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

Enhancing Transferability of Black-box Adversarial Attacks via Lifelong Learning for Speech Emotion Recognition Models

Zhao Ren[1], *Jing Han*[1], *Nicholas Cummins*[1,2], *Björn W. Schuller*[1,3]

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Department of Biostatistics and Health Informatics, IoPPN, King's College London, London, UK
[3] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

zhao.ren@informatik.uni-augsburg.de

Abstract

Well-designed adversarial examples can easily fool deep speech emotion recognition models into misclassifications. The transferability of adversarial attacks is a crucial evaluation indicator when generating adversarial examples to fool a new target model or multiple models. Herein, we propose a method to improve the transferability of black-box adversarial attacks using lifelong learning. First, black-box adversarial examples are generated by an atrous Convolutional Neural Network (CNN) model. This initial model is trained to attack a CNN target model. Then, we adapt the trained atrous CNN attacker to a new CNN target model using lifelong learning. We use this paradigm, as it enables multi-task sequential learning, which saves more memory space than conventional multi-task learning. We verify this property on an emotional speech database, by demonstrating that the updated atrous CNN model can attack all target models which have been learnt, and can better attack a new target model than an attack model trained on one target model only.

Index Terms: Speech Emotion Recognition, Black-box Adversarial Attacks, Transferability, Lifelong Learning

1. Introduction

With the development of artificial intelligence, Speech Emotion Recognition (SER) has been an essential component of Human-Computer Interaction (HCI) [1, 2] and beyond. Recently, deep learning has emerged as a promising technique to train more robust models by extracting highly abstract representations than conventional machine learning methods on big data [3, 4]. A range of deep learning topologies have been successfully applied to the task of SER, such as Convolutional Neural Networks (CNNs) [5], and Recurrent Neural Networks (RNNs) [6].

However, according to recent studies, deep learning models are vulnerable to external adversarial attacks [7], in which the generated examples have high similarities to the real data. This naturalness can be achieved by adding minimal and well-designed perturbations. Moreover, adversarial examples can make a deep learning model prone to misclassifications. In particular, adversarial attacks have the potential to be a threat to SER systems, generating invalid and misinterpreted interactions with users [8]. For instance, adversarial attacks could even be life-threatening in the case of the misdiagnosis of emotion-related mental diseases, e. g., depression [9] and bipolar disorder [10].

This work was supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project under grant agreement No. 766287 (TAPAS), and the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 115902, which receives support from the European Union's Horizon 2020 research and innovation program and EFPIA.

Generating artificial adversarial examples to simulate an attack procedure is an essential step in preventing the real-world attacks. Such can also help further improve the robustness of target deep learning models [11]. In this regard, a range of research works aimed to construct *white-box* and *black-box* adversarial attacks [12, 13]. White-box attacks require the data sources and complete parameters inside the target model, whereas either data or parameters are blind in black-box attacks [13]. While white-box attacks can be guided to generate examples by gradient descent methods [12], it is challenging to search for black-box attacks without or with partial knowledge of the target model.

