

Evaluating the Transferability of Adversarial Examples on Speech Models via Black-box Ensemble Attacks

Henry Bloom¹, Kathir Meyyappan¹

¹University of Chicago

hmbloom@uchicago.edu, kmeyyappan@uchicago.edu

Abstract

Adversarial attacks pose a significant threat to the robustness of automatic speech recognition (ASR) systems by introducing minimal input perturbations that cause misclassification. Studies such as [1] and [2] detail gradient-based attacks that succeed on ASR systems such as DeepSpeech [3] and are based on past literature attacking image classification models [4].

This study investigates the transferability of adversarial examples (AEs) in a black-box setting using ensemble attacks on ASR models. We use six DeepSpeech models across two architectures (v1 and v2 [5], available in PyTorch here), each trained on different speech datasets (LibriSpeech [6], AN4 [7], TEDLIUM [8]). For each version of DeepSpeech, we generate an adversarial attack on an ensemble of two models trained on two datasets and test its transferability to the model trained on the third dataset.

Using projected gradient descent with a constrained perturbation budget, we evaluate attack efficacy via edit distance and loss progression. We show that AEs fail to transfer even between models with identical architecture trained on different datasets. This implies that retraining with new data may enhance resilience in the case of model compromise. Our findings call for a broader evaluation using modern ASR architectures and diverse datasets to assess the transferability of ensemble-based adversarial attacks.

Index Terms: automatic speech recognition, adversarial machine learning, black-box attacks, machine learning security, gradient-based attacks

1. Introduction

ASR systems have become ubiquitous in modern society, and their accuracy and accessibility have improved significantly over time. However, as these systems have matured, researchers have successfully constructed adversarial attacks that exploit weaknesses in their underlying mechanisms.

The rise of adversarial machine learning has given birth to a discipline in speech recognition aimed at breaking ASR models: adversarial attacks add minimal perturbation to audio clips in order to make models misclassify their text content. For example, a *targeted* attack seeks to cause a specific misclassification, may perturb an audio clip that sounds like "hello world" to be recognized as "evil dot com" [9]. One of the foundational attacks is projected gradient descent (PGD), an algorithm designed to iteratively calculate the optimal perturbation to induce a misclassification by walking along the loss gradient on the target with respect to the input [4]. At each iteration, the perturbed input is clipped to be within a certain ϵ of the original input to keep the adversarial change imperceptible. A targeted projected gradient attack generally follows an algorithm similar to:

Algorithm 1 Targeted PGD

Require: Original input x , target label y_{target} , model f , step size α , iterations T , perturbation bound ϵ

- 1: $x_0 \leftarrow x$
- 2: **for** $t = 0$ to $T - 1$ **do**
- 3: $g_t \leftarrow \nabla_x \mathcal{L}(f(x_t), y_{\text{target}})$
- 4: $x_{t+1} \leftarrow \Pi_{\mathcal{B}_\epsilon(x)}(x_t - \alpha \cdot \text{sign}(g_t))$
- 5: **end for**
- 6: **return** x_T

Although the above algorithm was initially created for image classification systems, which use convolutional neural networks (CNNs), it has since been generalized and shown to work for speech recognition systems such as DeepSpeech, the model we investigate in this paper [1, 10]. As can be seen by the use of $\nabla_x \mathcal{L}(f(x_t), y_{\text{target}})$ here, we require white-box access to the model. That is, we utilize the architecture and weights of the model in order to store and calculate gradients to successfully compute an adversarial example. Black-box attacks refer to attacks that are conducted *without* access to model weights and typically involve approaches such as querying the target model or using surrogate models as white-box stand-ins.

Natural evolution strategies (NES), a popular black-box method, approximate the unknown target model gradient by repeatedly querying the model [11], perturbing the input with random noise, and estimating the loss gradient by observing accuracy changes. We initially attempted to run an NES attack; however, we were unable to find success due to the high dimensionality of the speech input.

Ensemble attacks are an alternate black-box adversarial attack strategy in which perturbations are computed against a set of white-box surrogate models and then applied to the unknown target. The goal is to produce *transferable* adversarial inputs which also work on the black-box target. The algorithm is very similar to the one detailed above for white-box PGD, except that it accounts for multiple models' loss when moving towards the target in the input space:

Algorithm 2 Targeted Ensemble PGD

Require: Original input x , target label y_{target} , models $f_1 \dots f_n$, weights $w_1 \dots w_n$, step size α , iterations T , perturbation bound ϵ

- 1: $x_0 \leftarrow x$
- 2: **for** $t = 0$ to $T - 1$ **do**
- 3: $g_t \leftarrow \sum_{i=1}^n w_i \cdot \nabla_x \mathcal{L}_i(f_i(x_t), y_{\text{target}})$
- 4: $x_{t+1} \leftarrow \Pi_{\mathcal{B}_\epsilon(x)}(x_t - \alpha \cdot \text{sign}(g_t))$
- 5: **end for**
- 6: **return** x_T

Here, the gradient is calculated on a weighted sum of the

ensemble models' losses. This allows us to conduct attacks on white-box models with the hope that our calculated adversarial example may transfer to a target model that we haven't seen.

