Trustworthy and Robust Machine Learning for Multimedia: Challenges and Perspectives

Katsuaki Nakano, Michael Zuzak, Cory Merkel, Alexander C. Loui
Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY USA
{kn9570, mjzeec, cemeec, acleec}@rit.edu

***Abstract*—Multimedia applications for machine learning models are characterized by the fusion of multiple modalities of data. In this work, we highlight the trust and robustness challenges of machine learning that arises from data fusion. To do so, we present three case studies demonstrating how multimedia applications exacerbate existing challenges of trustworthy and robust machine learning. For the first case study, we investigate the impact of fusion depth on the robustness of multi-modal machine learning models, observing that model architecture could impact robustness. For the second case study, we investigate the impact of fusion modality on the robustness of multi-modal machine learning models, observing that fusion models are only as robust as their most susceptible modality. For the third case study, we explore the impact of weight quantization techniques on the robustness of multi-modal models, observing the need for modality-based quantization schemes. Through these case studies, we hope to shed light on the unique trust and security challenges that arise in machine learning models when applied in multimedia applications and offer insights to fortify such systems in real-world scenarios.**

***Index Terms*—Machine Learning for Multimedia, Multi-Modal Fusion, Neural Network Robustness**

I. Introduction

Multimedia enriches machine learning applications by combining diverse modalities of data, such as text, image, video, and audio, that enhances a model's ability to understand and interpret complex real-world scenarios. This is supported by a large body of research indicating that the performance of machine learning models in many tasks substantially improves through the use of multiple data modalities [1]. Examples of such tasks include object tracking [2], image processing [3], intrusion detection [4], and others [5]. Similarly, the fusion of multiple modalities has been shown to improve the robustness of machine learning models [6], [7]. As a result, multimedia fusion is frequently used to enhance machine learning.

Consequently, the development of machine learning methods that operate on data from multiple modalities has emerged as a key research area [1]. The majority of the work in this space explores the use of multimedia through the lens of machine learning. This can be observed through the extensive study of different machine learning architectures to perform fusion (e.g., signal, feature, and decision fusion) [1], the adaptation of existing attacks on machine learning to multi-modal models [8]–[10], or the use of machine learning to address conventional challenges in multimedia (e.g., media analysis [11] or captioning [12]). While these avenues of research have driven substantial innovation and improvement in both multimedia and machine learning, there has been limited exploration of such problems through the lens of a multimedia problem, emphasizing particular aspects of the data modalities on the performance of the machine learning model. This motivates our work.

In this work, we aim to explore multi-modal fusion through the lens of the specific multimedia data being operated on. Based on this, we identify three open questions regarding the trust and robustness of multimedia fusion for machine learning applications and highlight their relevance with a corresponding case study with the goal of promoting further exploration by the research community. We pose each of these questions along with a brief summary of related work below.

First, we consider the question: *Does fusion depth (e.g., signal, feature, and decision fusion) in a machine learning model impact robustness, particularly to single-modal attacks?* Prior work has found accuracy improvements arising from earlier data fusion [13], however, they do not consider the corresponding impact on robustness. For this case study, we explore the impact of the depth of the fusion of multiple modalities on the adversarial robustness of neural networks against single and multi-modal attacks.

Second, we consider the question: *Can the inclusion of data modalities that are easy to perturb make a model less robust to adversarial attacks?* While prior work has demonstrated single modal attacks on fusion models [14], [15], they generally recognize and frame fusion as a defensive measure against attacks that improves robustness [16]. For this case study, we consider if a fusion model can be less robust than its un-fused counterpart to Projected Gradient Descent (PGD) [17] and Fast Gradient Sign Method (FGSM) [18] attacks.

Finally, we consider the question: *Does the impact of quantization on model robustness differ by data modality?* In neural networks, quantization is a technique employed to compress a model to require less resources and run faster. The multimedia community has long recognized that different data modalities can be subjected to different types and levels of compression [19]. Prior work has shown that the impact of quantization on the robustness of neural networks is complex, resulting in both increases and decreases depending on the scenario [20]. This suggests that different data modalities may lead to different adversarial robustness under quantization. For this case study, we consider how different levels of quantization impact attack susceptibility for two different modalities fused modalities. We summarize the contributions of the work as follows:

- We observe that fusion strategy impacts adversarial robustness to single-modal attacks and that this result appears to differ by data modality. This suggests that the available data modalities may be relevant when selecting a fusion strategy for multimedia applications.
- We observe that the robustness of multi-modal neural networks is limited by the easiest to attack modality, differing from the conventional view that multi-modal fusion inherently improves robustness.
- We observe that robustness to adversarial perturbations differs not only by data modality, but also by the level of quantization applied to the modality. This suggests that quantization in multimedia applications should consider quantization by data modality, possibly adopting different strategies or levels of precision for each.

