ON ADVERSARIAL ROBUSTNESS OF LARGE-SCALE AUDIO VISUAL LEARNING

Juncheng B Li, Shuhui Qu*, Xinjian Li, Po-Yao (Bernie) Huang, Florian Metze*

Carnegie Mellon University

ABSTRACT

As audio-visual systems are being deployed for safety-critical tasks such as surveillance and malicious content filtering, their robustness remains an under-studied area. Existing published work on robustness either does not scale to large-scale dataset, or does not deal with multiple modalities. This work aims to study several key questions related to multi-modal learning through the lens of robustness: 1) Are multi-modal models necessarily more robust than uni-modal models? 2) How to efficiently measure the robustness of multi-modal learning? 3) How to fuse different modalities to achieve a more robust multi-modal model? To understand the robustness of the multi-modal model in a large-scale setting, we propose a density-based metric, and a convexity metric to efficiently measure the distribution of each modality in high-dimensional latent space. Our work provides a theoretical intuition together with empirical evidence showing how multi-modal fusion affects adversarial robustness through these metrics. We further devise a mix-up strategy based on our metrics to improve the robustness of the trained model. Our experiments on AudioSet [1] and Kinetics-Sounds [2] verify our hypothesis that multi-modal models are not necessarily more robust than their uni-modal counterparts in the face of adversarial examples. We also observe our mix-up trained method could achieve as much protection as traditional adversarial training, offering a computationally cheap alternative.

1. INTRODUCTION

Nowadays, uploading a clip of video or audio to social media platforms such as Facebook or YouTube will trigger a multi-modal content filtering algorithm to proactively search for potentially policy-violating content; home monitoring devices such as NestCams or RingCams are presumably using audio-visual models to identify events in the monitored area. Multi-modal classification in safety critical, audio-visual tasks therefore calls for a thorough understanding of its robustness, besides its accuracy.

Many recent researches have documented neural network models could be vulnerable to adversarial attacks, manipulations of the input to a classifier specifically crafted to be inconspicuous to humans, but which cause the classifier to predict incorrectly [3, 4]. Concerns about potential adversarial examples have sparked a huge interest in the research community to study how can we train robust models that defend against potential perturbations [4, 5]. However, building such adversarially robust models is challenging [5]. A smaller but still substantial line of work has emerged to show that we could have formal verification of the robustness of neural network models [6, 7]. However, such methods are subjected to very tight constraints and are notoriously difficult to scale. So far, despite some successful large-scale empirical evaluations [8, 9] on image-only datasets, robustness of multi-modal learning has not been fully understood. As illustrated in Fig. 1, the major challenge to analyzing multi-modal models is the high non-convexity and non-linearity of the decision boundaries in high dimensional latent spaces.

In this work, we discuss robustness of multi-modal neural network models for classification tasks in the large-scale setting. We first measure *point-wise robustness* through the empirical maximum allowable perturbation in ℓ_p norm. Based on *point-wise robustness*, we show there exist counter examples to the general claim that multi-modal models are more robust compare to the uni-modal models. Due to the limitation of point-wise robustness in terms of scalability and generalizability, class-wise robustness is a necessary and practical compliment to tackle large-scale multi-modal robustness problems. Instead of measuring the accuracy drop caused by running universal adversarial perturbation in different magnitude and ℓ_p norm [3], we define the class-wise metric by using 1) the density of samples within the high-dimensional ball centered at the centroid of each class with a certain ℓ_p radius; 2) the convexity of samples in the high dimensional latent space. We evaluate our metric on the AudioSet [1] and Kinetic-Sounds dataset [2]. The results indicate that multi-modal models are only more robust measured by class-wise metrics for a limited number of classes. We also observe the point-wise robustness of classification results vary greatly depending on the variance of the data with specific class label.

Inspired by our observations, we propose a density-convexity-based mix-up fusion technique to smooth the decision boundary and add robustness to the fused model. Our proposed mixup could both improve class-wise robustness and point-wise robustness upon our baseline fusion model while increasing the accuracy compared to vanilla fusion methods. We also compare it to traditional adversarial training, where we also see competitive robust accuracy. These advances allow us to address adversarial robustness in large-scale multi-modal settings for the first time.

