Interspeech 2022
18-22 September 2022, Incheon, Korea

Hear No Evil: Towards Adversarial Robustness of Automatic Speech Recognition via Multi-Task Learning

Nilaksh Das, Duen Horng Chau

Georgia Institute of Technology, USA
{nilakshdas,polo}@gatech.edu

Abstract

As automatic speech recognition (ASR) systems are now being widely deployed in the wild, the increasing threat of adversarial attacks raises serious questions about the security and reliability of using such systems. On the other hand, multi-task learning (MTL) has shown success in training models that can resist adversarial attacks in the computer vision domain. In this work, we investigate the impact of performing such multi-task learning on the adversarial robustness of ASR models in the speech domain. We conduct extensive MTL experimentation by combining semantically diverse tasks such as accent classification and ASR, and evaluate a wide range of adversarial settings. Our thorough analysis reveals that performing MTL with semantically diverse tasks consistently makes it harder for an adversarial attack to succeed. We also discuss in detail the serious pitfalls and their related remedies that have a significant impact on the robustness of MTL models. Our proposed MTL approach shows considerable absolute improvements in adversarially targeted WER ranging from 17.25 up to 59.90 compared to single-task learning baselines (attention decoder and CTC respectively). Ours is the first in-depth study that uncovers adversarial robustness gains from multi-task learning for ASR.

Index Terms: ASR, adversarial robustness, multi-task learning

1. Introduction

Automatic speech recognition (ASR) systems have penetrated our daily lives with real-world applications such as digital voice assistants, IVR and news transcription. These are increasingly relying on deep learning methods for their superior performance. At the same time, a mature body of adversarial machine learning research has exposed serious vulnerabilities in these underlying methods [1, 2, 3, 4, 5, 6, 7], raising grave concerns regarding the trustworthiness of ASR applications. There is an urgent need to address these vulnerabilities for restoring faith in using ASR for safety-critical functions. Our study aims to push the envelope in this direction with a meticulous approach.

An adversarial attack on an ASR model allows the attacker to introduce faint noise to a speech sample that can influence the model into making an exact transcription of the attacker's choosing. Since this targeted adversarial scenario is considered more threatening for an ASR system as compared to adding noise that leads to some arbitrary prediction [3, 6], we focus on studying the characteristics of ASR models that can thwart *targeted* adversarial attacks. Research has shown that adversarial examples are manifestations of non-robust features learned by a deep learning model [8]. Hence, our objective is to regularize the ASR model training paradigm so as to learn robust features that can resist such attacks.

Multi-task learning (MTL) is one such approach that has shown some success in this aspect for computer vision tasks by making the underlying models more resilient to adversarial attacks [9, 10]. However, it is unclear whether such robustness would transfer to the audio modality. Moreover, hybrid ASR models are often trained with semantically equivalent tasks like CTC and attention decoding [11]. This raises interesting questions about their inherent adversarial robustness. In this work, we aim to study the impact of MTL on the adversarial robustness of ASR models, and compare it to robustness of single-task learning (STL). Here, we consider MTL as jointly training a shared feature encoder with multiple losses from diverse task heads. Our expectation is that MTL would induce the encoder to learn a robust feature space that is harder to attack [8]. We consider semantically equivalent as well as semantically diverse tasks for performing MTL. We find that a combination of both types of tasks is necessary to most effectively thwart adversarial attacks. We also discuss serious pitfalls related to inference for MTL models that have an adverse impact on robustness. Finally, we demonstrate remedies for such pitfalls that significantly make it harder for an attacker to succeed.

Contributions

- **First MTL Study of Adversarial Robustness for ASR.** To the best of our knowledge, this is the first work to uncover adversarial robustness gains from MTL for ASR models.
- **Extensive Evaluation.** We perform extensive experimentation with one of the most powerful adversarial attacks, and evaluate models trained with semantically equivalent as well as semantically diverse tasks across a wide range of training hyperparameters and strong adversarial settings.
- **Robustness of Hybrid ASR Inference.** Our study exposes the extreme vulnerability of using CTC head for inference in hybrid CTC/attention models; while showing that MTL training with CTC and attention loss improves resiliency to adversarial attacks if CTC head is dropped during inference.
- **Robust ASR with Semantically Diverse Tasks.** Our thorough analysis reveals that performing MTL with semantically diverse tasks such as combining ASR with accent classification makes it most difficult for an attacker to induce a maliciously targeted prediction. Our MTL approach shows considerable absolute improvements in adversarially targeted WER ranging from 17.25 up to 59.90 compared to STL baselines (attention decoder and CTC respectively).

