Trustworthy Machine Learning in Multimodal AI Applications: Case Studies and Perspectives

Alexander C. Loui

Rochester Institute of Technology

Rochester, NY USA

Contributors: Katsuaki Nakano, Michael Zuzak, Renaaron Ellis, and Cory Merkel

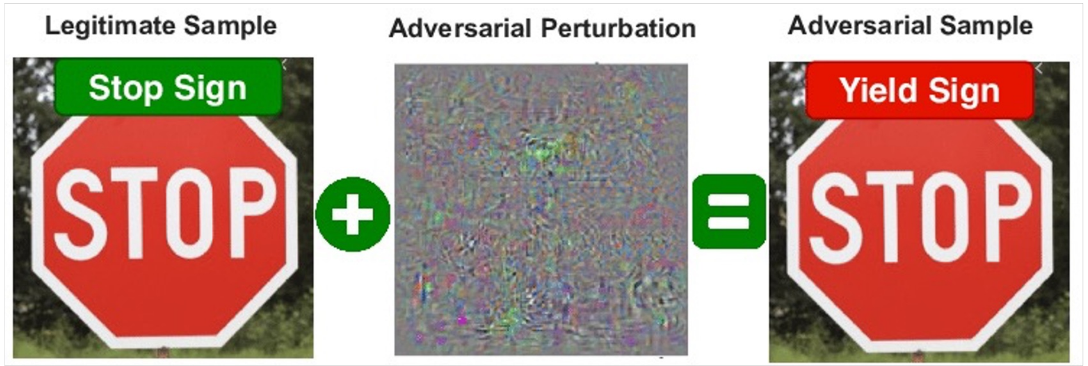
AIDA Symposium and Summer School on ‘AI/ML Cutting Edge Trends’ **(AIDA AICET’25)**

Outline

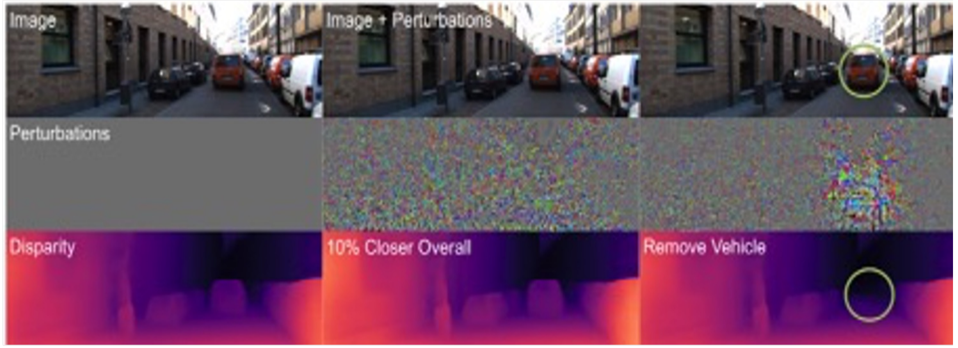
- Motivations
- Multimodality in AI and machine learning applications
- Adversarial attacks on different fusion architectures and models
- Case studies:
  - Does fusion depth in a ML model impact robustness, particularly to single-modal attacks?
  - Can the inclusion of data modalities that are more vulnerable to perturbation make a model less robust to adversarial attacks?
  - Does the impact of quantization on model robustness differ by data modality?
- Summary & future work

Adversarial Attacks on ML Models

Digital-Space Attacks:

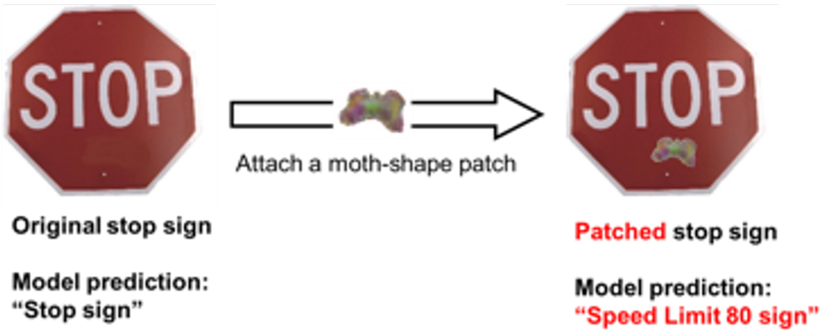


Wrong Traffic Sign Recognition



Wrong Distance (Depth) Estimation

Physical-World Attacks:



Wrong Traffic Sign Recognition

Wrong Distance (Depth) Estimation

Z. Cheng, et al, "Physical Attack on Monocular Depth Estimation with Optimal Adversarial Patches," *ECCV 2022*.

Adversarial Attacks on ML Models

- Multimodal fusion models can be **vulnerable** to adversarial attacks.
- Examples below show that when a patch is present in front, the pedestrians crossing the street cannot be detected by a fusion model anymore.