2. RELATED WORK

Previous works such as [7, 6, 10] focused on *point-wise* robustness, studying the maximum allowable radius of centered Chebyshev ball: *a ℓ_p ball centered at an input point, within which the output class of a given neural network with remains unchanged, treating the decision boundary of the model as a convex or non-convex polytope.* This formulation certainly provide a safe threshold to defend against adversarial attacks, whereas it involves expensive iterative computations on each anchor point, resulting in huge difficulty to scale [7, 6], most of them depend their claims on small-scale datasets like MNIST or CIFAR-10. For the large-scale multimodal datasets such as AudioSet, such a verification would hardly be feasible.

Some recent works are try to study defence methods in large scale, including adversarial training [9, 11], randomization [8], and model ensemble [12]. Most of them measured robustness by point-wise accuracy or attack success rate for specific attack budget ϵ and number of iterations. However, these metrics are often too general to reflect the classifier's behavior under the influence of adversarial perturbation. This motivates us to look into both class-wise accuracy changes along with point-wise accuracy change in order to have

* equal contribution

a better chance of understanding potential reasons of label change caused by adversarial perturbations.

Audio-visual learning [13] itself is more complicated than image classification, and the current focus still seems to be improving accuracy [14]. The robustness of multimodal classification models involving large-scale video-audio dataset has never been studied rigorously. [10] considered the robustness of deep neural networks on videos and experimented on UCF101 dataset [15] which is a relatively small dataset. They measured robustness by the maximum safe radius (point-wise), which computes the minimum distance from the optical flow sequence obtained from a given input to that of an adversarial example in the neighbourhood of the input. To our knowledge, our work is the first to comprehensively study the robustness of multi-modals models both consist of video and audio against adversarial perturbation cause changes to both modalities.

3. BACKGROUND

3.1. Universal Adversarial Perturbations

The problem of computing universal adversarial perturbation to attack a classification model f by maximizing the following [9]:

$$\begin{aligned} &\mathbf{E}_{(x,y)\sim\mathcal{D}} \max_{x'\in C(x)} [L(f(x'), y)] \\ &\text{subject to } C(x) = \{a \in \mathbb{R} : ||a - x||_p \leq \epsilon\}. \end{aligned} \tag{1}$$

where L is the loss function, x is input and y is label, $\mathcal{D}$ is the dataset, and $x' = x + \delta$ is our perturbed input. We want to find some perturbation x' that looks "indistinguishable" from x, yet is classified incorrectly by f even when x is classified correctly.

To solve such a constrained optimization problem, one of the most common methods utilized to circumvent the non-exact-solution issue is the Projected Gradient Descent (PGD) method [9]:

$$\delta := \mathcal{P}_\epsilon \left(\delta - \alpha \frac{\nabla_\delta L(f(x+\delta), y)}{\|\nabla_\delta L(f(x+\delta), y)\|_p} \right) \tag{2}$$

where $\mathcal{P}_\epsilon$ is the projection onto the ℓ_p ball of radius ϵ, and α is the gradient step size.

3.2. Multi-Modal Adversarial Perturbations

Under audio-visual multimodal learning setting, we formulate our loss as $L_{multi} = L(f(g(x_A) \oplus h(x_V)), y)$, over the classification function f, which can however readily be expanded to additional modalities. $g(x_A)$ denotes the encoding of audio features into a bottleneck representation with audio encoding networks (CSN) depicted in Figure 2, while $h(x_V)$ similarly represents the (R2+1D)CNN [16] encoded video representation, and $\oplus$ indicates concatenation of features. $\mathcal{D}_A$ and $\mathcal{D}_V$ indicating their individual dataset. This more complicated setting requires us to study the adversarial perturbation computed against both the audio input δ_A and the video input δ_V, and can be written out as:

$$\begin{aligned} &\mathbf{E}_{(x_A,y)\sim\mathcal{D}_\mathcal{A};(x_V,y)\sim\mathcal{D}_\mathcal{V}} \max_{\delta_A\in C(x_A),\delta_V\in C(x_V)} [L(f(x'), y)] \\ &\text{subject to } C(x) = \{a \in \mathbb{R} : ||a - x||_p \leq \epsilon\} \end{aligned} \tag{3}$$