2. Related Works

Several adversarial attacks have been proposed for maliciously influencing ASR models by leveraging model gradients to optimize a faint perturbation to the input speech [1, 2, 4, 5, 7]. Many such gradient-based adversarial techniques can be formulated as variations to the projected gradient descent (PGD) method [12], which is one of the strongest digital perturbation attacks proposed in the adversarial ML literature. In this work, we experiment with the *targeted* PGD attack, as this attack scenario is considered more threatening for ASR systems [3, 6].

10.21437/Interspeech.2022-11361

Defenses proposed to evade adversarial attacks on ASR models mostly employ input preprocessing [13, 14, 15, 16] that places an undue burden on inference-time computation. Adversarial training has also shown some success in improving adversarial robustness [12, 17]. However, these methods are extremely computationally expensive. MTL with a shared backbone has the potential to provide a reasonable middle ground as it is much less computationally expensive than adversarial training, and does not introduce any inference-time load. Many studies have indeed looked at MTL in the context of ASR robustness [11, 18, 19, 20, 21, 22]. However, the bulk of such works focus mostly on improvement in benign performance (when no attack is performed) or robustness to arbitrary background noise. Ours is the first work to study the impact of MTL on ASR models specifically in the context of adversarial attacks.

3. Approach

3.1. ASR with Multi-Task Learning

In this work, we study the impact of MTL on adversarial robustness of ASR models by jointly training a shared feature encoder ϕ. The encoder takes a speech sample x as input, and outputs $\phi(x)$, which can be considered as a latent sequence embedding in a shared feature space. The feature embedding is then passed to various task heads with corresponding losses. We consider semantically diverse tasks in addition to semantically equivalent tasks for performing MTL. The semantically equivalent task heads for ASR are: (1) CTC and (2) attention decoder. We also jointly train a discriminator for a semantically diverse task such as accent classification.

For the ASR task, we denote the ground-truth transcription as $\bar{y}$, and the CTC loss [23] and decoder attention loss [24] as $\mathcal{L}_{\text{CTC}}$ and $\mathcal{L}_{\text{DEC}}$ respectively. We compute the loss for ASR as:

$$\mathcal{L}_{\text{ASR}}(x, \bar{y}) = \lambda_C^{(t)} \mathcal{L}_{\text{CTC}}(x, \bar{y}) + (1 - \lambda_C^{(t)}) \mathcal{L}_{\text{DEC}}(x, \bar{y}) \quad (1)$$

Consequently, for performing joint ASR inference, we denote the CTC and decoder scoring functions [11] as f_{CTC} and f_{DEC} respectively. Finally, we determine the predicted output transcription $\hat{y}$ as follows:

$$\hat{y} = \lambda_C^{(i)} f_{\text{CTC}}\big(\phi(x)\big) + (1 - \lambda_C^{(i)}) f_{\text{DEC}}\big(\phi(x)\big) \quad (2)$$

Note here that $\lambda_C^{(t)}$ and $\lambda_C^{(i)}$ are training and inference weights respectively. Generally, we follow that $\lambda_C^{(i)} = \lambda_C^{(t)}$ in our experiments while performing inference, unless otherwise specified. Correspondingly, setting $\lambda_C^{(i)}$=1.0 allows us to use only the trained CTC head for inference, and vice-versa for the trained decoder head by setting $\lambda_C^{(i)}$=0.0.

For accent classification, we are given an accent label $\bar{z}$ that is to be predicted for a speech sample x. Besides being semantically diverse, accent classification is also functionally distinct as it is a *sequence-to-label* task, compared to the *sequence-to-sequence* task for ASR. Hence, denoting the cross-entropy loss for the discriminator head as $\mathcal{L}_{\text{DIS}}$, we compute the full MTL loss for jointly training the model as:

$$\mathcal{L}_{\text{MTL}}(x, \bar{y}, \bar{z}) = \lambda_A^{(t)} \mathcal{L}_{\text{ASR}}(x, \bar{y}) + (1 - \lambda_A^{(t)}) \mathcal{L}_{\text{DIS}}(x, \bar{z}) \quad (3)$$