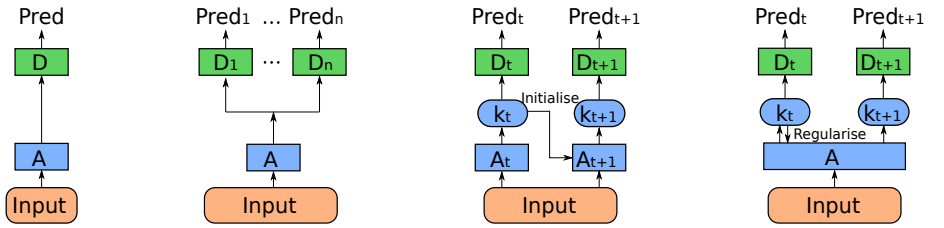
Improving transferability of black-box adversarial attacks promotes the generation of stronger adversarial data, which can transfer among different target models [14]. A highly transferable adversarial attack can deceive not only the already disposed target models but also a new target system [15]. Improving transferability is also helpful to save costs associated with learning a unified attacker when compared to training an independent attacker for each target model. While attacking a target model is typically viewed as a single task, multi-task learning [16, 17] optimises an attacker to cheat multiple target models simultaneously. However, there is increased time and space complexity with training multiple tasks. Transfer learning [18] can help fine-tune an attacker for a new target model with prior knowledge gained from the previous target; however, the attack model gradually forgets the prior knowledge. Inspired by lifelong learning paradigms [19], we propose, to the best of the authors' knowledge, for the first time to improve the transferability of black-box adversarial attacks by utilising a lifelong learning framework. Lifelong learning transfers the prior knowledge with a constraint, so that it is promising to overcome the shortcomings of multi-task and transfer learning, adapting an attack model for a new target model with reducing the time and memory, and without forgetting the prior knowledge.

With this goal in mind, this paper aims to improve the transferability of untargeted black-box adversarial attacks for deep SER models. This work has two main contributions. First, a lifelong learning framework is developed to train an attacker for a sequence of target models. Key results demonstrate that lifelong learning can effectively improve transferability. Second, performances on different target-model orders are compared and analysed. We observe that target-model sequences (shallow→deep) require a bigger constraint to the prior knowledge than those from deep to shallow, as an attacker for a shallow target model has a stronger transferability than that for a deep one.

2. Related Works

Recently, most studies of adversarial attacks were working on image processing tasks [20], whereas only a few works focused

http://dx.doi.org/10.21437/Interspeech.2020-1869



(a) Single-task learning (b) Multi-task learning (c) Transfer learning (d) Lifelong learning

Figure 1: *The frameworks of the four training methods. (a) The input data is fed into an (A)ttacker to generate fake data as the input of a (D)efence. (b) An attack model is trained to attack n defence models in one training procedure. (c) The learnt (k)nowledge k_t from the attacker for the t-th defence is transferred to initialise the $t+1$-th attack model. (d) One attack model is trained to deceivce multiple defences in turn, and the knowledge k_t is returned back to regularise the attacker itself.*

on adversarial attacks for SER systems. There have been a small number of works exploring white-box attacks to cheat a SER model in [8, 21]. A Generative Adversarial Network (GANs) based approach has also been proposed to generate black-box fake data [22]. The study in our paper focuses on generating black-box adversarial data to fool deep SER models.

In the generation of black-box adversarial examples, the original data is first fed into an attack model, which learns to output the required perturbations [23]. CNN models have been utilised as attackers to deceive image processing models [23, 24]. An encoder-decoder CNN was constructed in [23], and deconvolutional layers were also used to generate fake data in [24, 25]. In a previous study [26], an atrous CNN has shown a powerful capability of extracting high-level representations for audio processing. Our study proposes to generate perturbations of emotional speech data using an atrous CNN model.

Improving the transferability of adversarial attacks can promote to train robust deep learning models against practical attacks. Multi-task learning was proposed as an attack technique in a simultaneous segmentation and depth estimation task [27]. Alternatively, some research works have focused on either an ensemble of models or an ensemble of inputs [28, 29]. For example, in [14], an attack model was optimised with an ensemble of the target models. While in [30], the random transformations of the original input were fed to the attackers. However, multi-task learning and ensemble-based methods are time-consuming and require substantial memory space. Transfer learning and smooth regularisation on adversarial perturbations have been demonstrated to solve this problem in [18, 28]. However, it is difficult to train a transferable attack model, as the prior knowledge is forgotten in transfer learning. This study investigates to transfer an attack model across multiple target models with a low time and space complexity. Lifelong learning [19] is for the first time employed to learn a transferable adversarial attack model based on a sequence of target models.

3. Methodology

In this section, the generation of black-box adversarial attacks using a CNN architecture is first discussed. Then, we introduce the proposed lifelong learning framework aiming to improve the transferability of the attacks.