Here, $x' = g(x_A + \delta_A) \oplus h(x_V + \delta_V)$. The set $C(x)$ is usually defined as a ball of small radius of the perturbation size ϵ (of either ℓ_∞, ℓ_2 or ℓ_1) around x. Thus, to compute uni-modal audio perturbation to attack the multimodal model, our PGD step could be rewritten out as:

$$\delta_A := \mathcal{P}_\epsilon \left(\delta_A - \alpha \frac{\nabla_{\delta_A} L(f(g(x_A + \delta_A) \oplus h(x_V)), y)}{\|\nabla_{\delta_A} L(f(g(x_A + \delta_A) \oplus h(x_V)), y)\|_p} \right) \tag{4}$$

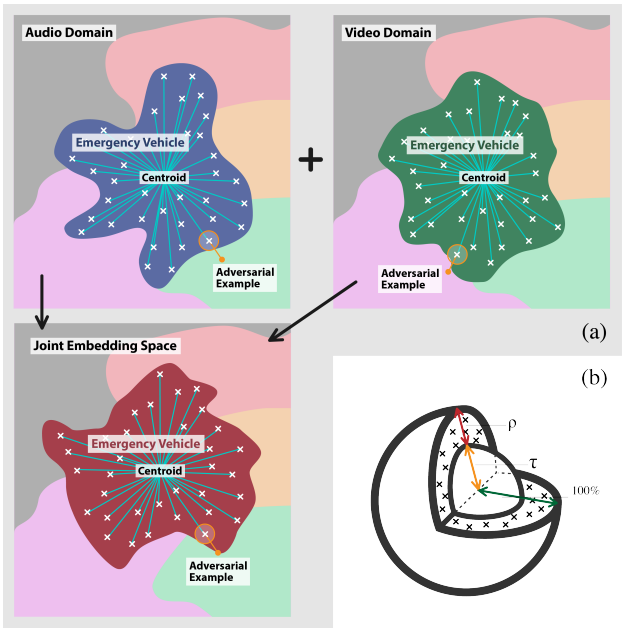

Fig. 1: (a) An illustration of multi-modal fusion. (b) Illustration of the centroid based density metric $\rho_c^{R_\tau, p}$.

Accordingly, to compute video perturbation against video classifier $g(x)$, we perform the following PGD step:

$$\delta_V := \mathcal{P}_\epsilon \left(\delta_V - \alpha \frac{\nabla_{\delta_V} L(f(h(x_V + \delta_V) \oplus g(x_A)), y)}{\|\nabla_{\delta_V} L(f(h(x_V + \delta_V) \oplus g(x_A)), y)\|_p} \right) \tag{5}$$

In the more complicated multi-modal case, we jointly optimize δ_A and δ_V, where:

$$\delta_A, \delta_V := \mathcal{P}_\epsilon(\delta_{(V,A)} - \alpha \frac{\nabla_{\delta_{(V,A)}} L(f(h(x_V + \delta_V) \oplus g(x_A + \delta_A)), y)}{\|\nabla_{\delta_{(V,A)}} L(f(h(x_V + \delta_V) \oplus g(x_A + \delta_A)), y)\|_p}) \tag{6}$$

4. CHALLENGE COMMON ASSUMPTIONS IN MULTIMODAL LEARNING

There is a vague notion that multimodal systems are generally more robust compared to unimodal models [13]: "having access to multiple modalities that observe the same phenomenon may allow for more robust predictions." From an information retrieval perspective, this statement is theoretically true since the same information was captured twice in different modality, improving the robustness of multimodal models. However, this is not always true if we rigorously consider how adversarial perturbations affect the neural network model as follows.