From Equations (1) and (3), we can see that modulating the $\lambda_A^{(t)}$ and $\lambda_C^{(t)}$ weights allows us to independently modulate the effect of various heads during training, *e.g.*, setting $\lambda_A^{(t)}$=1.0 and $\lambda_C^{(t)}$=1.0 allows us to train using only the CTC head. Conversely, we can train only the decoder head by setting $\lambda_A^{(t)}$=1.0 and $\lambda_C^{(t)}$=0.0. These can also be considered as the single-task learning (STL) baselines. With $\lambda_A^{(t)} < 1.0$, we can train the model with the discriminator head included. We perform extensive experiments across a range of these hyperparameters to study the impact of MTL on ASR robustness to attacks.

3.2. Adversarial Attack on ASR with PGD

An adversarial attack on ASR introduces an inconspicuous and negligible perturbation δ to a speech sample that confuses the ASR model into making an incorrect prediction. In this work, we focus on the projected gradient descent (PGD) attack [12]. Specifically, we consider the *targeted* PGD attack, as it is considered more malicious and threatening for ASR systems [3, 6]. Given, a target transcription $\tilde{y}$, the targeted attack aims to minimize the following inference loss function so as to force the ASR model into making a prediction of the attacker's choosing:

$$\mathcal{L}_{\text{ADV}}(x, \tilde{y}) = \lambda_C^{(i)} \mathcal{L}_{\text{CTC}}(x, \tilde{y}) + (1 - \lambda_C^{(i)}) \mathcal{L}_{\text{DEC}}(x, \tilde{y}) \quad (4)$$

The PGD attack is an iterative attack that consists of two main stages. The first stage is the perturbation stage, wherein a small perturbation of step size α is computed in the direction of the gradient of $\mathcal{L}_{\text{ADV}}$ with respect to the sample from the previous iteration. This perturbation having a magnitude of α is added to the sample. The second stage is the projection stage that ensures that the perturbed sample remains within an ϵ-ball of the original input. This is also called the $L2$ threat model, as it uses the $L2$ norm for limiting the perturbation. Hence, the targeted PGD attack optimizes the perturbation δ as:

$$x_{\text{ADV}} = \arg\min_{\delta} \mathcal{L}_{\text{ADV}}(x + \delta, \tilde{y}), \text{ s.t. } \|\delta\|_2 \leq \epsilon \quad (5)$$

The perturbation and projection stages are performed iteratively, and the computational cost to the attacker increases with increasing number of iterations. In order to isolate and study the impact of our MTL training on adversarial robustness, we perform greedy ASR inference while implementing the attack. Since the adversarial objective is to induce a maliciously targeted prediction as opposed to untargeted arbitrary predictions, we analyze the MTL robustness to specifically evade such targeted attacks. Hence, we examine the adversarially targeted word error rate (abbreviated as **AdvTWER** hereon) in this work, which reports the word error rate of the prediction $\hat{y}$ with respect to the adversarial target transcription $\tilde{y}$, *i.e.,* a higher AdvTWER implies that the model is more robust in evading the attack. Conversely, a lower AdvTWER means that the attack was more successful. We report the adversarially targeted word error rate as we find it to be more compelling in studying the MTL robustness. Correspondingly, we also observe equivalent robustness trends with respect to the benign word error rate.

4. Experiment Setup

4.1. Data

We use annotated speech data from Mozilla's Common Voice dataset [25] for our experiments. Common Voice consists of naturally spoken human speech in a variety of languages. The dataset also includes demographic metadata like age, sex and accent that is self-reported by speakers, and the speech is validated by annotators. Specifically, we use the English language speech and extract accent-labeled speech samples for US and Indian accents, which are among the most abundantly available accented speech in the dataset. Using the splits as defined by the dataset, we get $\sim$260K samples for training and $\sim$1.2K samples for validation. Finally, we report the performance metrics on $\sim$1K samples obtained from the test split.