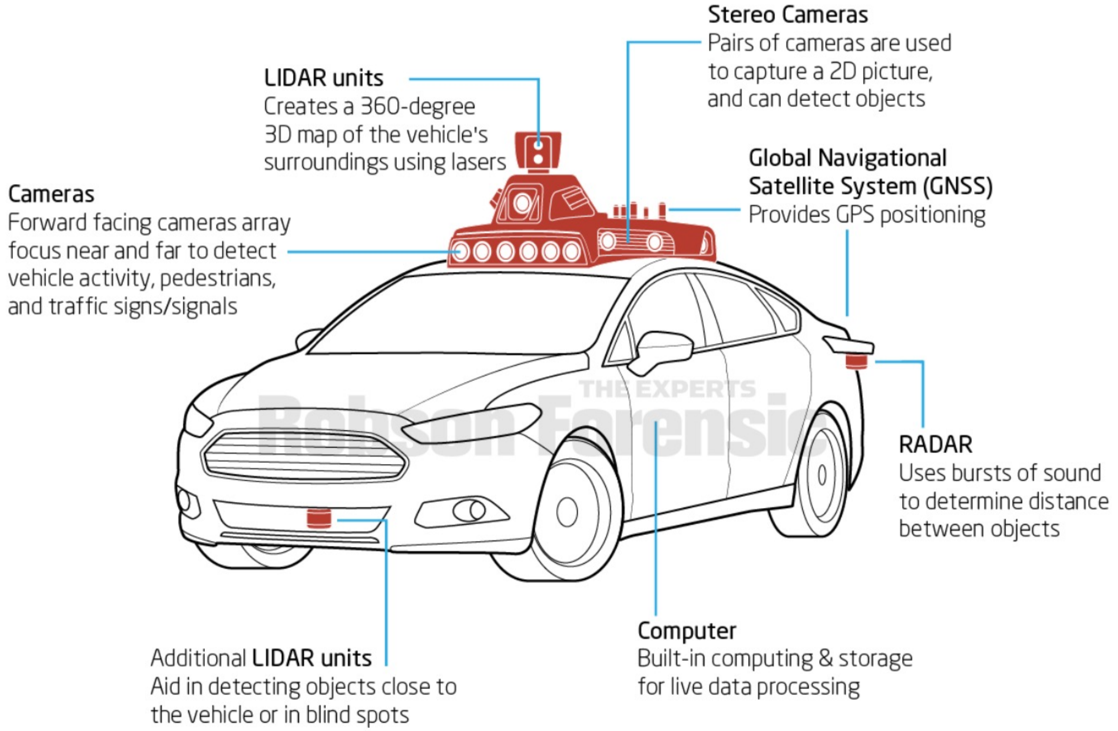
(a) Benign Scenario (b) Patch on the Ground (c) Adversarial Scenario

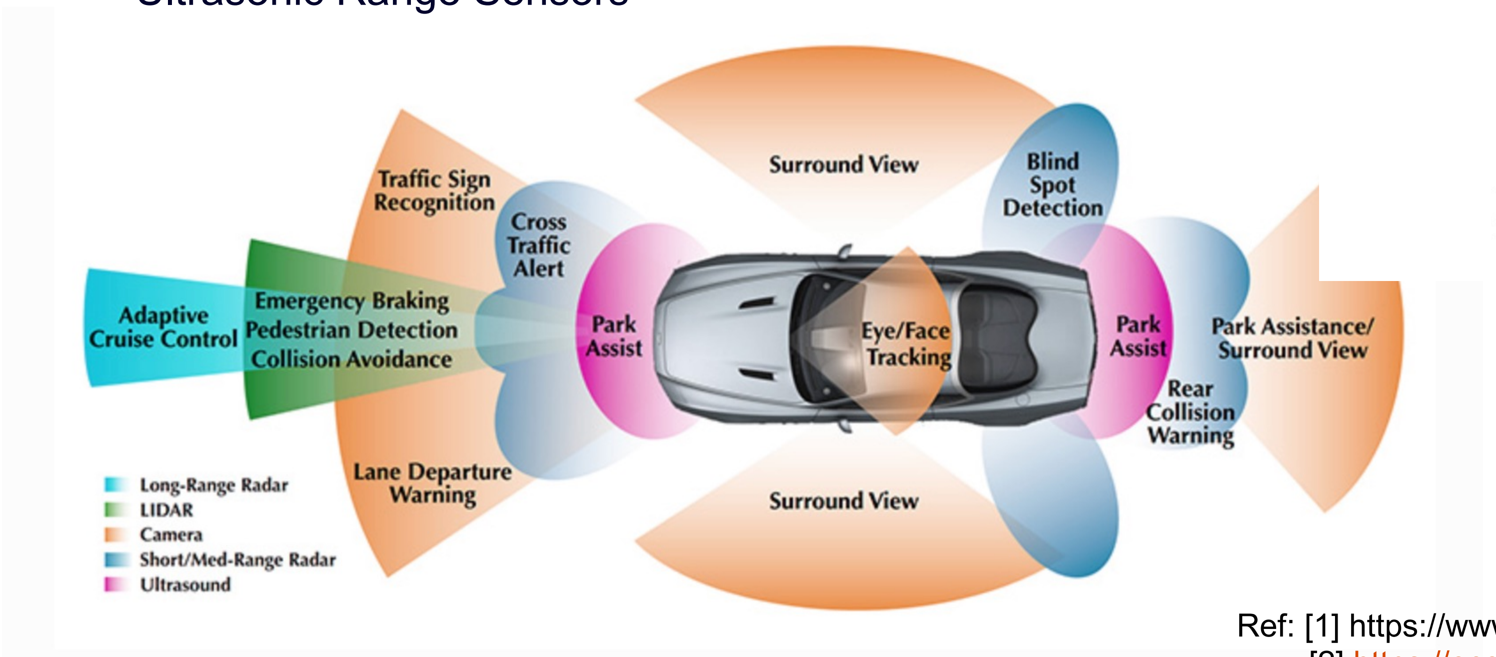
Z. Cheng et al, Fusion Is Not Enough: Single Modal Attacks on Fusion Models for 3D Object Detection, ICLR 2024