II. Preliminaries

A. Multi-Modal Fusion in Machine Learning Applications

Multi-modal fusion in machine learning is the concept of merging multi-modal data sets composed of data obtained from different sensors with the goal of predicting an output value: a class (e.g., 0 to 9 numbers), or a continuous value (e.g., similarity between handwritten numbers). Three primary advantages of multi-modal fusion are driving interest in machine learning [21]. First, having access to various modalities could help us gather complementary data, which is not always visible when using just one modality. Second, multi-modal fusion produce more robust machine learning models if there are several modalities available for observing the same phenomenon. Third, even if one of the modalities is absent, a multi-modal system can still function. For instance, multi-modal fusion models can still identify emotions from a person's facial expressions even when they are silent [22].

In this work, we separate neural network architectures that perform multi-modal fusion into three classes based on the work in [1], [23]: early (i.e., signal), intermediate (i.e., feature), and late (i.e., decision) fusion. Early (i.e., signal) fusion is applied before entering a recognition network. It converts unprocessed data into an intermediate, more condensed form. Intermediate (i.e., feature) fusion is fusion that occurs inside recognition models. To create a new representation that is more expressive than the individual representations from which it originated, this fusion combines the characteristics that set each type of data apart [23]. Late (i.e., decision) fusion is a strategy that happens outside of the classification models using single modality. It creates new selections that are more accurate and dependable by combining the choices made by each classifier. When compared to using a single representation alone, these fusion methods can produce strong results [13].

B. Security, Privacy, and Trust in Multimedia

Multimedia systems that leverage state-of-the-art deep learning techniques are vulnerable to a number of adversarial attacks which can compromise the security, performance, privacy, and overall trustworthiness of the system and its users. Evasion, poisoning, and privacy attacks are three categories that have received significant research attention over the past several years [24]–[26]. Evasion attacks exploit small-margin decision boundaries by perturbing legitimate inputs just enough to move them to a different decision region in the input space. Poisoning attacks typically use modified labeling or addition of training data to reduce the margins of decision boundaries or insert new boundaries that cause misclassifications and/or make evasion attacks easier to perform. Privacy attacks steal information about the machine learning model (parameters, etc.) or training data through statistical analysis of query results or side channel information.

We are particularly interested in the robustness of multimedia systems to evasion attacks, which have been demonstrated on deep learning models for multiple modalities, including image [27], audio [28], text [29], physiological data [30], and more. The goal of an evasion attack can be expressed as an optimization problem, where, for some model Π, a correctly-classified input $\mathbf{u}$, usually from the test or training set, is perturbed by $\mathbf{r}^*$ to maximize a loss function $\mathcal{L}$ and cause Π's classification of $\mathbf{u}' = \mathbf{u} + \mathbf{r}^*$ to be different from $\mathbf{u}$'s ground truth label. Specific choices of the optimization procedure used and the constraints placed on $\mathbf{r}^*$ lead to different specific attack variants [26]. Recently, there has also been considerable interest in the impacts of machine learning model compression techniques on their robustness to these types of adversarial attacks. For example, parameter quantization, which is a popular compression technique [31], was recently shown to either improve or degrade the robustness of neural networks depending on the strength of the attack (length of $\mathbf{r}^*$) [20].

C. Threat Model and Scope

In this work, we consider a white-box scenario in which the attacker has complete knowledge of the model being evaluated, including all model parameters. This threat model is commonly adopted in prior work exploring adversarial attacks on fusion models [7], [9], [15]. Despite the explicit consideration of a white-box attacker, we aim to make the corresponding preliminary analysis and observations drawn from each case study to be largely attacker-agnostic.

III. Motivation

Given the widespread adoption of multi-modal machine learning for critical decision-making tasks (e.g., autonomous driving, healthcare, predictive maintenance, etc. [2]–[5]), there is a strong need to understand the trust and robustness ramifications of multimedia on machine learning models. If we follow the conventional wisdom that more data produces better models, then incorporating multiple data modalities together is a clear benefit. However, this conventional wisdom overlooks so-called *catastrophic fusion*, which has long been recognized by multimedia researchers [32]. In this work, we aim to draw on past research in the multimedia community to identify similar scenarios where the use of multimedia may lead to non-intuitive outcomes with a focus on trust and robustness. For each scenario, we present a case study and suggest future research directions based on our observations.