Table 1: *AdvTWER (↑ is more robust) with $\lambda_A^{(t)}$=1.0. Training with CTC and dropping CTC for inference ($\lambda_C^{(i)}$=0.0) is more robust.*

	$\lambda_C^{(i)} = \lambda_C^{(t)}$					$\lambda_C^{(i)} = 0.0$		
	STL (Decoder) ← MTL with joint inference →				STL (CTC)	MTL with Decoder inference		
$\lambda_C^{(t)} \rightarrow$	0.0	0.3	0.5	0.7	1.0	0.3	0.5	0.7
PGD-500	93.35	59.87	61.55	76.64	37.53	96.80	**99.95**	97.06
PGD-1000	72.31	35.21	37.49	53.91	15.19	79.36	81.29	**82.45**
PGD-1500	57.57	23.94	26.09	42.44	8.99	67.16	68.45	**69.57**
PGD-2000	49.79	19.22	20.40	36.99	7.14	60.15	**63.29**	60.56

gray = STL; highlighted = MTL > STL; **bold** = highest in row

4.2. Model Architecture and Training

We leverage a pre-trained hybrid CTC-attention model that is publicly available [26], and fine-tune it using multi-task learning. The model consists of a conformer-based encoder [27] that is shared by multiple task heads. The encoder has 12 conformer blocks, each with 8 attention heads and a hidden unit size of 2048. The output size of the encoder is 512, which is consequently the size of the sequence embeddings from the shared feature space. The MTL model employs a CTC head and a decoder head for ASR. The decoder has 6 transformer blocks, with 8 attention heads and a hidden unit size of 2048. The ASR heads output scores for 5000 subword units. For accent classification, we use a dense feed-forward discriminator that passes the mean sequence embedding through 5 fully connected layers, and finally outputs class labels corresponding to the accents. The MTL models are implemented using the ESPnet2 toolkit [28]. We train several MTL models by extensively modulating the $\lambda_A^{(t)}$ and $\lambda_C^{(t)}$ weights. Each MTL model is trained with a learning rate of 10^{-3} for 30 epochs, and we pick the model with the lowest validation loss from all the epochs.

4.3. Adversarial Setting

For the targeted PGD attack, we generate a fixed set of adversarial transcriptions having varying lengths. For each sample, the target transcription $\tilde{y}$ is chosen in a deterministic manner by finding the adversarial transcription from the fixed set having the closest length to the original transcription $\bar{y}$. To ensure there is minimal word overlap between the original transcriptions and the adversarial transcriptions, we use a lorem-ipsum generator for generating the adversarial transcriptions. We leverage the targeted PGD attack implemented using the Adversarial Robustness Toolbox [29] for our experiments. While attacking the ASR model, we focus solely on the MTL robustness by performing greedy inference for determining the gradients, and no additional LMs are used. The targeted PGD attack is performed using an $L2$ threat model in a white-box attack for inference. We use an extremely strong perturbation limit of ϵ=2.0, beyond which we observed that the perturbed samples would no longer remain within natural constraints. As the attack's computational cost increases with increasing number of steps, we report the AdvTWER metric for performing multiple attack steps, up to as high as 2000 steps. We use a step size α=0.05, allowing the attack to make fine-grained perturbations at each step.

5. Results

Our extensive experimentation reveals that increasing the proportion of multi-task learning (MTL) while training significantly improves the model's resiliency to adversarial attacks. Before analysing the adversarial performance of the models, we first briefly discuss the benign performance, *i.e.*, when no attack is performed. On the test set, the single-task learning (STL) baselines have a benign word error rate (WER) of 20.85 and 16.62 for the CTC and the decoder heads respectively. As has been documented by previous studies [18, 30, 20], models trained with MTL also show better benign performance compared to STL. For example, we find that training an MTL model with a decoder and discriminator ($\lambda_A^{(t)}$=0.8; $\lambda_C^{(t)}$=0.0) yields a benign WER of 15.86 on the test set. For the accent classification task as well, we observe a reasonable bening accuracy of ∼90% for the MTL models with $\lambda_A^{(t)} < 1.0$.

We now shift our focus to the adversarial performance of ASR. An overview of the MTL robustness trends is also depicted in Figure 1, comparing different MTL combinations with STL baselines. We see that training with MTL makes it consistently harder for an attacker to succeed.