Existing research has detailed the challenge of creating transferable ensemble attacks for ASR models [12]. Guo et al. demonstrated that transferability in the audio domain is markedly more difficult than in images due to the irregularity and instability of ASR model decision boundaries, as well as the critical role of data context in speech recognition. They further showed that applying vanilla ensemble methods does not guarantee success, but carefully designed strategies such as random gradient ensemble (RGE) or dynamic gradient weighting (DGWE) can moderately improve black-box transferability. Success can also be enhanced by taking advantage of noise injection, dropout, and scale-invariance assumptions. Notably, the authors only consider a singular target sentence, "turn off the light", when constructing adversarial inputs and evaluating transferability.

In this work, we investigate the transferability of ensemble attacks where the target model has the same architecture as the models in the ensemble, yet is trained on a different dataset. By testing ensemble transferability between identical models with different training sets, we hope to investigate whether the challenge of transferability remains for this specific adversarial situation. If transferability still proves difficult, this would suggest a potential easy way to slow down adversaries who may have access to a previously compromised speech model by retraining models on new data. We seek to thoroughly test these results across a variety of input waveforms and target sentences to increase the robustness of our results.

2. Methodology

Our experimental framework evaluates the transferability of AEs across speech recognition models using black-box ensemble attacks. We simulate a scenario where an adversary has white-box access to a limited ensemble of models with the same architecture yet trained on different datasets and attempts to attack a held-out target model with the same model architecture, without direct access to its gradients or parameters.

2.1. Model Architecture and Dataset Configuration

We use Mozilla's DeepSpeech models v1 and v2, representing two different ASR architectures. Each version is trained on three distinct speech datasets:

- **LibriSpeech** – a large corpus of read English audio-books [6],
- **AN4** – a small command-and-control dataset [7],
- **TEDLIUM** – a corpus of TED talk transcriptions [8].

In total, we used six models: v1+LibriSpeech, v1+AN4, v1+TEDLIUM, v2+LibriSpeech, v2+AN4, and v2+TEDLIUM. For each DeepSpeech version, we designate two of the three models as ensemble models and the third as the target. Each permutation allows us to isolate the effects of dataset variation while controlling for architecture. The ensemble-target pairs look like this:

Target Model	Ensemble Models
v1+LibriSpeech	v1+AN4, v1+TEDLIUM
v1+AN4	v1+LibriSpeech, v1+TEDLIUM
v1+TEDLIUM	v1+LibriSpeech, v1+AN4
v2+LibriSpeech	v2+AN4, v2+TEDLIUM
v2+AN4	v2+LibriSpeech, v2+TEDLIUM
v2+TEDLIUM	v2+LibriSpeech, v2+AN4

Table 1: *Ensemble-target model pairings for each attack. Each ensemble contains two models with the same architecture as the target model that were trained on different datasets.*

By testing with every ensemble-target model combination, we seek to reduce potential biases in our results due to the inherent differences between the models trained on each dataset.

2.2. Adversarial Attack Setup

We generate targeted adversarial examples using PGD, which optimizes perturbations to drive the model's output toward a specific target sentence. For the black-box ensemble attack:

- PGD is applied over a weighted sum of gradients from two white-box surrogate models. We weighted each model in our ensemble equally after finding that their losses were of comparable magnitude.
- The objective is to minimize loss with respect to a pre-defined target transcription.
- The loss function used is the Connectionist Temporal Classification loss, which is standard for ASR models.

We constrain the perturbation under an L_∞ norm bound of 1.5% of the input signal amplitude (range of 0.03 centered around 0), and run PGD for 300 iterations with a fixed step size ($\alpha = 10^{-3}$).

2.3. Sampling and Target Selection

We curated a set of ten target sentences from the "Harvard sentences," a collection of phonetically balanced sentences commonly used for speech model testing [13] and five input waveforms from Librispeech. For each ensemble-target model configuration, we sample 20 pairs of target sentences and input waveform, create an adversarial input by running PGD on the two ensemble models, and test the transferability on the target model. This design ensures a reproducible evaluation of attack transferability that accounts for the large variance in typical audio and target sentences.