Specifically, we consider the following three scenarios and present a corresponding case study for each:

- **Case Study 1:** How does data fusion architecture impact model robustness, particularly against single-modal attacks?
- **Case Study 2:** Can the inclusion of easy to perturb data modalities make a model less robust to attacks?
- **Case Study 3:** Does the impact of quantization on the model robustness differ by data modality?

To begin, we outline the dataset and machine learning models used to perform each case study presented in the work.

A. Dataset

To evaluate multi-modal fusion models, we use the written and spoken digits dataset for multi-modal learning [33]. It is a constructed dataset based on existing written and spoken digits datasets. The written digits dataset is the original MNIST dataset [34] including 70000 digit images. The spoken digits dataset was extracted from Google Speech Commands [35]. 38908 utterances of the ten digits are associated with written digits of the same class. All spoken digit data was subjected to pre-processing by extracting the Mel Frequency Cepstral Coefficients (MFCC) with standardization and normalization.

B. Multimedia Fusion Architecture for Case Studies

We selected a TinyML architecture based on ResNet_v1 from the MLPerf Tiny Deep Learning benchmark [36]. This architecture was chosen for two reasons. 1) TinyML is becoming increasingly prevalent, particularly in privacy-sensitive applications where trust and robustness are critical, because it does not require data to move off-site [37]. 2) The use of a smaller architecture allows us to train more models for each case study and aggregate results. Using the ResNet_v1 architecture as a baseline, we generated three fusion architectures using different approaches from prior art [1], [23]. We depict each model and label them as ①, ②, and ③ in Figure 1:

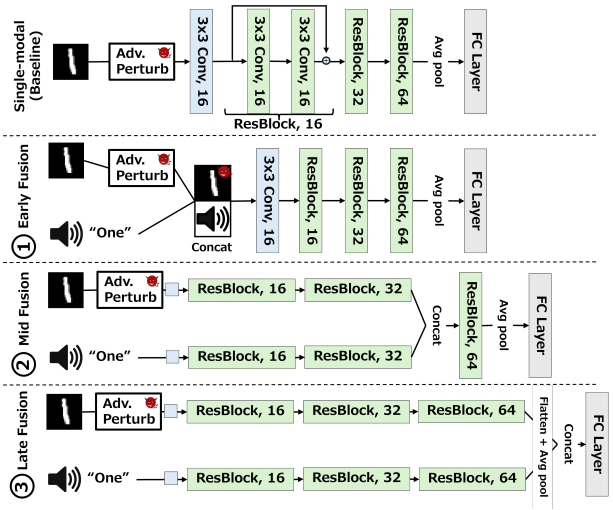

Fig. 1. Overview of case study 1.

① **Early Fusion Model:** For this model, written digit image and spoken digit MFCC are concatenated prior to being used as input to the model. Both the image and audio data are transformed to a 28 by 28 matrix and concatenated. The extra part of audio data to make the 28 by 28 shape is filled with zeros. Likewise, the shape of the audio data is changed into 28 by 28 for intermediate and late fusion model for consistency.

② **Intermediate Fusion Model:** For this model, we employed two residual blocks from the original ResNet8_v1 for each uni-modal stream. After processing each modality, these results are concatenated (i.e., fused) and fed into the remaining layers of the model, which has one residual block and a fully-connected layer at the end.

③ **Late Fusion Model:** For this model, two separate ResNet8_v1 models are used for each modality (i.e., one for image, one for audio) with the final fully-connected layer removed. The outputs of each model are concatenated (i.e., fused) and fed into a fully-connected layer.

C. Adversarial Attacks

To assess the robustness of our fusion models, we employed two adversarial attacks, PGD [17] and FGSM [18]. FGSM is an algorithm that generates an adversarial example using the gradient of a Neural Network. Since the parameter θ of the learned model can be treated as a constant, the loss is increased by adjusting the input data x. Unlike FGSM, PGD iterates the FGSM process over and over to create strong perturbations. As shown in Figure 2, the hyperparameter ϵ dictates the magnitude of the adversarial perturbation applied to the input signals. The ϵ value for PGD and FGSM was swept from 0.01 to 0.1 in increments of 0.01 in all experiments. For all experiments, we used FGSM and PGD algorithms from CleverHans [38].