5.1. MTL with CTC and Attention Decoder

We first study the adversarial performance of the semantically equivalent task heads by training jointly using the CTC and decoder heads. We train multiple models by setting $\lambda_A^{(t)}$=1.0 and modulating $\lambda_C^{(t)}$ from values $\{0.0, 0.3, 0.5, 0.7, 1.0\}$. We perform the PGD attack on each of these models as decribed in Section 4.3. For performing the attack, we follow two inference modes: (1) using the CTC head with same inference weights as training, *i.e.*, $\lambda_C^{(i)} = \lambda_C^{(t)}$; and (2) drop the CTC head for inference, *i.e.*, $\lambda_C^{(i)}$=0.0. It will become clear in the following discussion why we follow this. The adversarial performance for these inference modes is reported in Table 1 using the AdvTWER metric for various attack steps.

When the trained CTC head is included for inference, *i.e.*, $\lambda_C^{(i)} = \lambda_C^{(t)}$, we see from Table 1 that performing MTL with a combination of CTC and decoder heads has no robustness benefit compared to the STL decoder baseline. In fact, the STL CTC baseline itself is significantly more vulnerable to the adversarial attacks. Performing MTL moderately improves the robustness compared to the STL CTC baseline but it is still not better than the STL decoder baseline. This implies that the attacker is able to easily influence the CTC head into hijacking the overall joint prediction. This can be seen in Figure 1 with the clear gap between the gray STL baselines and the MTL model with CTC included in inference ($\lambda_A^{(t)}$=1.0; $\lambda_C^{(t)}$=$\lambda_C^{(i)}$=0.5). Decreasing $\lambda_C^{(t)}$ from 0.7 to 0.3 clearly shows worsening robustness in Table 1. This means that attacking a less trained CTC head (with lower $\lambda_C^{(t)}$) having a lower impact on the overall prediction (due to correspondingly lower $\lambda_C^{(i)}$) is still able to overpower the more trained decoder head scores. This evident vulnerability of the CTC head may be attributed to the well studied peaky behavior of the CTC loss [31, 32] that would allow the attacker to induce

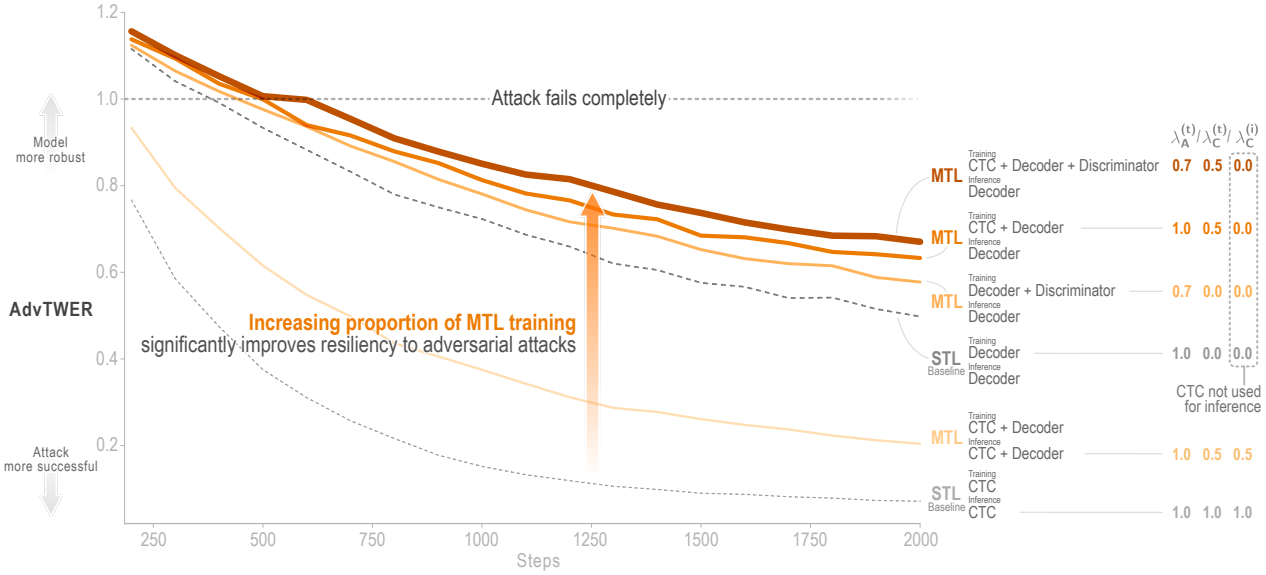

Figure 1: *Adversarial performance for various MTL combinations. MTL outperforms STL baselines when CTC is dropped for inference.*