3.1. Black-box Adversarial Attacks

During training a black-box attack model in this study, the parameters of the target model are unknown for the attacker. The input data is first fed into the attack model to generate fake data, which is similar to the original input data (Fig. 1 (a)). The target model (i. e., defence) then fails to produce correct predictions on the generated data. In the following subsections, the attack and defence models are presented respectively.

3.1.1. Attack

Regarding the input, log-Mel spectrograms are utilised due to their strong performance in the task of SER [8, 31]. The real log-Mel spectrograms and the labels are denoted as $(\boldsymbol{x}, y)$, and the generated adversarial data as $\boldsymbol{x'}$. A CNN model can then be trained to output two-dimensional adversarial perturbations $\boldsymbol{\eta}$. Finally, the adversarial data is obtained by $\boldsymbol{x'} = \boldsymbol{x} + \boldsymbol{\eta}$.

An atrous CNN model is proposed to generate the adversarial perturbations due to its ability to output feature maps with the same size as the input [26]. To effectively train an atrous CNN attack model, the loss function has two objects: one is to cheat the defence f, and the other is to minimise the difference between the fake and real data. Hence, the loss function is a weighted sum of the loss functions for the two goals:

$$loss = \alpha loss_{cla}(f(\boldsymbol{x'})) + (1-\alpha) loss_{MSE}(\boldsymbol{x'}, \boldsymbol{x}), \quad (1)$$

$$loss_{cla}(f(x')) = \max(f(x')_l - \max(f(x')_{other}), 0), \quad (2)$$

where α is a hyperparameter that balances the two loss functions $loss_{cla}$ and $loss_{MSE}$. The loss function $loss_{cla}$ aims to pull down the classification performance of defence on the fake data, and $loss_{MSE}$ aims to improve the similarity of the fake and real data using Mean Squared Error (MSE). The function $loss_{cla}$ is defined by the difference between the probability of the correct label l and the maximum probability of $other$ classes. The computation of $loss_{cla}$ is referred to as the Carlini-Wagner (C&W) loss function [32]. This approach has been used to generate adversarial attacks in image processing tasks [20, 33].

3.1.2. Defence

With the generated adversarial data, the defence model is also a CNN architecture, as CNNs are suitable for extracting high-level features from log-Mel spectrograms for classification [26]. As training an attacker is the main focus in this study, the defence models are pre-trained to classify emotional classes, and employed to help train a highly transferable attack model only.

3.2. Lifelong Learning

While training an attack model, improving the attacker's capability to deceive a new defence or multiple defences is a challenging task. Training an attack model for one defence can be viewed as a task, and more defences lead to multiple tasks. Single-task learning (Fig. 1 (a)) is time consuming, as it learns one attacker for each target model. Multi-task learning (Fig. 1 (b)) is also enduring and requires a big memory when training an

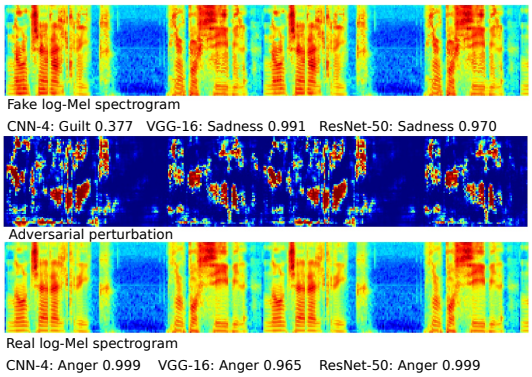


Figure 2: *The real (bottom), fake (top) log-Mel spectrograms, and adversarial perturbation (middle) of NP_m_19_ang08b.wav. The three CNN models classify the real data correctly as anger, but give wrong predictions of guilt or sadness on the fake data.*

attacker for multiple defences simultaneously. Transfer learning approaches (Fig. 1 (c)) can be used to fine-tune the attacker for a new defence; however, it results in prior knowledge being forgotten. Different from the three learning strategies, lifelong learning (Fig. 1 (d)) trains one attacker for each defence individually, transferring the prior knowledge back to the attack model and utilising regularisation in order to retain prior knowledge.