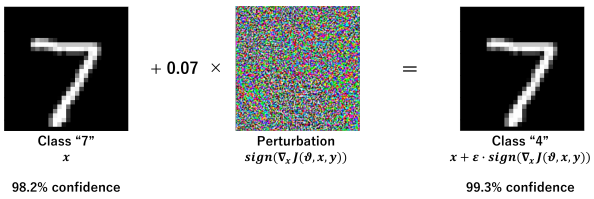

Fig. 2. An overview of the FGSM attack proposed in [18].

IV. Case Study 1: Fusion Architecture and Adversarial Robustness

For this case study, we explore the question: *How does data fusion architecture impact neural network robustness, particularly against single-modal attacks?* To do so, we launch our adversarial attacks, PGD [17] and FGSM [18], against our three fusion models outlined in Section III-B that each employ a different fusion methodology. Specifically, as shown in Figure 1, the following experiments were performed on fusion models ①, ②, and ③. 1) A single-modal attack was performed by adding adversarial perturbations to only the image input or the audio input for each fusion model. 2) A multi-modal attack was performed by adding adversarial perturbations to both the image and audio inputs of each

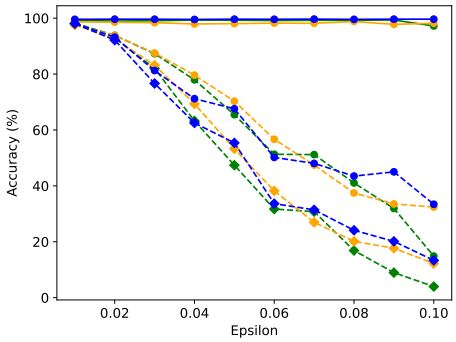

Fig. 3. Results of attacks on both modalities.

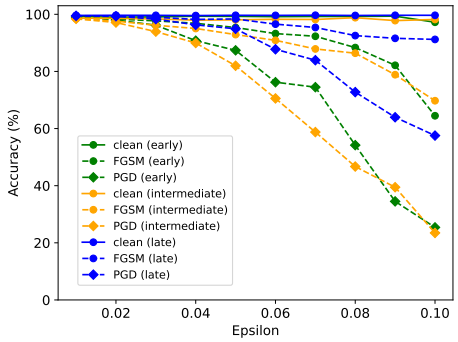

Fig. 4. Results of attacks on image input.

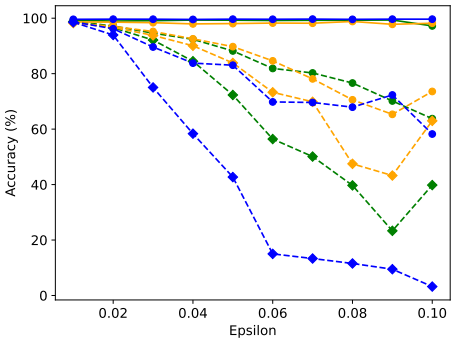

Fig. 5. Results of attacks on audio input.

model simultaneously. To serve as a baseline, two single-modal ResNet8_v1 models were trained using only the image or audio data and attacked using PGD and FGSM as well. For all experiments, the ϵ value for PGD and FGSM was swept from 0.01 to 0.1 in increments of 0.01. The results of these experiments are contained in Figures 3, 4, 5, and 6.

A. Analysis of Case Study

Figure 6 shows the accuracy of two single-modal ResNet8_v1 models that were trained using only the image *or* audio modality after FGSM and PGD attacks. Based on these results, attacking the audio input degrades the accuracy more than attacking the image input does, regardless of attack type. Figure 5 contains the resulting accuracy of the three fusion models after a single-modal FGSM and PGD attack on only the audio input. At lower epsilon values, the single-modal ResNet8_v1 model trained only on audio inputs exhibited a sharp accuracy degradation, dropping to $\sim 60\%$ accuracy at the smallest tested $\epsilon = 0.01$. Comparatively, all fusion models exhibited above 60% accuracy against even multi-modal attacks until $\epsilon = 0.04$. Hence, fusion appears to improve the robustness of the model to single-modal adversarial attacks, regardless of architecture. This observation holds regardless of the considered modality, attack, or fusion architecture.