Multimodal AI Applications

Autonomous Driving

- LiDAR
- Video cameras
- Radar
- GNSS/GPS
- Ultrasonic Range Sensors



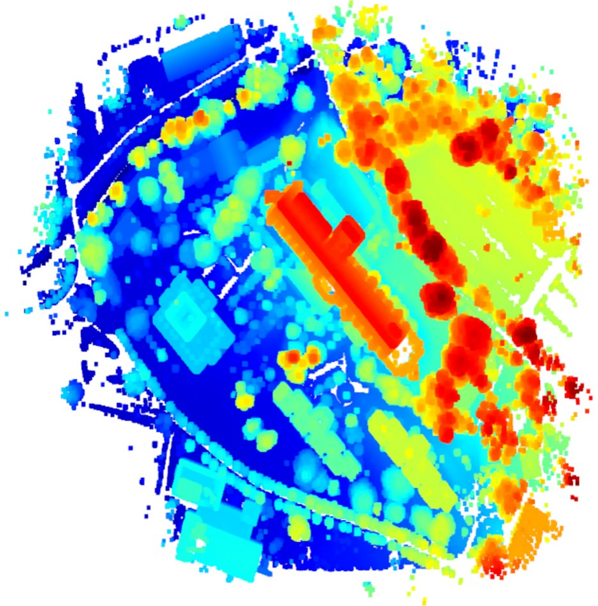


Ref: [1] https://www.robsonforensic.com/articles/autonomous-vehicles-sensors-expert
[2] https://ecotron.ai/blog/introduction-to-autonomous-driving-sensors/

Multimodal AI Applications

Multispectral Image Segmentation

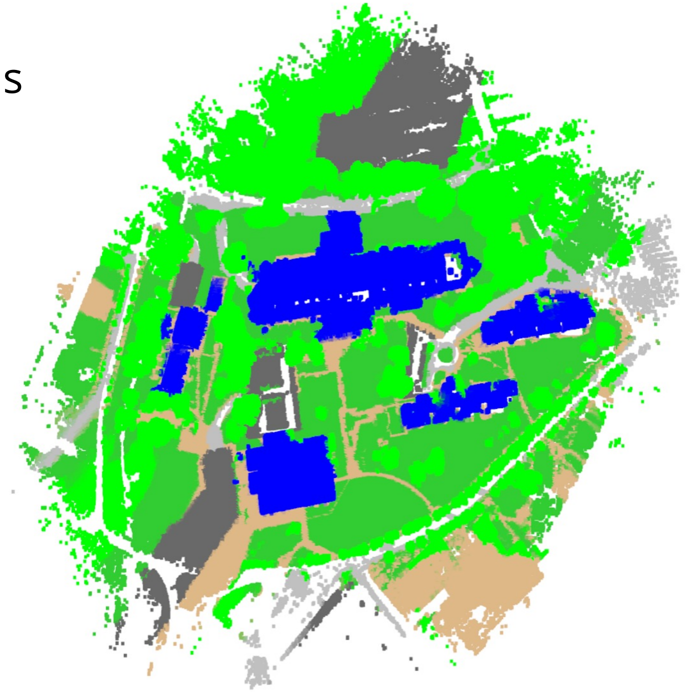
RGB Point Cloud



LiDAR Point Cloud

Ground Truth Labels

- Trees
- Grass
- Parking lot
- Roadway
- Walkway
- Buildings
- Car



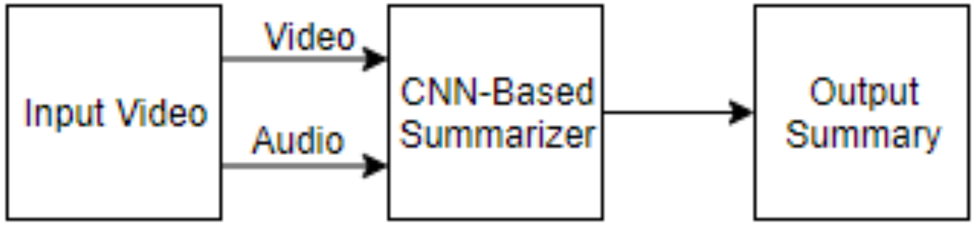
Semantic Segmentation

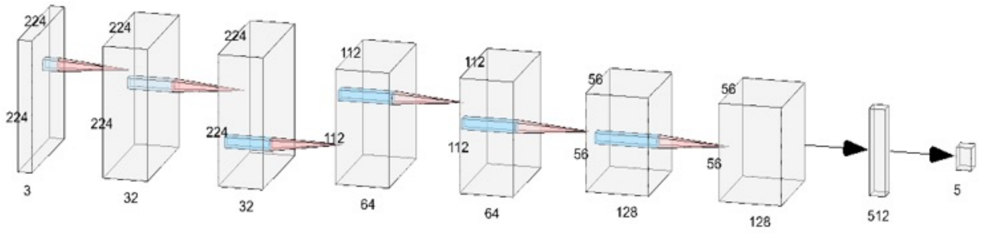
Other image inputs: Near Infrared, Red Edge ($\lambda \sim 0.717$), Edge Map

Multimodal AI Applications

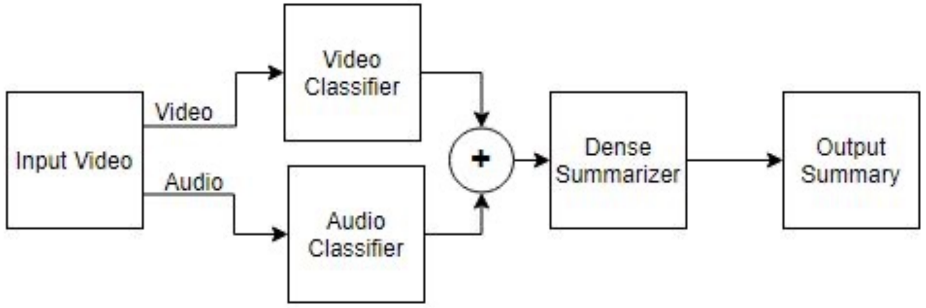
Video Summarization

- A process of taking a video and creating a shorter summary based on significant/interesting parts:
  - Video summary, image summary, text summary