2.4. Evaluation Metrics

We assess attack success using the following metrics:

- **Levenshtein (Edit) Distance.** Levenshtein distance measures dissimilarity between the model's output and the target sentence by calculating the minimum number of 1-character changes needed to match the two. For example, the strings "kitten" and "sitting" have a Levenshtein distance of three. Our goal when attacking a model is to reduce the edit distance as much as possible.
- **Loss Progression.** We track the CTC loss over the course of PGD iterations for our ensemble and target models. The CTC loss is correlated with the final Levenshtein distance, and is what we are directly optimizing during PGD.

CTC-based models output character-wise, often in the form of phonetics, so we use Levenshtein distance instead of Word Error Rate (WER) as we would expect the WER of our model outputs to be incredibly poor. These metrics are recorded over each iteration of PGD for every attack on all ensemble-target model specifications. Successful transfer is indicated by reduced edit distance.

3. Results

3.1. Transferability Across Identical Architectures

We found that adversarial examples generated on two models sharing the same architecture but trained on different datasets did *not* reliably transfer to a third model with the same architecture. Despite the perturbations consistently misleading the ensemble models and achieving low edit distances, the target models’ transcriptions remained largely unaffected and yielded high edit distances. This result suggests that data differences alone are sufficient to break transferability in naive PGD settings, even when ensemble models share the same architecture as the target model.

3.2. An Attack Example

For each of the 120 attacks we carried out across our 6 target models, we tracked the average CTC loss and edit distance over PGD iterations for both the ensemble and target models. Figure 1 shows one such attack’s edit distance progression across PGD iterations for the computed AE:

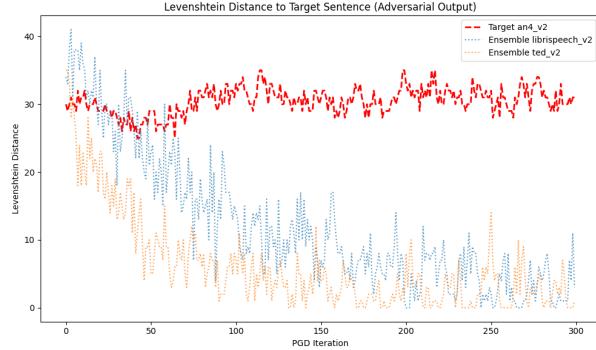


Figure 1: AE output closeness to target sentence for ensemble models v2+TEDLIUM, v2+LibriSpeech, and target model v2+TEDLIUM.

Clearly, the computed AE managed to achieve near-perfect edit distance on the two ensemble models, yet failed to reliably guide the prediction of the target model towards the target.

3.3. Aggregate Evaluation

To summarize performance across all attack scenarios, we plotted the distribution of the final edit distances between the target models’ transcriptions of the final adversarial input and the desired transcription for each attack:

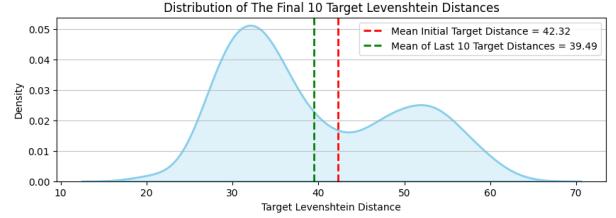


Figure 2: Distribution of the final ten Levenshtein distances of each attack’s adversarial example output in its target model.

As can be seen in Figure 2, edit distances remained extremely high after 300 iterations, signaling a failure to successfully fool the target models into transcribing the audio as the target sentences. While there was a minor numerical improvement (mean edit distance reduced from 42.32 to 39.49), the change was insufficient to be considered a successful attack.

Figures 3 and 4 showcase the mean progression of the Levenshtein distance and loss, respectively, of the target and ensemble models with respect to the adversarial input at that iteration of PGD.

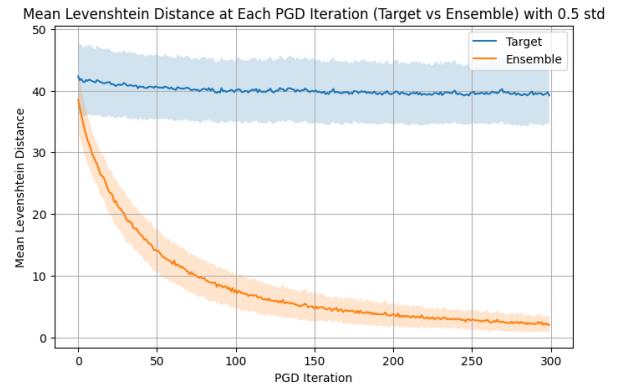


Figure 3: The average edit distance between the target models’ transcription of the adversarial input and the desired label for both ensemble and target models across all attacks. A half standard deviation of data is also included to showcase the relative consistency of the results.