Next, we consider the relative impact of single and multiple modality attacks on each fusion model. Figure 3 contains the multi-modal attack results. When compared to the image-only and audio-only attack results contained in Figures 4 and 5, multi-modal attacks resulted in greater accuracy degradation. This is unsurprising because the multi-modal attacks could perturb both input modalities, making them a super-set of the single-modal attack strategies. This result held regardless of the fusion architecture evaluated, suggesting that single-modal attacks are less effective than multi-modal attacks.

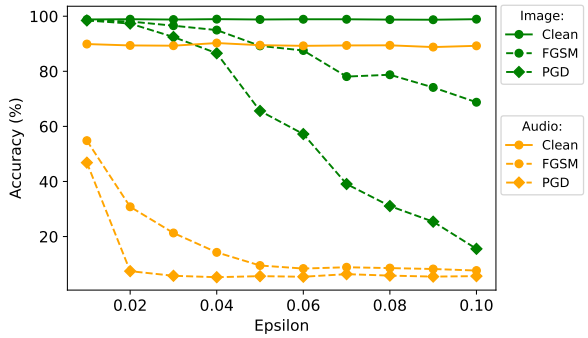

Fig. 6. Accuracy degradation from attacks on single-modal models.

Finally, we consider the impact of fusion architecture on model robustness. For attacks on the image modality contained in Figure 4, the late fusion model appears to be more robust than either other fusion strategy (i.e., early and intermediate). For attacks on the audio modality contained in Figure 5, the intermediate fusion model appears more robust against the PGD attack than the early and late fusion models. This indicates that fusion architecture may have some impact on model robustness to single-modal attack strategies.

B. Discussion and Future Research Directions

The results of the case study suggest that fusion architecture may impact the robustness of machine learning models. Previous research has shown that early fusion can enhance accuracy [13]. However, our findings suggest that early fusion does not provide similarly clear-cut improvements against single-modal adversarial attacks. We also observed differences in the robustness of each fusion strategy by the attacked data modality. This suggests that it may be advantageous to consider the specific data modalities being used when selecting the fusion architecture. Given the growing use of multimedia in critical machine learning applications [2]–[5], understanding the implications of fusion depth on model robustness, especially against single-modal attack strategies, is important. Hence, further research into how fusion architecture affects adversarial robustness in multimedia models is warranted.

V. Case Study 2: Modality Selection and Fusion Model Robustness

For this case study, we explore the question: *Can the inclusion of data modalities that are easy to perturb make a model less robust to adversarial attacks?* To evaluate this, we compare the adversarial robustness of a single-modal model using only the image modality to a multimedia model fusing both image and audio data with an early fusion approach. This is illustrated in Figure 7. In such a scenario, conventional wisdom assumes that the multimedia model will exhibit more

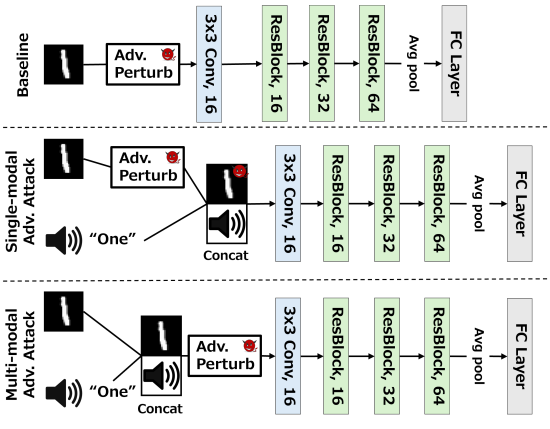

Fig. 7. Overview of case study 2.

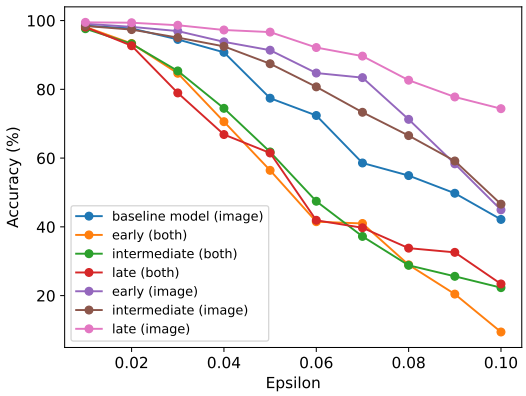

Fig. 8. Model accuracy after single and multi-modal adversarial attacks.

robustness to adversarial attacks. However, the prior case study indicates that the audio modality is more susceptible to adversarial perturbation. To assess this, we launch PGD and FGSM attacks against our single-modal and multimedia fusion models. For the multimedia model, both single-modal attacks on the image modality and multi-modal attacks on both image and audio modalities were considered.