Elastic Weight Consolidation (EWC) [34] is applied in this study to achieve lifelong learning. The idea of EWC is to give elastic constraints to the parameters of the current trained model, optimising the important parameters to be close to those of the previous model. The important parameters are given high constraints, and unimportant parameters are given low constraints. The importance of each parameter is computed by the diagonal of the Fisher information matrix [35]. This step is equivalent to the second derivative of the loss function close to the minimum value. Hence, applying EWC helps remember the important parameters while attacking the previous defence, so that the attack model can deceive both the previous and the current target models. EWC is achieved by adding a regularisation term to the loss function:

$$loss_{ewc} = loss + \lambda \sum_i F_i(\theta_i - \theta_i^*)^2, \tag{3}$$

where $loss$ is the loss function of the current attacker computed by Eq. (1), λ is a hyperparameter deciding the global importance of the parameters of the attack model for previous defence, and $\boldsymbol{F}$ denotes the Fisher information matrix. The square of the difference between the current parameters $\boldsymbol{\theta}$ and previous parameters $\boldsymbol{\theta}^*$ is minimised to remember $\boldsymbol{\theta}^*$ according to $\boldsymbol{F}$. Finally, the regularisation leads to the learning of a new attack model to fool both target models.

An example of training an attack model using lifelong learning is shown in Fig. 2. The three CNN models (CNN-4, VGG-16, and ResNet-50), which were trained on a real emotional speech data, predict the real log-Mel spectrogram correctly to *anger* [8]. However, all of the three CNN models fail to predict the correct emotional label on the adversarial log-Mel spectrogram.

4. Experimental Results

4.1. Database

An Italian emotional speech corpus – the Database of Elicited Mood in Speech (DEMoS) [36] is used throughout this work. During recording the speech signals, the subjects' emotions were induced by an arousal-valence progression. Totally, 332 neutral and 9 365 emotional speech samples were collected from 68 speakers (23 females, 45 males). The neutral speech samples were not considered in the experiments, as *neutral* is more minor than other classes. Therefore, this study works on DEMoS with seven classes: *anger*, *disgust*, *fear*, *guilt*, *happiness*, *sadness*, and *surprise*. The data was partitioned speaker-independently into train, development, and test sets (cf. [8]).

4.2. Experimental Setup

As in [8], the data is resampled into 16 kHz, and then the log-Mel spectrograms are extracted with a size of $(373, 64)$. As mentioned in Section 3.1.1, an atrous CNN model is employed to train an attacker. The atrous CNN model consists of four convolutional layers with channel numbers of $\{64, 128, 64, 1\}$, dilation values of $\{1, 2, 4, 8\}$, and a kernel size of $(5, 5)$. Each convolutional layer is followed by a batch normalisation and a 'relu' (for the first three layers) or a 'sigmoid' (for the final layer) function. As the output of the attacker is a minimal adversarial perturbation, the data values in $[0, 1]$ produced by a 'sigmoid' function is more helpful for the optimisation than the data values in $[0, +\infty]$ from a 'relu' function. To validate the effectiveness of the fake data, three pre-trained defence models in [8] are utilised, including CNN-4, VGG-16, and ResNet-50. These CNNs have four, 16, and 50 convolutional layers respectively, and a log-softmax as the final layer for classification.

During training an atrous CNN attack model, the optimiser is 'Adam' with a learning rate of 0.0001, and the training procedure stops at the $10\,000$ iterations. To stabilise the training models, the learning rate is reduced into $90\,\%$ of its current value at each $1\,000$ iterations. The attacker is then trained for $1\,000$ iterations for a new defence using lifelong learning. The hyper-parameter is set as $\alpha = 0.02$ experientially, and λ is optimised from $\{1e4, 1e5, 1e6, 1e7\}$ on the development set.