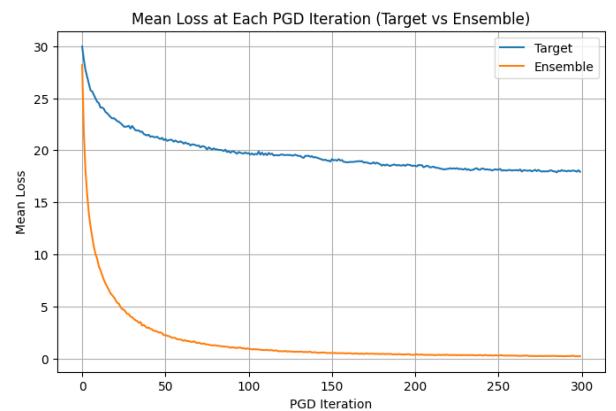


Figure 4: The average loss for ensemble models and target models across all attacks.

As expected, the ensemble models generally showed a steady decrease in both metrics over the course of the attack, indicating the high success rate and convergence of our PGD attack on the white-box models. In contrast, the loss and edit distance calculated on the target models showed little to no improvement throughout the attacks, revealing the adversarial examples' limited transferability. Testing with more PGD iterations did not improve performance.

Both the mean loss and edit distance curves illustrate the divergence between ensemble and target behavior, providing additional evidence that adversarial perturbations may be overfit to the training distributions of the ensemble models, leaving them unable to transfer to the target model.

3.4. Larger Ensembles

We investigated whether a larger ensemble including models with the other DeepSpeech version may improve transferability by avoiding local optima in our adversarial input. Notably, the ensembles contained a model which had been trained on the same dataset as the target model, yet was a different version of DeepSpeech. However, we found a similar lack of transferability in this setting. In Figure 5, we plot the Levenshtein distance for an adversarial attack on target model v2+LibriSpeech using an ensemble consisting of the rest of the models.

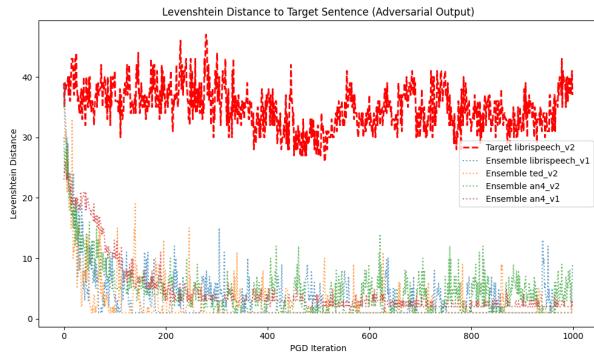


Figure 5: AE result's closeness to target sentence for target model (v2+LibriSpeech) and ensemble models (other 5) over PGD iterations

Clearly, despite the ensemble containing models with the same architecture (v2+AN4, v2+TEDLIUM) and training data (v1+LibriSpeech) as the target model, the AE still does not transfer.

4. Conclusion

4.1. Interpreting Results

Our results indicate the infeasibility of transferring AEs computed through naive PGD-based ensemble attacks even when using an ensemble of models with identical architecture to the target model. Despite prior research showcasing the challenge of creating transferable AEs for ASR models, this is a surprising result given the belief that model decision boundaries will be much more similar for models with the same architecture [14].

The low success rate of these ensemble attacks suggests that adversarial vulnerabilities in ASR systems are highly dependent on the model's training data distribution. Differences in learned

representations appear sufficient to disrupt the generalization of adversarial perturbations, even when the ensemble and target models have identical architecture. This is in contrast to findings in the image domain, where adversarial examples often generalize between similar models trained on similar data and ensemble attacks are a popular black-box approach [14]. We posit that ASR models may encode input features in a more data-specific manner, potentially due to the structured nature of speech, which combines acoustic and linguistic information in a tightly coupled way. The inputs to these models are extremely high-dimensional, more so than in the image domain, which may contribute to the complexity of model decision boundaries and prevent transferability.