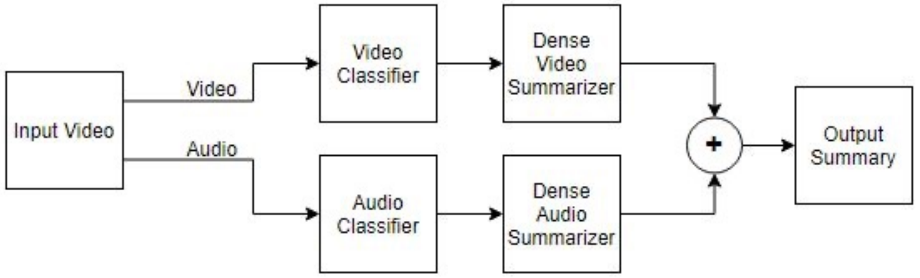




Pre-Fusion

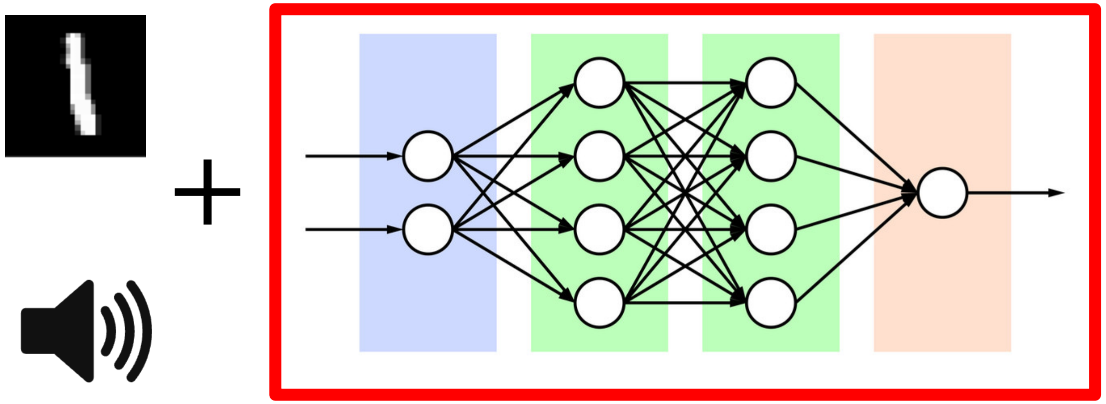


Post-Fusion



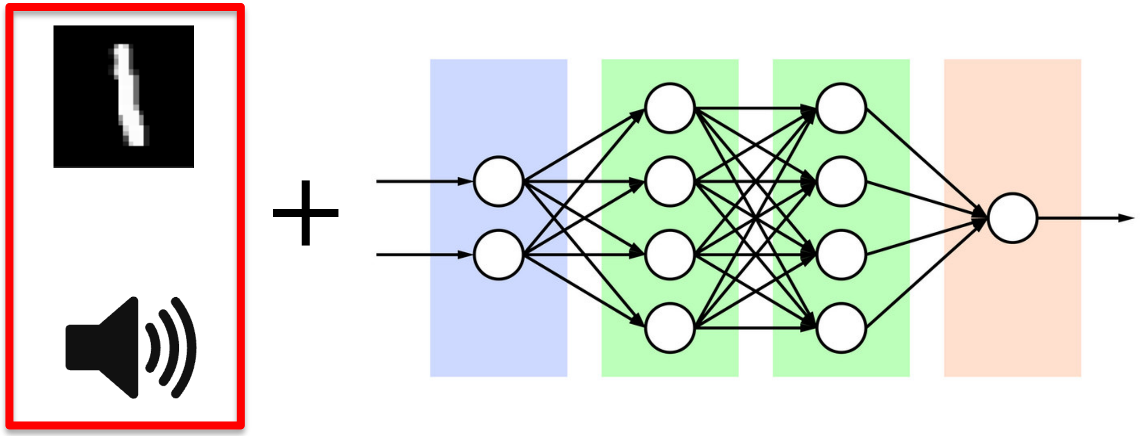
Background

Major Work:



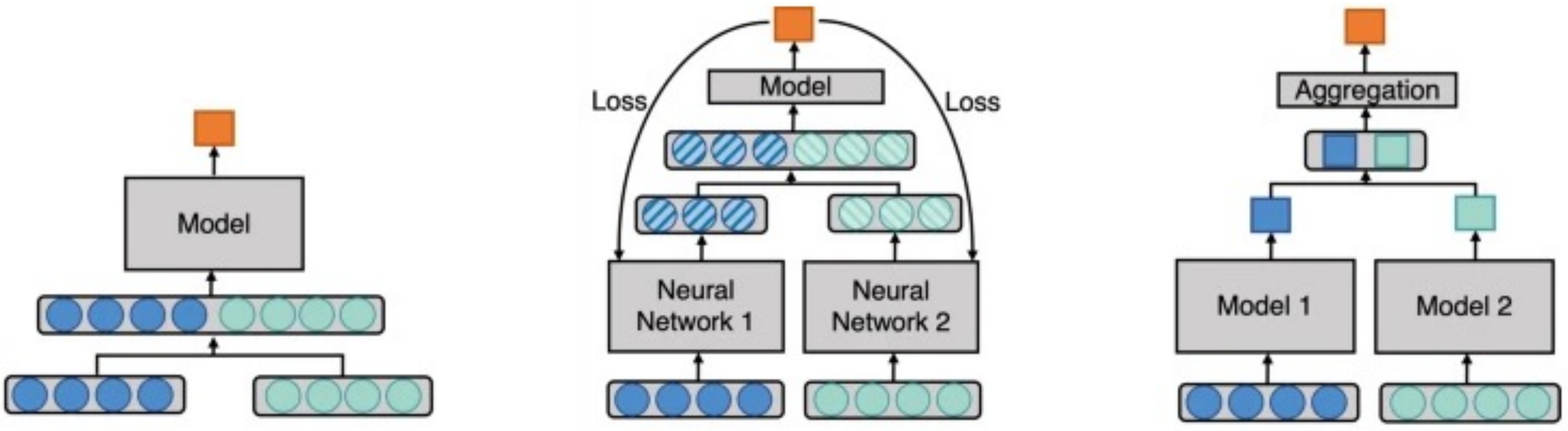
- Multimedia through **Machine Learning**
- Fusion architectures (signal, feature, and decision fusion)