To analyse the effect of defence orders in lifelong learning, an attacker is trained on two types of defence sequences: one is going deeper (CNN-4 $\rightarrow$ VGG-16 $\rightarrow$ ResNet-50) (i. e., clockwise), and the other is going shallower ((ResNet-50 $\rightarrow$ VGG-16 $\rightarrow$ CNN-4)) (i. e., counterclockwise). For each sequence, the attackers trained on both the first two defences and the whole sequence are validated to attack the three CNN defences. Finally, the Unweighted Average Recall (UAR) is utilised as the evaluation metric with the consideration of class imbalance.

4.3. Results

In the experimental results, Fig. 3 shows the attack performance of the atrous CNN models on the development set, where a lower UAR value is corresponding to a stronger attacker. We can see that, in each defence sequence, a bigger value of λ leads to more constraint to the parameters of the attacker, so that more prior knowledge from the previous target model is remembered. For example, in the clockwise sequence, the attack model can better fool CNN-4 with λ increasing. While comparing the attackers trained on two target models, the attacker in the clockwise sequence performs better on the third model ResNet-50 than the attacker in the counterclockwise sequence working on CNN-4. This indicates that a shallow target model helps train a more transferable attacker than a deep one. In this regard, remembering more prior knowledge of a shallow attack model is necessary to improve the attack transferability. To perform well on all three defences, the attacker trained on the clockwise sequence requires a bigger value of λ ($1e6$ on the two-defence sequence, and $1e5$ on three) than that on the counterclockwise sequence ($1e4$ on the

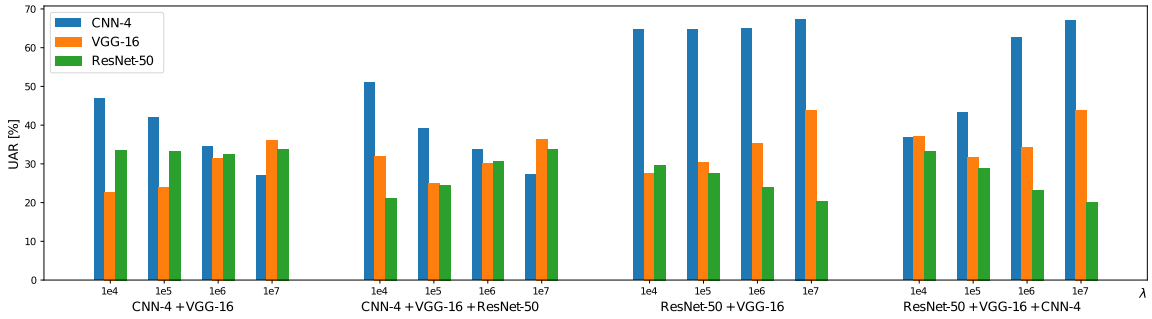


Figure 3: *The performances of lifelong learning with different λ values on the development set. The attack models are trained on the four sequences of target models (bottom), and the transferability is tested on three target models (CNN-4, VGG-16, and ResNet-50).*

Table 1: *The performance comparison of four learning methods on the (dev)elopment and test sets. The attackers are learnt on different defence sequences, and then the generated fake data are ultilised to attack the three target models (CNN-4, VGG-16, and ResNet-50).*