A. Analysis of Case Study

Figure 8 shows the relevant data extracted from the graphs in Figures 3, 4, and 6. Figure 8 includes multi-modal attacks on three fusion models, single-modal attacks to the image input on three fusion models, and a single-modal attack to the image input on the baseline (i.e., without fusion) model. Each data point is the average of the results of both evaluated attack strategies, FGSM and PGD. Single-modal attacks on the three fusion models do not degrade accuracy more than a single-modal attack does on the baseline model. This suggests that introducing audio fusion makes the model more robust against single-modal image attacks.

However, when compared the baseline model, multi-modal attacks on the three fusion models degrades accuracy more than the single-modal attack does on the baseline model. This connects to the result that introducing audio, which based on case study 1 is a modality more susceptible to adversarial attacks, results in a less robust model when both modalities can be attacked. Therefore, this indicates a counter-intuitive scenario, where introducing an additional complementary modality actually leads to a less robust model overall if both modalities can be perturbed.

B. Discussion and Future Research Directions

The results of this case study highlight the importance of considering the relative adversarial robustness of candidate data modalities in multimedia fusion. Specifically, when adapting a single-modal model into a multimedia fusion model, incorporating a data modality with lower adversarial robustness (i.e., audio in this case) may compromise the overall robustness of the fusion model. This serves as a counterexample to the conventional wisdom that adding complementary data enhances model robustness [6], [7]. This case study suggests a more complex scenario where the inclusion of additional, complementary data modalities does not always lead to a more robust multimedia model.

This implies that the robustness of fusion modalities is a critical factor in the deployment of neural networks in multimedia applications. It points to the need for further research into how the robustness of individual modalities affects the overall robustness of a multimedia model. Based on this, we propose two potential factors that could impact model robustness in multimedia settings as future research directions:

- *The adversarial robustness of a candidate modality.* Introducing a new modality that is robust to perturbations can enhance the overall model's resistance to adversarial attacks, including multi-modal ones. For example, in our case study, if the original model utilized only audio, adding the image modality, which exhibited higher robustness, would have improved the model's overall defense against against both a single and multiple modalities. If this scenario generalizes, one could strive to fuse data modalities that are resistant to adversarial perturbation to improve model robustness.
- *The relative difficulty of performing adversarial perturbation to a candidate data modality.* In scenarios where only single-modal attacks are feasible, our case study suggests that incorporating this modality may enhance overall model robustness. This finding implies that fusing even a less robust modality may still be beneficial if attacks on it are sufficiently difficult. For instance, tasks like object detection, which typically rely on image data, are vulnerable to attacks using small, strategically placed patches [39]. However, by integrating extra data modalities, such as Lidar, which requires a complex physical setup to attack [40], the robustness of the model could be improved. This is because multi-modal attacks would be more challenging to execute.

VI. Case Study 3: Quantization and Fusion Model Robustness

For this case study, we explore the question: *Does the impact of quantization on model robustness differ by data*

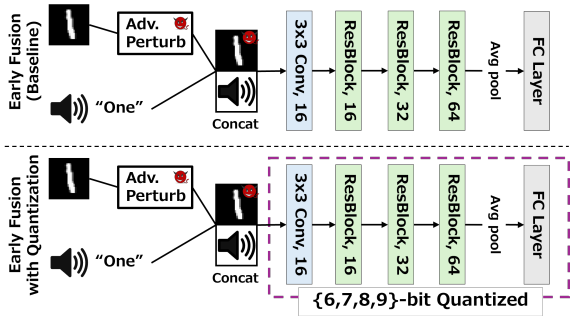

Fig. 9. Overview of case study 3.