Theorem 1 *There exists a sample $x_i \in \mathcal{D}$, and a unimodal sample-wise attack $\exists ||\delta_{A,i}||_p \leq \epsilon_A$ or $\exists ||\delta_{V,i}||_p \leq \epsilon_V$ that can break a multimodal fusion network $f((x_{V,i} \oplus x_{A,i}), y_i)$, changing its prediction label y_i.*

Here, $\mathcal{D}$ is the dataset, and ϵ_A and ϵ_V are the point-wise robustness threshold for each uni-modal of sample x_i. Therefore, as a conjecture, a unimodal attack can break a multimodal model, which we empirically verified the existence of such cases in our experiments. The proof of Theorem 1 can be found at [17].

5. METRICS FOR CLASS-WISE ROBUSTNESS

Point-wise robustness is limited in terms of scalability and generalizability. Class-wise robustness metric is a more efficient for a large-scale dataset. Instead of exhaustively running universal adversarial perturbation we define two metrics to capture the main robustness property of each class.

5.1. Centroid-based density metric

We calculate the class-wise density of the class's high dimensional l_p norm ball by a function of number of samples n_c in the class c and the volume of the l_p norm ball. In this work, n_c is the number of samples of each class in Audioset.

The centroid of a class $\odot_c$ is the mean of bottleneck features $l = g(x)$ of samples x in the class c: $\odot_c = \frac{\sum_{i=1}^{n_c} l_{i,c}}{n_c}$, where n_c is the number of samples in the class c. In our case, it is $l = g(x_A)$ for audio modality or $l = g(x_V)$ for video modality. For each class, we calculate the distance of samples in c to the centroid $\odot_c$. The radius of a class $R_{p,c}$ on l_p norm ball is the maximum distance of all samples in c to the centroid $\odot_c$: $R_{p,c} = \max\|l_{i,c} - \odot_c\|_{p,i=1,...,n_c}$. The radius of first τ percentage of samples closing to the centroid is $R_{\tau,p,c}$, and the $n_{\tau,c} = \tau \times n_c$ is the number of τ percentage samples close to the centroid. According to [18], the volume $V_d^p(R)$ of the d-dimensional l_p norm ball with a radius R is: $V_d^p(R) = \frac{(2\Gamma(\frac{1}{p}+1)R)^2}{\Gamma(\frac{d}{p}+1)}$, where, d is the dimension of the ball, R is the radius, p is the ℓ_p-norm, and Γ is the Gamma function[1]. Now, we formally define the robustness of a class c with regard to τ quantile of the class sample x_c's distance to the centroid $\odot_c$ of the class c by:

$$\rho_c^{R_\tau,p,c} = \frac{n_c - n_{\tau,c}}{log(V_d^p(R_{p,c})) - log(V_d^p(R_{\tau,p,c}))} \tag{7}$$

where the numerator is the number of class samples whose l_p distance to centroid larger than τ quantile of samples in c; $R_{\tau,p,c}$ is the τ quantile of all class sample's l_p distance to the class's centroid. We perform the log operation on the volume to reduce the scale of Γ function for the ease of computation. This can be intuitively interpreted as the density in the outer crust of a ball as is shown in Fig. 1(b). Generally, the higher the density of the crust, the more robust the samples within/below the crust are.

5.2. Convexity-based metric

The convex set C in geometry is defined as a set where given any two points $x_1, x_2 \in C$ in the set, the set contains the whole line segment $x = \theta x_1 + (1-\theta)x_2$, with $0 \leq \theta \leq 1$. Based on our observations and conjecture, we propose the convexity-based metric as one of the robustness measurement of the class. We construct the convex set $S = \{\hat{x}_s | \hat{x}_s = \theta x_1 + (1-\theta)x_2, \theta \sim U[0,1], \forall x_1, x_2 \in C\}$, and sample n points from it $\{\hat{x}_1, ..., \hat{x}_n | \hat{x}_i \in S\}$. The metric is as follows:

$$\kappa_c = \frac{\sum_{i=1}^{n} \mathbb{1}\{f(\hat{x}_i) = c\}}{n} \tag{8}$$

where y_c is the class label. The higher the κ_c is, the more convex the decision boundary of class c is. In this work, we set $n = 2000$. Therefore, we use this metric as a proxy to measure how convex the neural network is. We hope to see positive correlation between convexity and robustness.