Our Work:



- Explore multi-modal fusion through the lens of **Data Modalities**
- Trust and robustness of multimedia fusion model

Fusion Architectures

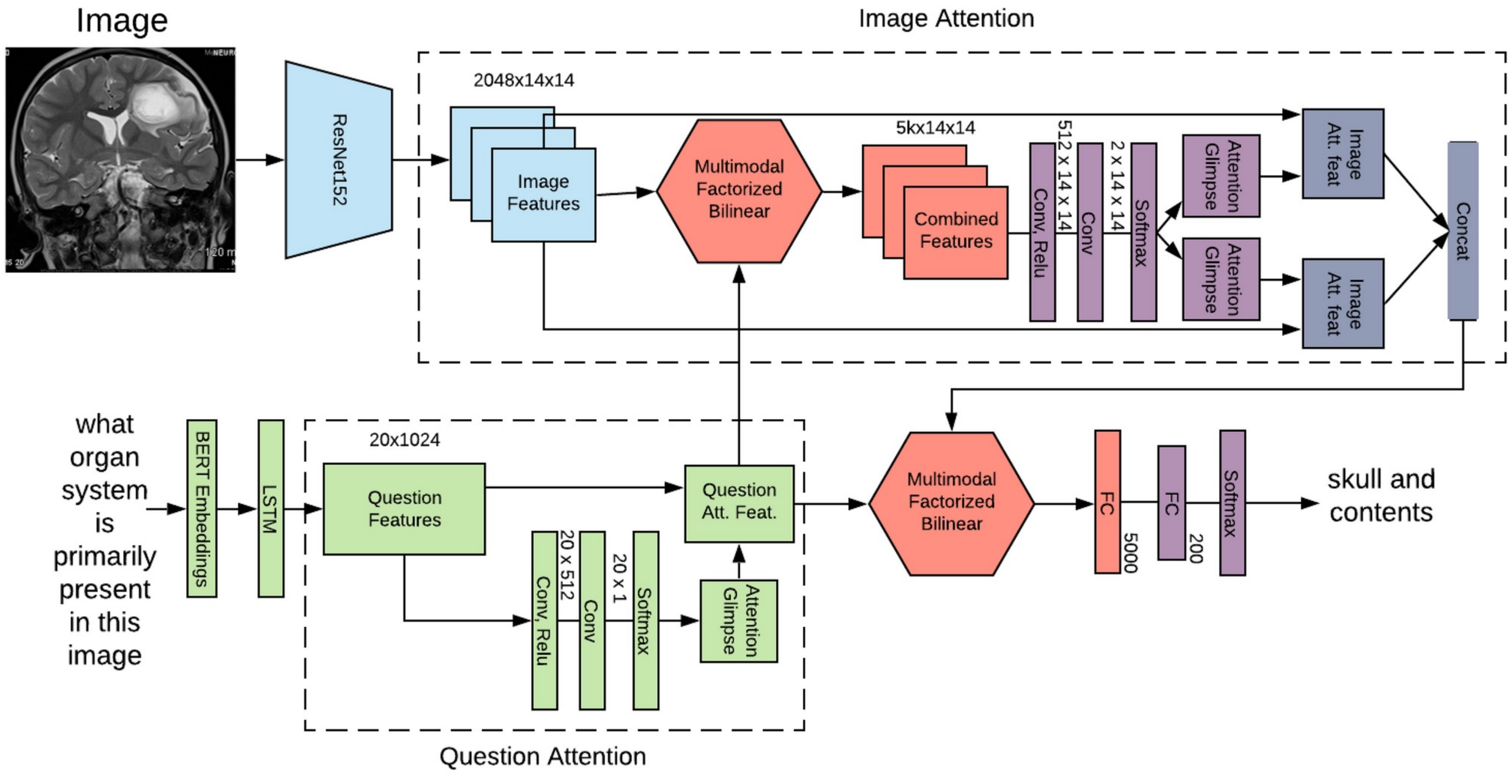


Early Fusion Mid Fusion Late Fusion

- **Three types of fusion architectures in Deep Learning**
  - **Early Fusion** concatenates original or extracted features at the input level
  - **Intermediate Fusion** joints feature representations from intermediate layers of neural networks
  - **Late Fusion** combines the predictions of multiple models

Huang, SC., Pareek, A., Seyyedi, S. *et al.* Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3**, 136 (2020).