The initial decrease in average CTC-loss for target models could be attributed to the shared feature space of architecturally identical models; initial PGD iterations would calculate a gradient that moves the computed AE broadly in the direction of the desired target though the specificities of each models' decision boundary soon prevents transferability. This early progress may reflect high-level similarities in low-layer representations, such as acoustic features or phoneme alignment strategies, but the lack of convergence suggests that deeper or later-stage model behavior diverges as a function of training data. In other words, even if initial perturbation steps overlap in effect, the optimization path quickly becomes tailored to the ensemble's learned mapping, and that no longer benefits the held-out model.

4.2. Discussion

These findings have important implications for the robustness of ASR systems. Specifically, they suggest that adversarial vulnerability may be highly data-dependent, and that retraining compromised models on new data could serve as a lightweight defense mechanism. For example, if a proprietary speech model with previously unknown weights were to be compromised, an attacker could easily run PGD on it to create AE's for the model. With our results, it follows that it may be sufficient to simply retrain the model on another large dataset to deter the adversary rather than rebuilding the model architecture to prevent transferability.

However, further testing with modern ASR architectures and a broader range of training corpora is necessary to assess whether these observations generalize beyond DeepSpeech.

5. Future Work

Future research should incorporate more advanced ensemble attack methods such as RGE and DGWE in the setting of same-architecture ensemble and target models to compare transferability to the general case where the ensemble and target models have different architectures. These methods have been shown to enhance transferability when the ensemble models are distinct from the target models [12]. Further, use of other optimization methods, such as Adam, when executing PGD on our ensemble models should be investigated.

In addition, testing with architecturally identical models that were trained on the same dataset yet fine-tuned with different data is a natural next step. If ensemble attacks reveal a similar lack of transferability, then fine-tuning a compromised model may serve as an efficient and sufficient defense from transferring adversarial examples.

6. References

- [1] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” 2018. [Online]. Available: <https://arxiv.org/abs/1801.01944>
- [2] P. Zelasko, S. Joshi, Y. Shao, J. Villalba, J. Trmal, N. Dehak, and S. Khudanpur, “Adversarial attacks and defenses for speech recognition systems,” no. arXiv:2103.17122, Mar. 2021, arXiv:2103.17122. [Online]. Available: <http://arxiv.org/abs/2103.17122>
- [3] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, “Deep speech: Scaling up end-to-end speech recognition,” no. arXiv:1412.5567, Dec. 2014, arXiv:1412.5567. [Online]. Available: <http://arxiv.org/abs/1412.5567>
- [4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” no. arXiv:1706.06083, Sep. 2019, arXiv:1706.06083. [Online]. Available: <http://arxiv.org/abs/1706.06083>
- [5] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, “Deep speech 2: End-to-end speech recognition in english and mandarin,” no. arXiv:1512.02595, Dec. 2015, arXiv:1512.02595. [Online]. Available: <http://arxiv.org/abs/1512.02595>
- [6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, p. 5206–5210. [Online]. Available: <https://ieeexplore.ieee.org/document/7178964>
- [7] A. Acero and R. Stern, “Environmental robustness in automatic speech recognition,” in *International Conference on Acoustics, Speech, and Signal Processing*, Apr. 1990, p. 849–852 vol.2. [Online]. Available: <https://ieeexplore.ieee.org/document/115971>
- [8] A. Rousseau, P. Deléglise, and Y. Estève, “Ted-lium: an automatic speech recognition dedicated corpus,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association (ELRA), May 2012, p. 125–129. [Online]. Available: <https://aclanthology.org/L12-1405/>
- [9] N. Carlini. [Online]. Available: https://nicholas.carlini.com/code/audio_adversarial_examples/
- [10] N. Das and D. H. Chau, “Hear no evil: Towards adversarial robustness of automatic speech recognition via multi-task learning,” no. arXiv:2204.02381, Apr. 2022, arXiv:2204.02381. [Online]. Available: <http://arxiv.org/abs/2204.02381>
- [11] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” 2018. [Online]. Available: <https://arxiv.org/abs/1804.08598>
- [12] F. Guo, Z. Sun, Y. Chen, and L. Ju, “Towards the transferable audio adversarial attack via ensemble methods,” no. arXiv:2304.08811, Apr. 2023, arXiv:2304.08811. [Online]. Available: <http://arxiv.org/abs/2304.08811>
- [13] IEEE No 297-1969, p. 1–24, Jun. 1969. [Online]. Available: <https://ieeexplore.ieee.org/document/7405210>
- [14] Y. Liu, X. Chen, C. Liu, and D. Song, “Delving into transferable adversarial examples and black-box attacks,” no. arXiv:1611.02770, Feb. 2017, arXiv:1611.02770. [Online]. Available: <http://arxiv.org/abs/1611.02770>