For both metrics, we use the bottleneck feature l to calculate the value of the metric. In later experiments, we empirically show the effectiveness of our metrics by constrasting with the accuracy drop caused by universal perturbation [3].

[1] https://wikipedia.org/wiki/Gamma_function

6. DENSITY-CONVEXITY BASED MIX-UP

As is noted by [19], mix-up techniques could potentially smoothen the decision boundary via generating virtual training samples by weighted sum of existing training samples, which improves generalizability. Inspired by our density-based and convexity-based metric, we employ a simple adjustment to mixup:

$$\begin{aligned} \tilde{x}_A &= \alpha x_{Ai} + (1-\alpha)x_{Aj}; \\ \tilde{x}_V &= \alpha x_{Vi} + (1-\alpha)x_{Vj}; \\ \tilde{y} &= \alpha y_i + (1-\alpha)y_j; \end{aligned} \tag{9}$$

where $\alpha \in [0,1]$,(x_{Ai}, x_{Vi}, y_i) and (x_{Aj}, x_{Vj}, y_j) are two training samples, with both audio and video inputs drawn from 2 different classes y_i and y_j, subject to $\kappa_c < T$ (T is an empirical threshold on the convexity) and $\rho_c^{R_\tau,p} > D$ (D is an empirical threshold on the density), for both classes. Both T and D are dataset dependent parameters, $T = 0.5, D = 8$ in this work. Effectively, we are augmenting the less convex classes of training data with more samples from the "denser" samples which are closer to the center of its feature space.

7. EXPERIMENTS

7.1. Dataset and Model Setup

AudioSet [1] contains 2 Million 10-second YouTube video clips, summing up to 5,800 hours annotated with 527 types of sound events. We train and test the models according to the train and test split described in [20]. The input for the audio branch are matrices of Mel filter bank features. For the visual branch, we employ a (R2+1D)CNN + transformer backbone [16] to encode the spatial-temporal context. The clean performance (with no data augmentation) of our unimodal audio model and audio-visual model are listed in Table 1 (italic font). *Kinetics-Sounds* [2] is a subset of Kinetics [21] that contains 34 classes of audio-related events (22,107 train, 1,504 validation). We preprocess the Kinetic-sound dataset in the similar fashion. Our clean multimodal basline: *86.5%*. We use Kinetic-sounds only for audio-visual performance since its audio-unimodal performance is low and thus not representative.

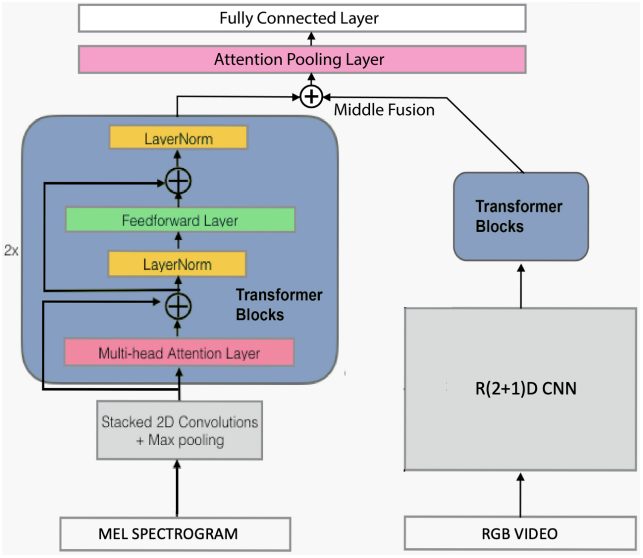

Fig. 2: The overall architecture, the audio branch (left) uses Convolution self-attention architecture, video branch is on the right. Mid fusion involves the concatenation step described in §3.2.