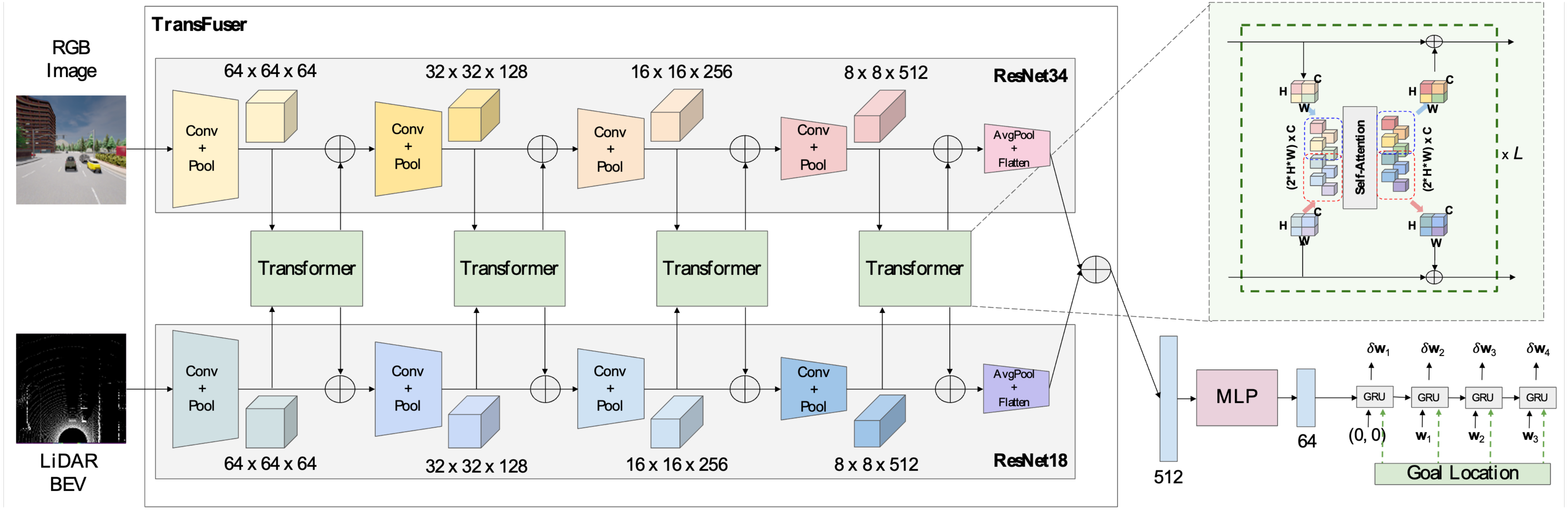
CNN-based Fusion Example



- **Visual Q&A is a representative task with multi modal features**
  - Image features are extracted by ResNet
  - Question features are extracted by LSTM
  - These features are concatenated in the middle of the architecture

Sharma, D., Purushotham, S. & Reddy, C.K. MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Sci Rep* **11**, 19826 (2021).

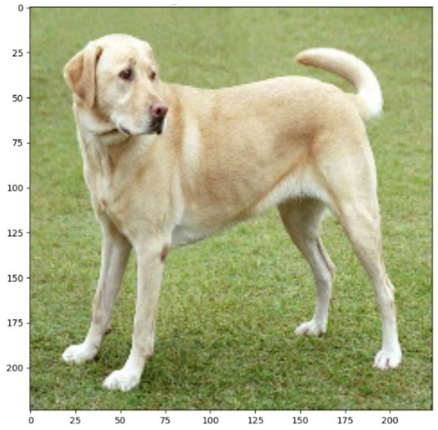
Transformer-based Fusion Example



- **Integrating multiple features is essential to perform autonomous driving**
  - Both RGB image and LiDAR data are processed by CNN and Transformer layers
  - Transformers in the middle share these features in 4 different levels
  - Each ResNet stream is concatenated at the end of the process

A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7073-7083 (2021).

Adversarial Attack Methods

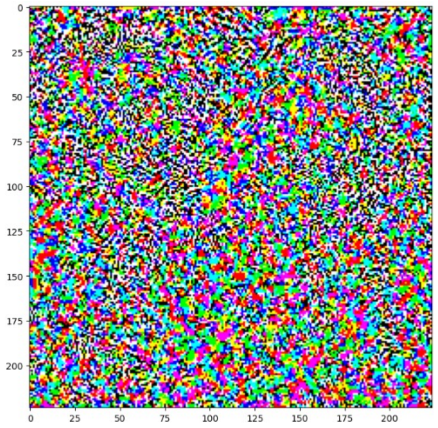


Class: "Labrador retriever"

x

41.82% Confidence

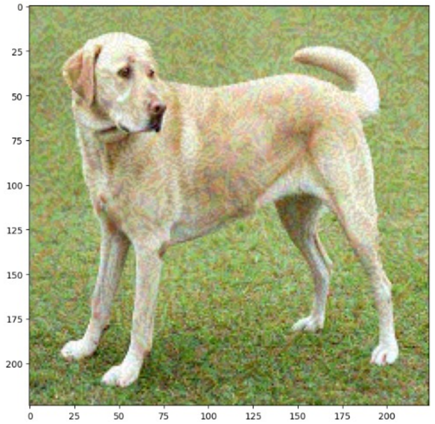
+ ε x



Perturbation

$sign(\nabla_x J(\theta, x, y))$

=



Class: "Saluki"