Table 2: *AdvTWER (↑ is more robust) with* $\lambda_C^{(t)}$=0.0. *Training the decoder with a discriminator shows better robustness.*

$\lambda_A^{(t)} \rightarrow$	1.0	0.9	0.8	0.7	0.6	0.5
PGD-500	93.35	89.60	95.48	**97.60**	97.04	90.94
PGD-1000	72.31	68.67	73.02	78.07	**78.57**	70.25
PGD-1500	57.57	54.88	60.88	**65.24**	63.15	57.88
PGD-2000	49.79	47.28	51.23	**57.75**	55.61	50.43

gray = STL (Decoder); highlighted = MTL > STL; **bold** = highest in row

Table 3: *AdvTWER (↑ is more robust) with* $\lambda_A^{(t)}$=0.7, $\lambda_C^{(i)}$=0.0. *Training all heads combined shows most superior robustness.*

$\lambda_C^{(t)} \rightarrow$	0.0	0.3	0.5	0.7
PGD-500	97.60	97.02	100.65	**101.04**
PGD-1000	78.07	76.83	**85.06**	83.69
PGD-1500	65.24	62.75	**73.70**	70.67
PGD-2000	57.75	56.57	**67.04**	65.79

gray = Baseline; highlighted = MTL > Baseline; **bold** = highest in row

overconfident prediction scores through the CTC head.

Next we study the effect of performing MTL by training the CTC head, but dropping the CTC head during inference, *i.e.*, $\lambda_C^{(i)}$=0.0. This means that the attacker can no longer manipulate the CTC head during inference, but the shared encoder is still jointly trained using the CTC loss. Immediately, we see the adversarial performance of the MTL models beat both the STL baselines in Table 1. It is consistently harder for the attacker to attack the MTL models across many attack steps. This is also visualized in Figure 1 for $\lambda_A^{(t)}$=1.0, $\lambda_C^{(t)}$=0.5 and $\lambda_C^{(i)}$=0.0.

5.2. Robust ASR with Semantically Diverse Tasks

We now study the impact of performing MTL with semantically diverse tasks. As we saw in Section 5.1 that the CTC head is extremely vulnerable to adversarial attacks, we first examine MTL with the decoder and discriminator combination, *i.e.*, we set $\lambda_C^{(t)}$=0.0. Table 2 shows the adversarial performance for training the decoder and discriminator heads by modulating $\lambda_A^{(t)}$ from values $\{1.0, 0.9, 0.8, 0.7, 0.6, 0.5\}$. We see that AdvTWER for $\lambda_A^{(t)}$=0.8, 0.7, 0.6 consistently outperforms the STL decoder baseline. This implies that the $\lambda_A^{(t)}$ should not be too high (less MTL) or too low (less ASR training) for optimal MTL robustness. We can also see this robustness in Figure 1 by comparing the STL decoder baseline with the MTL model for $\lambda_A^{(t)}$=0.7 and $\lambda_C^{(t)}$=$\lambda_C^{(i)}$=0.0. However, we see that decoder/discriminator MTL is still not able to beat CTC/decoder MTL when the CTC head is dropped for inference.

Therefore, we next study the combination of all three heads (CTC, decoder and discriminator) for performing MTL while dropping the CTC head during inference ($\lambda_C^{(i)}$=0.0). For these experiments, we first set $\lambda_A^{(t)}$=0.7 which shows superior adversarial robustness in Table 2. We then modulate $\lambda_C^{(t)}$ from values $\{0.0, 0.3, 0.5, 0.7\}$. Table 3 shows the adversarial performance results for this setting. Similar to Table 1, we can see that increasing MTL training weight for the CTC head but dropping the CTC head during inference improves the overall robustness, with $\lambda_C^{(t)}$=0.5 showing the best robustness of all MTL models. Hence, combining all three heads and performing MTL with semantically diverse tasks makes the model most resilient to the adversarial attacks consistently across multiple attack steps. This can also be clearly seen in Figure 1 where the MTL model corresponding to $\lambda_A^{(t)}$=0.7, $\lambda_C^{(t)}$=0.5 and $\lambda_C^{(i)}$=0.0 outperforms all other MTL combinations and baselines.