Models	Attack	mAP	AUC	d-prime
Audio UniModal (PANNS) [23]	No	0.383	0.963	2.521
Audio UniModal	Yes	0.183	0.895	1.770
Mid Fusion (G-blend) [14]	No	**0.427**	0.971	2.686
Mid Fusion	Yes A+V	0.182	0.889	1.836
Mid Fusion	Yes V-only	0.339	0.954	2.441
Mid Fusion	Yes A-only	0.310	0.940	2.276
Mid Fusion mixup	No	0.424	0.972	2.711
Mid Fusion mixup	Yes A+V	0.234	0.891	1.983
Mid Fusion *AT*	No	0.397	0.964	2.530
Mid Fusion *AT*	Yes A+V	0.199	0.900	1.861

Table 1: Performance of our best performing CSN models on AudioSet, and their performance against the adversarial perturbation, using the overall architecture shown in Figure 2. Here, mAP is the mean average precision, AUC is the area under the false positive rate and true positive rate (recall). The d-prime can be calculated from AUC [1]. AT denotes adversarial training. A red text color indicates the most potent perturbation against the model.

Models	Attack	Acc	mAP	AUC
Mid Fusion	No	**86.5%**	0.853	0.987
Mid Fusion	Yes A+V	5.0%	0.513	0.824
Mid Fusion	Yes V-only	6.6%	0.541	0.848
Mid Fusion	Yes A-only	85.6%	0.851	0.987
Mid Fusion mixup	No	85.5%	0.854	0.987
Mid Fusion mixup	Yes A+V	15.2%	0.734	0.623

Table 2: Performance of CSN models on Kinetics-Sounds.

7.2. Adversarial Perturbation & Adversarial Training

All of the adversarial perturbations in this work are computed with the PGD method [9], with $\epsilon = 0.1$, ℓ_2, trained over 20 iterations. Attack budget (ϵ and # of iteration) are constrained to be the same across unimodal, multimodal experiments. Adversarial Training is performed by the methods mentioned in [22].

7.3. Results and Discussion

As we can see from Tables 1 and 2, we find empirical evidence supporting Theorem 1, as uni-modal attacks (Video-only, Audio-only) can successfully break the multimodal network. Multimodal attacks are however more potent than unimodal attacks given the same attack budget, with the same ϵ and #iterations causing more accuracy drop compared to a unimodal attack.

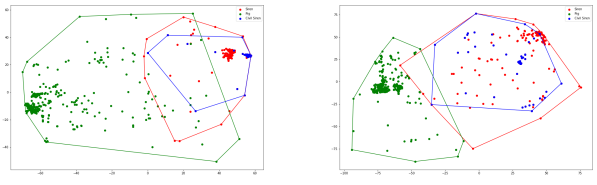

Fig. 3: Convex hull of samples. Left: TSNE of the vanilla fusion feature. Right: TSNE of the mixup fusion feature. Three classes: Red → Siren; Blue → Civil Siren (subclass of Siren); Green → Pig.

We could also see the benefit of our proposed mixup, which could help minimize the accuracy loss significantly on both AudioSet and Kinetics-Sounds, compared to vanilla mid fusion and traditional adversarial training (AT). Interestingly, among AudioSet where audio event classification should be the dominant task, audio modality showed on-par robustness compared to video. While in Kinetics-Sounds dataset where video event classification is the dominant task, audio attacks did not affect the classifier at all, video modality dominantly decides the robustness.

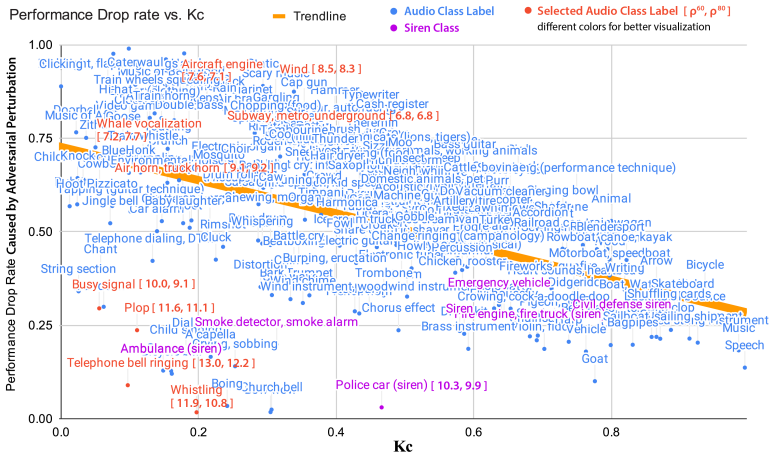

Fig. 4: Performance Drop Rate% VS Convexity (κ_c) for all audio classes in AudioSet. Density $\rho_c^{R,p}$ partially shown due to space limit.