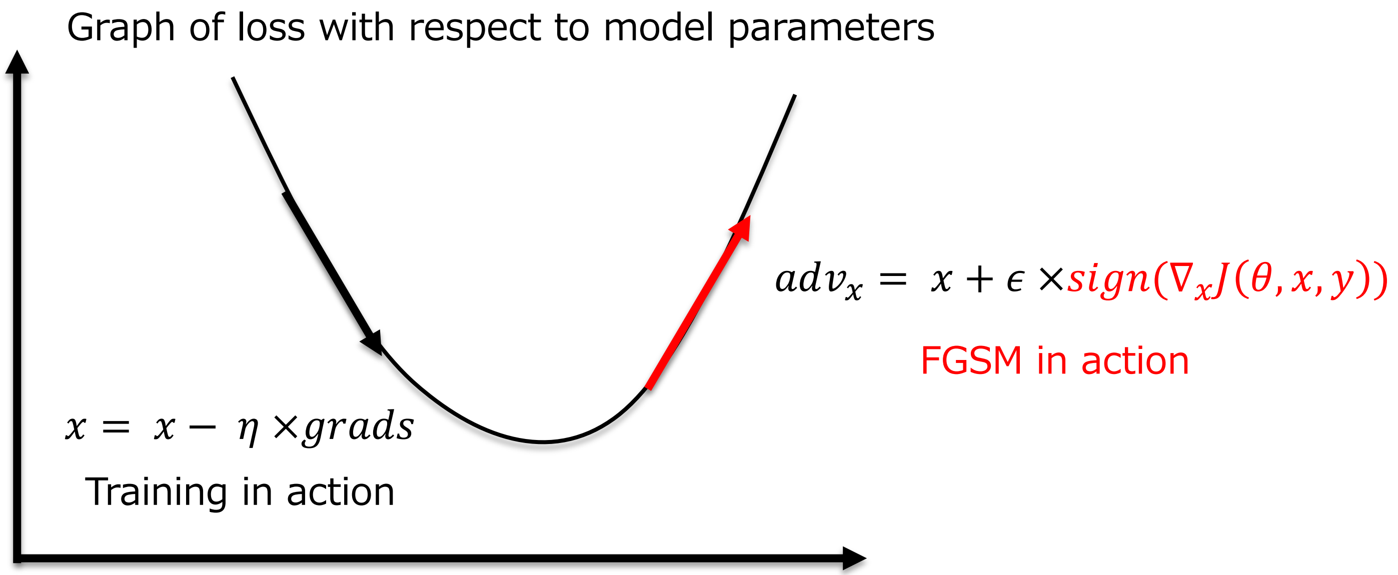
$x + \epsilon \times sign(\nabla_x J(\theta, x, y))$

13.08% Confidence

- **An adversarial example: the original image + perturbation**
- **Methods to generate perturbation with known model parameters**
  - Fast Gradient Sign Method (FGSM)
  - Projected Gradient Descent (PGD)

Fast Gradient Sign Method (FGSM)

In FGSM, nudge the pixels of the image slightly in the direction of the calculated gradients that maximize the loss calculated.

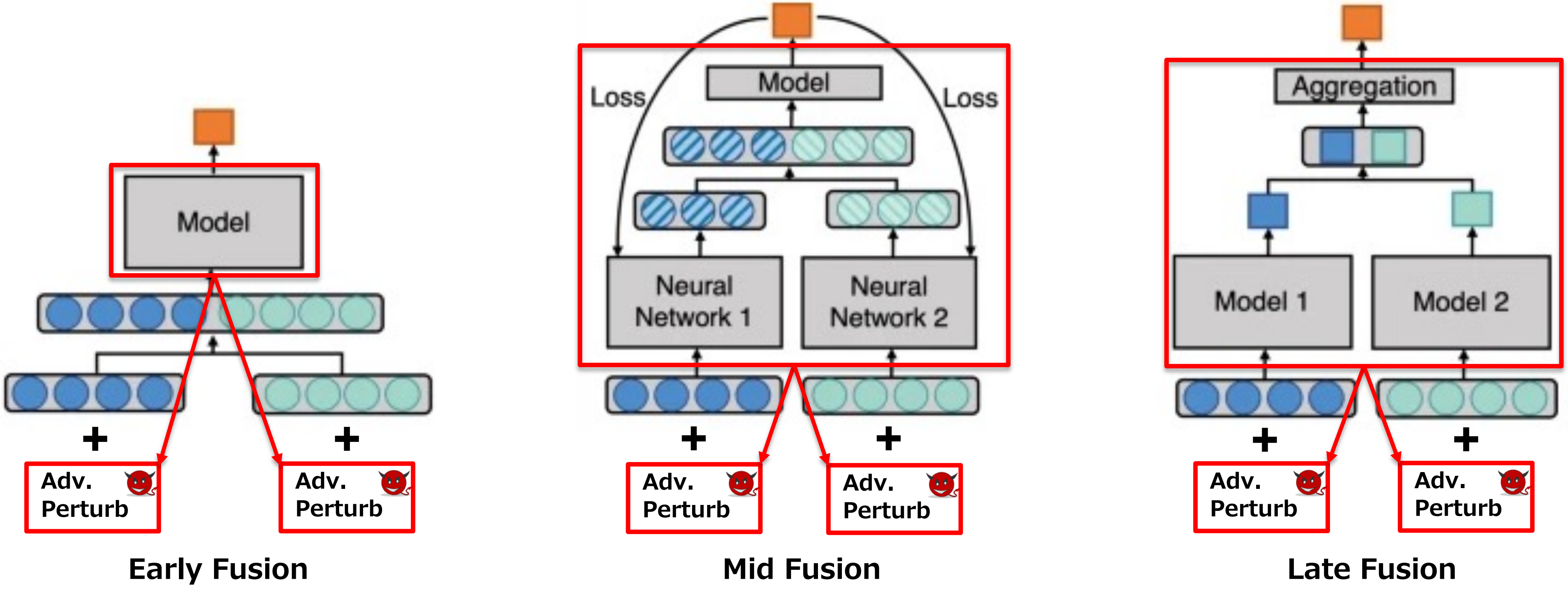


- The Fast Gradient Sign Method (FGSM) works by using the gradients of the neural network to create an adversarial sample.
- For an input image, the method uses the gradients of the loss (∇_x) with respect to the input image to create a new image that maximizes the loss. This new image is called the adversarial image, adv_x.
- The noise on the resulting image depends on the epsilon, ϵ
  - The larger the value, the more noticeable the noise

Projected Gradient Descent (PGD)

- Projected Gradient Descent (PGD) is an iterative method used in adversarial machine learning to create adversarial samples.
- PGD is a variant of FGSM applied iteratively with projection.
- PGD operates by applying small but iteratively adjusted perturbations to the input data, aimed at maximizing the model's prediction error.
- Specifically, the update rule for PGD is defined as
  - $x'_{t+1} = P(x_t + \alpha \cdot sign(\nabla_x J(\Theta, x_t, y)))$, where, x_t is the input at iteration t, α is the step size, $\nabla_x J(\Theta, x_t, y)$ is the gradient of the loss with respect to the input, and P is the projection operator ensuring perturbed input stays within predefined bounds.
- PGD is generally considered more effective in creating adversarial examples

Model-based Adversarial Attacks



- **Adversarial perturbations will be added to both inputs or either one of them**
  - These perturbations are created based on models in the case of white-box attack

Huang, SC., Pareek, A., Seyyedi, S. *et al.* Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3**, 136 (2020).