6. Conclusion

In this work, we study the impact of multi-task learning (MTL) on the adversarial robustness of ASR models. We perform extensive experimentation with multiple training and inference hyperparameters as well as wide-ranging adversarial settings. Our thorough empirical testing reveals that performing MTL with semantically diverse tasks consistently makes it harder for an attacker to succeed across several attack steps. We also expose the extreme vulnerability of the CTC loss, and discuss related pitfalls and remedies for using the CTC head during training with MTL. In future work, we aim to investigate regularization methods that can reduce the peaky behavior of CTC so as to induce more adversarially robust hybrid inference.

7. References

[1] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured prediction models," *arXiv preprint arXiv:1707.05373*, 2017.

[2] D. Iter, J. Huang, and M. Jermann, "Generating adversarial examples for speech recognition," *Stanford Technical Report*, 2017.

[3] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? adversarial examples against automatic speech recognition," *arXiv preprint arXiv:1801.00554*, 2018.

[4] 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.

[5] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "CommanderSong: A systematic approach for practical adversarial voice recognition," in *27th USENIX security symposium (USENIX security 18)*, 2018, pp. 49–64.

[6] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri, "Targeted adversarial examples for black box audio systems," in *2019 IEEE security and privacy workshops (SPW)*. IEEE, 2019, pp. 15–20.

[7] Q. Wang, B. Zheng, Q. Li, C. Shen, and Z. Ba, "Towards query-efficient adversarial attacks against automatic speech recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 896–908, 2020.

[8] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[9] C. Mao, A. Gupta, V. Nitin, B. Ray, S. Song, J. Yang, and C. Vondrick, "Multitask learning strengthens adversarial robustness," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 12347. Springer, 2020, pp. 158–174.

[10] S. Ghamizi, M. Cordy, M. Papadakis, and Y. L. Traon, "Adversarial robustness in multi-task learning: Promises and illusions," *arXiv preprint arXiv:2110.15053*, 2021.

[11] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[13] N. Das, M. Shanbhogue, S.-T. Chen, L. Chen, M. E. Kounavis, and D. H. Chau, "Adagio: Interactive experimentation with adversarial attack and defense for audio," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 677–681.

[14] S. Hussain, P. Neekhara, S. Dubnov, J. McAuley, and F. Koushanfar, "WaveGuard: Understanding and mitigating audio adversarial examples," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2273–2290.

[15] A. Sreeram, N. Mehlman, R. Peri, D. Knox, and S. Narayanan, "Perceptual-based deep-learning denoiser as a defense against adversarial attacks on asr systems," *arXiv preprint arXiv:2107.05222*, 2021.

[16] P. Żelasko, S. Joshi, Y. Shao, J. Villalba, J. Trmal, N. Dehak, and S. Khudanpur, "Adversarial attacks and defenses for speech recognition systems," *arXiv preprint arXiv:2103.17122*, 2021.

[17] C.-H. Yang, J. Qi, P.-Y. Chen, X. Ma, and C.-H. Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3107–3111.

[18] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning." in *Interspeech*, 2018, pp. 2454–2458.

[19] Y. Adi, N. Zeghidour, R. Collobert, N. Usunier, V. Liptchinsky, and G. Synnaeve, "To reverse the gradient or not: An empirical comparison of adversarial and multi-task learning in speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3742–3746.

[20] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, "Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning," in *Proc. Interspeech 2021*, 2021, pp. 1314–1318.

[21] L. Li, Y. Kang, Y. Shi, L. Kürzinger, T. Watzel, and G. Rigoll, "Adversarial joint training with self-attention mechanism for robust end-to-end speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–16, 2021.

[22] Y. Tang, J. Pino, C. Wang, X. Ma, and D. Genzel, "A general multi-task learning framework to leverage text data for speech to text tasks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6209–6213.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[25] "Mozilla Common Voice," https://commonvoice.mozilla.org.

[26] N. Kamo, "ESPnet2 pretrained model," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4604011

[27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[28] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[29] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards, "Adversarial robustness toolbox v1.2.0," *CoRR*, vol. 1807.01069, 2018. [Online]. Available: https://arxiv.org/pdf/1807.01069

[30] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4854–4858.

[31] H. Liu, S. Jin, and C. Zhang, "Connectionist temporal classification with maximum entropy regularization," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[32] A. Zeyer, R. Schlüter, and H. Ney, "Why does ctc result in peaky behavior?" *arXiv preprint arXiv:2105.14849*, 2021.