In Fig. 4, Performance Drop Rate:= $\frac{\text{clean performance} - \text{performance under attack}}{\text{clean performance}}$. The density $\rho_c^{R,p}$ (see §5.1) is measured at 60 and 80 quantile of the class in ℓ_2 norm. We can observe a negative correlation between drop rate and the convexity of the class, suggesting a positive correlation between robustness and convexity. We could also see higher density would correlate to more robustness for an audio class, some denser classes tend to be more robust despite having low convexity. Intriguingly, we see the pattern that simpler waveforms such as siren (highlighted), bell sounds, child crying are more robust than complex signals like traffic/mechanics sounds. Our empirical findings verify our conjecture that both the density and the convexity of the data are simple factors that could severely affect its robustness. Fig. 3 shows the convex hull of bottleneck feature of mixup fusion model and vanilla mid fusion model, where we observe that the mixup model clearly pushes samples of a class to form a denser outer crust and encourages a clearer boundary between classes resulting in a higher convexity within selected classes.

8. CONCLUSION

In contrast to some common notion, our work shows that multimodal systems are not always more robust to adversarial attacks. In this work, we propose alternative metrics to understand the adversarial robustness of large-scale multi-modal models, and we empirically show the effectiveness of our density and convexity metric. We further propose a novel multi-modal mix-up method that selectively augments the denser samples in less convex classes to compensate for the robustness, and which outperforms simple adversarial training with respect to robustness. Our experiments on AudioSet [1] and Kinetic-Sounds dataset [2] verify our hypothesis and the effectiveness of the mix-up strategy.

9. ACKNOWLEDGMENTS

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center (PSC).

10. REFERENCES

[1] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[2] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[3] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[5] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry, "Robustness may be at odds with accuracy," *arXiv preprint arXiv:1805.12152*, 2018.

[6] Eric Wong and Zico Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5286–5295.

[7] Matt Jordan, Justin Lewis, and Alexandros G Dimakis, "Provable certificates for adversarial examples: Fitting a ball in the union of polytopes," in *Advances in Neural Information Processing Systems*, 2019, pp. 14082–14092.

[8] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter, "Certified adversarial robustness via randomized smoothing," *arXiv preprint arXiv:1902.02918*, 2019.

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[10] Min Wu and Marta Kwiatkowska, "Robustness guarantees for deep neural networks on videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu, "Large-scale adversarial training for vision-and-language representation learning," *arXiv preprint arXiv:2006.06195*, 2020.

[12] Sanchari Sen, Balaraman Ravindran, and Anand Raghunathan, "Empir: Ensembles of mixed precision deep networks for increased robustness against adversarial attacks," *ICLR*, 2020.

[13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[14] Weiyao Wang, Du Tran, and Matt Feiszli, "What makes training multi-modal classification networks hard?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12695–12705.

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[16] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[17] Juncheng Li, Shuhui Qu, Xinjian Li, Po-Yao Huang, and Florian Metze, "Appendix: On adversarial robustness of large-scale audio visual learning," `https://lijuncheng16.github.io/ICASSP2022_proof.pdf`, 2021.

[18] Michael Jorgensen, "Volumes of n-dimensional spheres and ellipsoids," "https://www.whitman.edu/documents/Academics/Mathematics/2014/jorgenmd.pdf".

[19] Saehyung Lee, Hyungyu Lee, and Sungroh Yoon, "Adversarial vertex mixup: Toward better adversarially robust generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[20] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[22] Eric Wong, Leslie Rice, and J Zico Kolter, "Fast is better than free: Revisiting adversarial training," *arXiv preprint arXiv:2001.03994*, 2020.

[23] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *arXiv preprint arXiv:1912.10211*, 2019.