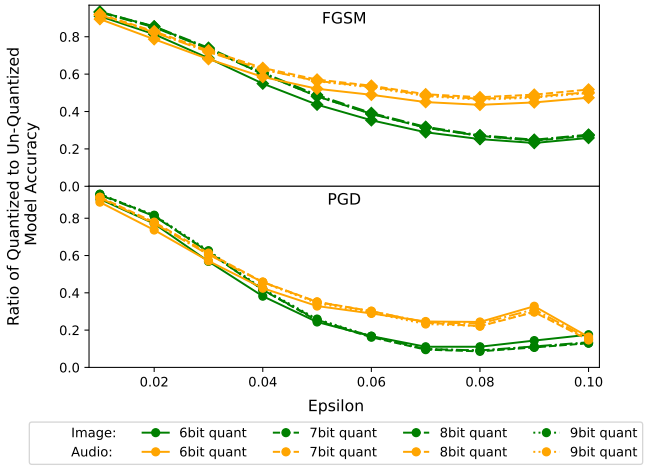

Fig. 10. Accuracy of quantized model after single-modal FGSM/PGD attack.

modality? To evaluate this, we quantized our mulitmedia model employing early fusion to $\{6, 7, 8, 9\}$ bits. For each quantized model, we launched a single-modal attack, applying adversarial perturbation to only the image or audio inputs using PGD [17] and FGSM [18]. The resulting accuracy was aggregated to determine whether quantization had disparate impact on model robustness between the image and audio modality. Figure 9 contains an overview of this experiment.

A. Quantization Technique

To quantize a model, we performed post training quantization using min-max scaling [41]. To perform n-bit quantization, the minimum and maximum parameter value for each layer was used to determine the range for the quantization of each layer. This range was then sliced into 2^n discrete values and all layer parameters were rounded to the nearest one.

B. Analysis of Case Study

Figure 10 displays the ratio of quantized and un-quantized model accuracy after single-modal FGSM and PGD attacks. This ratio is used to isolate the reduction in accuracy caused by quantization. For both the FGSM and PGD attack, a single-modal attack on the image input of the quantized models reduces the inference accuracy more than a single-modal attack on the audio input of the quantized models. However, we note that this result is more pronounced for the FGSM attack and higher ϵ values. Based on this result, quantization appears to degrade the robustness of the image modality more than the audio modality. This suggests that quantizing different modalities may result in a disparate impact on robustness.

C. Discussion and Future Research Directions

The results of this case study indicate that quantization impacts model robustness differently across data modalities. Specifically, quantization reduced adversarial robustness in the image modality more than in the audio modality. Therefore, just as there are modality-dependent compression algorithms for signal processing (e.g., JPEG), our findings suggest that similar modality-dependent quantization algorithms could benefit multimedia machine learning applications. Furthermore, modality-dependent, mixed-precision quantization approaches may also be advantageous. Existing algorithms for mixed-precision quantization, such as [42], could be adapted to tailor quantization strategies to the specific characteristics of each data modality. This area of research is particularly promising as edge-deployments of machine learning, where quantization is common, are predicted to substantially increase [37].

VII. Discussion of Limitations

The case studies presented in this work provide small, intuitive examples to highlight interesting concepts at the intersection of multimedia, machine learning, and security. While we have made observations based on these case studies, they are insufficient to draw any definitive conclusions for the considered research questions. Rather, each case study is presented as anecdotal data to showcase how aspects of multimedia may impact machine learning in a non-intuitive fashion with the goal of highlighting possible directions for future research on multimedia.

VIII. Conclusion

In this work, we presented three case studies exploring the impact of multimedia on machine learning applications. Each case study indicates that multimedia applications may exacerbate existing challenges of trustworthy and robust machine learning. For the first case study, we observe that the fusion architecture of multi-modal machine learning models (i.e., early, intermediate, or late fusion) greatly impacts susceptibility to adversarial attacks. For the second case study, we observed that the use of fusion can in fact degrade the overall robustness of a model to attacks. For the third case study, we observed that quantization impacted the robustness of machine learning models differently based on robustness. Alongside each case study, we identified candidate areas of future research on trustworthy and robust machine learning for multimedia with the hope of shedding light on the unique trust and security challenges that arise when machine learning is applied to multimedia applications.

IX. Acknowledgements

This work was supported by the National Science Foundation (NSF) under Grant 2245573.

REFERENCES

[1] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, pp. 115–129, 2020.

[2] Y. Cui *et al.*, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.

[3] H. Kaur, D. Koundal, and V. Kadyan, "Image fusion techniques: a survey," *Archives of computational methods in Engineering*, vol. 28, no. 7, pp. 4425–4447, 2021.

[4] G. Li, Z. Yan, Y. Fu, H. Chen *et al.*, "Data fusion for network intrusion detection: a review," *Security and Communication Networks*, vol. 2018, 2018.