		CNN-4		VGG-16		ResNet-50	
Method	UAR [%]	Dev.	Test	Dev.	Test	Dev.	Test
Single-task learning	CNN-4	28.5	21.9	47.5	46.1	39.3	54.5
	ResNet	68.8	56.0	49.7	42.1	20.0	20.7
Multi-task learning	CNN-4 + VGG-16	29.4	23.0	21.8	17.9	31.5	49.1
	ResNet-50 + VGG-16	63.5	52.6	24.0	19.8	21.2	22.9
	CNN-4 + VGG-16 + ResNet-50	33.1	28.0	22.9	18.5	21.0	23.9
Transfer learning	CNN-4 + VGG-16	45.2	48.6	19.3	16.1	30.8	54.7
	CNN-4 + VGG-16 + ResNet-50	52.7	41.4	30.7	28.8	18.5	18.9
	ResNet-50 + VGG-16	59.3	53.0	22.9	18.1	25.1	37.6
	ResNet-50 + VGG-16 + CNN-4	24.9	22.0	30.2	29.0	24.0	34.2
Lifelong learning	CNN-4 + VGG-16	34.6	30.1	31.4	29.7	32.5	46.9
	CNN-4 + VGG-16 + ResNet-50	39.2	38.5	25.0	26.4	24.5	33.4
	ResNet-50 + VGG-16	64.7	54.1	27.4	21.7	29.5	36.3
	ResNet-50 + VGG-16 + CNN-4	43.3	33.9	31.6	29.4	28.8	30.9

two-defence sequence, and $1e5$ on three). Finally, both attack models trained on the whole of two sequences can successfully fool the three target models.

Further, the results of lifelong learning are compared to the other three learning frameworks in Table 1. Single-task learning performs the worst on new target models as the attack model learnt from one target model has a low transferability. Multi-task learning trains the most transferable attack model to deceive the three target models, but it is time and space consuming as aforementioned. While using transfer learning, the attack model forgets the prior knowledge during fine-tuning, especially in the results of the clockwise sequences. Compared to transfer learning, lifelong learning trains the attack model with remembering the prior knowledge, so that the attacker can cheat all target models which have been trained on.

Moreover, while comparing the results of two types of sequences using transfer and lifelong learning, lifelong learning performs better than transfer learning on the clockwise sequences, and the performances on the counterclockwise using lifelong learning are comparable with those by transfer learning. This is in consistency with the aforementioned analysis, which is that the attackers trained on the clockwise sequences require more constraint than those on the counterclockwise sequences, because a shallower target model leads to a more transferable attacker. Transfer learning can be viewed as a special case of lifelong learning without constraint, so that it adapts the attacker to fool shallow defence models in the counterclockwise sequences, remembering less prior knowledge from deep defences than typical lifelong learning. Although the order of target models can affect the attack performance, it is always unknown how deep a new defence is in practical. Therefore, it is difficult to train a transferable attack model using transfer learning when the new target model is shallower than the existed defences. Hence, lifelong learning can help the attack model adapt to fool the new target model and the target models which have been learnt on.

5. Conclusions and Future Work

In summary, we trained an atrous Convolutional Neural Network (CNN) as a black-box adversarial attack model, and improved its transferability using lifelong learning. The proposed lifelong learning framework trained the attack model successfully cheating the three CNN classifiers of CNN-4, VGG-16, and ResNet-50. Further, the effect of different defence orders was analysed, discovering that a sequence from shallow to deep models needs a bigger constraint than an inverse one.

In future efforts, more lifelong learning methods will be investigated to further improve the transferability, and more defences will be utilised to test the transferability of attackers. As the attackers in this study generated a corresponding adversarial perturbation for a real data, universal black-box attacks will be focused on to generate a universal perturbation for all real data. Lifelong learning will be further applied to improve the transferability of universal black-box adversarial attacks.

6. References

[1] R. Khalil, E. Jones, M. Babar, T. Jan, M. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, Aug. 2019.

[2] J. Han, Z. Zhang, M. Schmitt, Z. Ren, F. Ringeval, and B. Schuller, "Bags in bag: Generating context-aware bags for tracking emotions from speech," in *Proc. INTERSPEECH*, Hyderbad, India, 2018, pp. 3082–3086.

[3] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, May 2018.

[4] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1691–1695.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[6] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97 515–97 525, July 2019.

[7] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning models," in *Proc. CVPR*, Salt Lake City, UT, 2018, pp. 1625–1634.

