

5.3. Comparative experiments

To quantitatively measure the performance of *AS-Net*, we also employed other countermeasures, including APE-GAN [24], feature-squeezing [25] and adversarial training (Adv-Train) [11, 26]. APE-GAN [24] used generative adversarial network (GAN) [27] to eliminate adversarial perturbations. Feature-squeezing [25] determined whether the input was an adversarial example by comparing the feature-squeezing example's prediction with the original example's prediction. Adversarial training [11, 26] combined adversarial examples to retrain the model with the ability to correctly classify adversarial examples.

Tables 2 and 3 compared the performance of different countermeasures. We applied *AS-Net* to the speaker recognition system (Ours). When compared with the system without adversarial countermeasures (None), *AS-Net* achieves relative error reduction rate of 78.8% under different adversarial attacks. The experimental results proved that *AS-Net* could effectively enhance the safety and robustness of speaker recognition systems against adversarial examples. Besides, our approach significantly reduced the EER against MT adversarial examples. This

Table 2: *Comparison of different adversarial defense countermeasures in **SE-Resnet** based speaker recognition system.*

EER (%)	FGSM	PGD	DDN	MT
None	13.81	16.66	24.75	12.96
Adv-Train	6.05	4.48	17.76	4.43
Feature-Squeezing	11.81	13.32	21.39	11.16
APE-GAN	13.51	15.10	15.64	11.53
AS-Net (ours)	**3.62**	**1.94**	**4.51**	**2.68**

Table 3: *Comparison of different adversarial defense countermeasures in **SE-Resnext** based speaker recognition system.*

EER (%)	FGSM	PGD	DDN	MT
None	13.67	13.72	26.65	12.24
Adv-Train	**3.69**	3.05	12.52	2.88
Feature-Squeezing	10.88	10.13	20.38	8.90
APE-GAN	14.10	13.83	13.32	11.45
AS-Net (ours)	3.72	**2.65**	**6.75**	**2.59**

indicated that *AS-Net* was also effective for unknown adversarial attacks that may appear in real-world application scenarios.

According to the results of Tables 2 and 3, the *AS-Net* proposed was superior to similar defense algorithms in performance. Specifically, *AS-Net* achieves relative error reduction rate of 73.8% compared with APE-GAN. The main reason is that *AS-Net* adds a supervision mechanism for recognition results, which greatly improves the quality of the embeddings generated by DNN. Besides, our proposed approach outperforms the Feature-Squeezing by 73.6% relative error reduction. The reason is that the effect of block-structured smoothing in Feature-Squeezing decreases when processing time-structured speech data. Moreover, when recognizing FGSM, PGD and MT adversarial examples, *AS-Net* has 30.0% relative error reduction compared to Adv-Train. However, when recognizing DDN adversarial examples, the relative error reduces by 62.8%. The reason is that the robustness of Adv-Train is greatly dependent on the transferability of the adversarial examples in the training set. This causes Adv-Train's performance to oscillate when dealing with different types of adversarial examples.

The performances of adversarial examples with different perturbation intensities were also evaluated, as shown in Fig. 2. While perturbation intensity increases, it is observed that the EER of *AS-Net* is significantly lower than other methods. Meanwhile, the EER of *AS-Net* shows no dramatic fluctuation within a certain range. And this means that *AS-Net* achieves the best performance under different perturbation intensities.

5.4. Ablation study

For a systematic analysis, we also carried out ablation experiments on compression structure and speaker quality loss.

We compared *AS-Net* with the other two mechanisms. One is the case that no compression structure (W/o Compression) is applied where the speaker quality loss is preserved while replacing the compression structure with the transformation network that has no data dimension change. The other one is the absence of speaker quality loss (W/o Speaker Loss) where only compression structure is applied. Our method sets the weight of the speaker quality loss to 0 during the training to keep *AS-Net*'s parameters free from being affected by the loss updating.

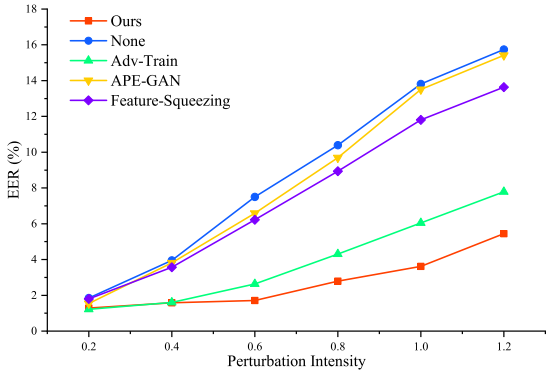

Figure 2: *Results of speaker recognition system (**SE-Resnet**) with different adversarial countermeasures against FGSM with different perturbation intensities.*

Table 4: *Results of speaker recognition system (**SE-Resnet**) in ablation study.*

EER (%)	FGSM	PGD	DDN	MT
W/o Compression	12.99	15.46	18.35	12.02
W/o Speaker Loss	12.42	14.43	10.34	11.12
AS-Net (ours)	**3.62**	**1.94**	**4.51**	**2.68**

Table 4 shows the results of ablation study. It is observed that the relative error improves by 361.3% after removing the compression structure (W/o Compression). The model with no compression structure needs to perform transformations in the entire data domain, which makes model learning to be difficult. Besides, the results also show that after removing the speaker quality loss (W/o Speaker Loss), the recognition system's EER under different adversarial examples also increases significantly. This is because the mechanism for supervising *AS-Net* to generate examples that can be correctly embedded by the DNN is missing. The results of ablation study indicated that the two mechanisms could effectively improve the defensive effect of *AS-Net* on speaker recognition systems.

6. Conclusions

In this study, we proposed a novel countermeasure to enhance the safety and robustness of speaker recognition systems against adversarial examples. In particular, we designed an adversarial separation network (*AS-Net*) to eliminate adversarial perturbations and restore natural clean speeches. The test speech must be processed by *AS-Net* before input it to the system for recognition. To further improve the performance, we also optimized the *AS-Net* by introducing compression structure and speaker quality loss. We compared *AS-Net* with other state-of-the-art countermeasures on the public speaker dataset. The results indicated that *AS-Net* could effectively enhance the safety of speaker recognition systems in different adversarial situations, and significantly outperformed other countermeasures.

7. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

8. References

[1] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks." in *Interspeech*, 2017, pp. 82–86.

[2] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge." *Proc. Interspeech 2019*, pp. 1033–1037, 2019.

[3] F. Alegre, R. Vipperla, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012, pp. 1688–1691.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[5] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[6] S. Liu, H. Wu, H.-y. Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," *arXiv preprint arXiv:1910.08716*, 2019.

[7] Q. Wang, P. Guo, S. Sun, L. Xie, and J. H. Hansen, "Adversarial regularization for end-to-end robust speaker verification," *Proc. Interspeech 2019*, pp. 4010–4014, 2019.

[8] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 1–7.

[9] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 582–597.

[10] J. Lu, T. Issaranon, and D. Forsyth, "Safetynet: Detecting and rejecting adversarial examples robustly," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 446–454.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[12] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *arXiv preprint arXiv:1611.03814*, 2016.

[13] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[14] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[20] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[21] J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4322–4330.

[22] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[23] G. W. Ding, L. Wang, and X. Jin, "Advertorch v0. 1: An adversarial robustness toolbox based on pytorch," *arXiv preprint arXiv:1902.07623*, 2019.

[24] G. Jin, S. Shen, D. Zhang, F. Dai, and Y. Zhang, "Ape-gan: Adversarial perturbation elimination with gan," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3842–3846.

[25] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[26] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.