[5] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[6] M. Bednarek, P. Kicki, and K. Walas, "On robustness of multi-modal fusion—robotics perspective," *Electronics*, vol. 9, no. 7, p. 1152, 2020.

[7] S. Wang, T. Wu, A. Chakrabarti, and Y. Vorobeychik, "Adversarial robustness of deep sensor fusion models," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2387–2396.

[8] J. Tu *et al.*, "Exploring adversarial robustness of multi-sensor perception systems in self driving," *arXiv preprint arXiv:2101.06784*, 2021.

[9] Y. Cao *et al.*, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks," in *S&P*, 2021.

[10] M. Abdelfattah, K. Yuan, Z. J. Wang, and R. Ward, "Towards universal physical attacks on cascaded camera-lidar 3d object detection models," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3592–3596.

[11] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intelligent Systems*, vol. 36, no. 2, pp. 88–95, 2021.

[12] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–39, 2023.

[13] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd international conference on information fusion (FUSION)*. IEEE, 2020, pp. 1–6.

[14] K. Yang, W.-Y. Lin, M. Barman, F. Condessa, and Z. Kolter, "Defending multimodal fusion models against single-source adversaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3340–3349.

[15] Z. Cheng *et al.*, "Fusion is not enough: Single modal attack on fusion models for 3d object detection," in *The Twelfth International Conference on Learning Representations*, 2023.

[16] T. Liang *et al.*, "BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework," in *NeurIPS*, 2022.

[17] 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. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[19] V. Bhaskaran and K. Konstantinides, "Image and video compression standards: algorithms and architectures," 1997.

[20] M. Gorsline, J. Smith, and C. Merkel, "On the adversarial robustness of quantized neural networks," in *Proceedings of the 2021 on Great Lakes Symposium on VLSI*, 2021, pp. 189–194.

[21] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[22] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, 2015. [Online]. Available: https://doi.org/10.1145/2682899

[23] S. Boulahia, A. Amamra, M. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, 11 2021.

[24] A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, *Adversarial Machine Learning*. Cambridge University Press, 2018.

[25] Y. Vorobeychik and M. Kantarcioglu, "Adversarial machine learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 12, no. 3, pp. 1–169, 2018.

[26] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[27] J. C. Costa, T. Roxo, H. Proença, and P. R. Inácio, "How deep learning sees the world: A survey on adversarial attacks & defenses," *IEEE Access*, 2024.

[28] X. Liu, K. Wan, Y. Ding, X. Zhang, and Q. Zhu, "Weighted-sampling audio adversarial example attack," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4908–4915.

[29] M. Behjati, S.-M. Moosavi-Dezfooli, M. S. Baghshah, and P. Frossard, "Universal adversarial attacks on text classifiers," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7345–7349.

[30] X. Han *et al.*, "Deep learning models for electrocardiograms are susceptible to adversarial attack," *Nature medicine*, vol. 26, no. 3, pp. 360–363, 2020.

[31] A. Gholami *et al.*, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*. Chapman and Hall/CRC, 2022, pp. 291–326.

[32] J. R. Movellan and P. Mineiro, "Modularity and catastrophic fusion: A bayesian approach with applications to audio-visual speech recognition," *Departement of Cognitive Science, USCD, San Diego, CA, Tech. Rep*, vol. 97, 1997.

[33] L. Khacef, L. Rodriguez, and B. Miramond, "Written and spoken digits database for multimodal learning," 2019. [Online]. Available: https://hal.science/hal-02327938

[34] Y. LeCun and C. Cortes, "Mnist handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[35] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 04 2018.

[36] C. Banbury *et al.*, "Mlperf tiny benchmark," *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[37] Y. Abadade *et al.*, "A comprehensive survey on tinyml," *IEEE Access*, 2023.

[38] N. Papernot *et al.*, "Technical report on the cleverhans v2.1.0 adversarial examples library," *arXiv preprint arXiv:1610.00768*, 2018.

[39] Z. Cheng *et al.*, "Physical attack on monocular depth estimation with optimal adversarial patches," in *European conference on computer vision*. Springer, 2022, pp. 514–532.

[40] Y. Cao *et al.*, "Adversarial sensor attack on lidar-based perception in autonomous driving," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 2267–2281.

[41] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," *arXiv preprint arXiv:1806.08342*, 2018.

[42] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Hawq: Hessian aware quantization of neural networks with mixed-precision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 293–302.