[8] Z. Ren, A. Baird, J. Han, Z. Zhang, and B. Schuller, "Generating and protecting against adversarial attacks for deep speech-based emotion recognition models," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 7184–7188.

[9] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proc. AVEC*, Nice, France, 2019, pp. 3–12.

[10] Z. Ren, J. Han, N. Cummins, Q. Kong, M. Plumbley, and B. Schuller, "Multi-instance learning for bipolar disorder diagnosis using weakly labelled speech data," in *Proc. DPH*, Marseille, France, 2019, pp. 79–83.

[11] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *Proc. ICLR*, Vancouver, Canada, 2018, 17 pages.

[12] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, San Diego, CA, 2015, 11 pages.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICML*, Sydney, Australia, 2017, 10 pages.

[14] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. ICLR*, Toulon, France, 2017, 14 pages.

[15] L. Wu, Z. Zhu, C. Tai, and W. E., "Understanding and enhancing the transferability of adversarial examples," arXiv preprint arXiv:1802.09707, 2018, 15 pages.

[16] M. Zhao, B. An, Y. Yu, S. Liu, and S. Pan, "Data poisoning attacks on multi-task relationship learning," in *Proc. AAAI*, New Orleans, LA, 2018, pp. 2628–2635.

[17] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 6705–6709.

[18] B. Wang, Y. Yao, B. Viswanath, H. Zheng, and B. Zhao, "With great training comes great vulnerability: Practical attacks against transfer learning," in *Proc. SEC*, Baltimore, MD, 2018, pp. 1281–1297.

[19] G. Parisi, R. Kemker, J. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, May 2019.

[20] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, Feb. 2018.

[21] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," in *Proc. DYNAMICS*, San Juan, PR, 2017, 8 pages.

[22] S. Latif, R. Rana, and J. Qadir, "Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness," arXiv preprint arXiv:1811.11402, 2018, 7 pages.

[23] G. Zhao, M. Zhang, J. Liu, and J.-R. Wen, "Unsupervised adversarial attacks on deep feature-based retrieval with GAN," arXiv preprint arXiv:1907.05793, 2019, 10 pages.

[24] J. Hayes and G. Danezis, "Learning universal adversarial perturbations with generative models," in *Proc. SPW*, San Francisco, CA, 2018, pp. 43–49.

[25] S. Baluja and I. Fischer, "Adversarial transformation networks: Learning to generate adversarial examples," arXiv preprint arXiv:1703.09387, 2017, 13 pages.

[26] Z. Ren, Q. Kong, J. Han, M. Plumbley, and B. Schuller, "Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 56–60.

[27] Z. Zhang, X. Zhu, Y. Li, X. Chen, and Y. Guo, "Adversarial attacks on monocular depth estimation," arXiv preprint arXiv:2003.10315, 2020, 8 pages.

[28] W. Zhou, X. Hou, Y. Chen, M. Tang, X. Huang, X. Gan, and Y. Yang, "Transferable adversarial perturbations," in *Proc. ECCV*, Munich, Germany, 2018, pp. 452–467.

[29] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. CVPR*, Salt Lake City, UT, 2018, pp. 9185–9193.

[30] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proc. CVPR*, Long Beach, CA, 2019, pp. 2730–2739.

[31] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, Oct. 2017.

[32] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. SP*, San Jose, CA, 2017, pp. 39–57.

[33] A. Bose and P. Aarabi, "Adversarial attacks on face detectors using neural net based constrained optimization," in *Proc. MMSP*, Vancouver, Canada, 2018, pp. 1–6.

[34] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[35] G. Desjardins, K. Simonyan, R. Pascanu, and K. Kavukcuoglu, "Natural neural networks," in *Proc. NIPS*, Montréal Canada, 2015, pp. 2071–2079.

[36] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. Schuller, "DEMoS: An Italian emotional speech corpus," *Language Resources and Evaluation*, pp. 1–43, Feb. 2019.