Research Questions

Research Questions

- Question 1: Does fusion depth in a ML model impact robustness, particularly to single-modal attacks?

Research Questions

- Question 1: Does fusion depth in a ML model impact robustness, particularly to single-modal attacks?

- Question 2: Can the inclusion of data modalities that are more vulnerable to perturbation make a model less robust to adversarial attacks?

Research Questions

- Question 1: Does fusion depth in a ML model impact robustness, particularly to single-modal attacks?
- Question 2: Can the inclusion of data modalities that are more vulnerable to perturbation make a model less robust to adversarial attacks?
- Question 3: Does the impact of quantization on model robustness differ by data modality?

Research Questions

- Question 1: Does fusion depth in a ML model impact robustness, particularly to single-modal attacks?
- Question 2: Can the inclusion of data modalities that are more vulnerable to perturbation make a model less robust to adversarial attacks?
- Question 3: Does the impact of quantization on model robustness differ by data modality?

Case Study 1: Overview

Question 1: Does fusion depth in a ML model impact robustness, particularly to single-modal attacks?

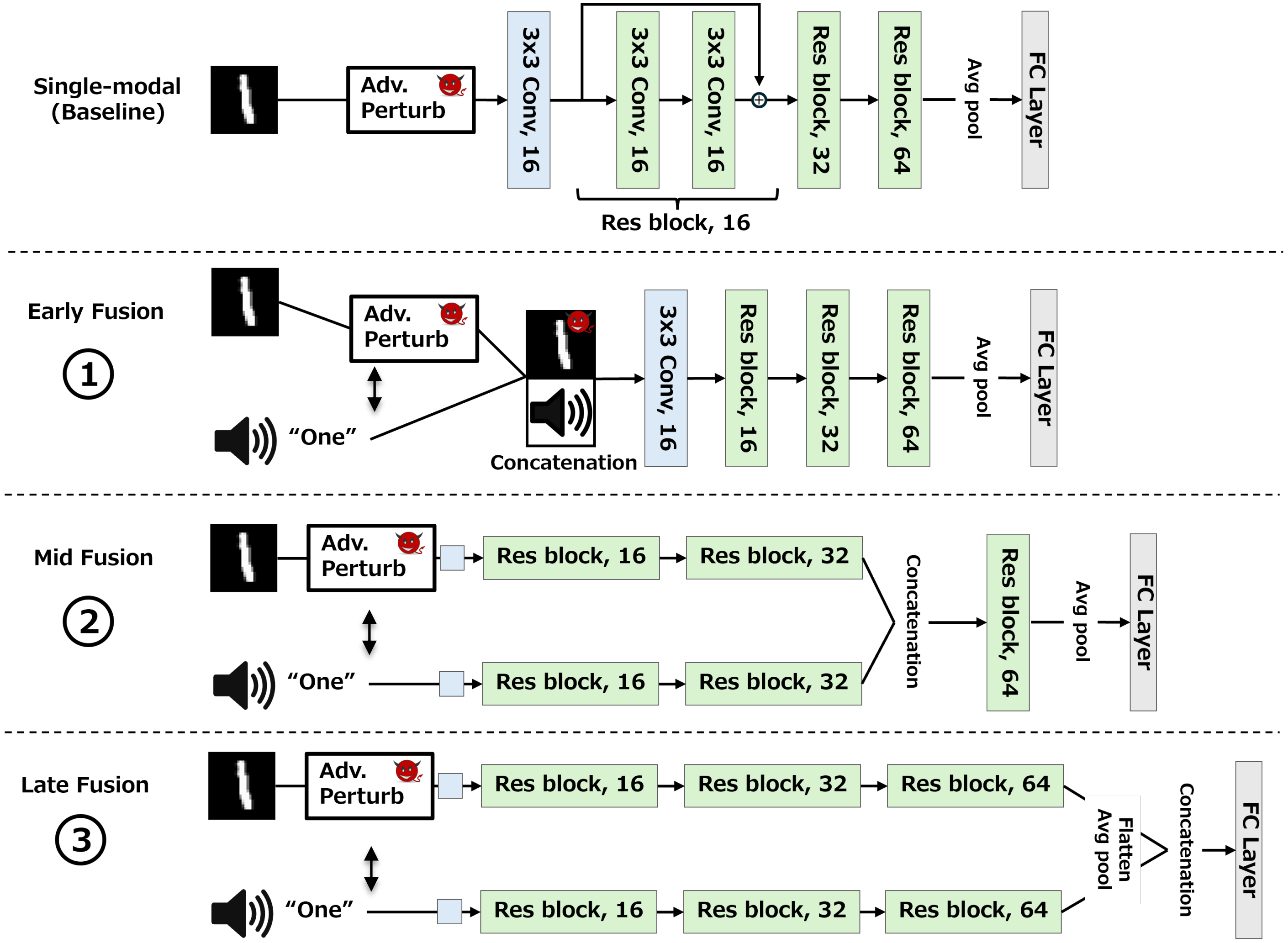
Model: Resnet 8

Modalities: Audio, Image

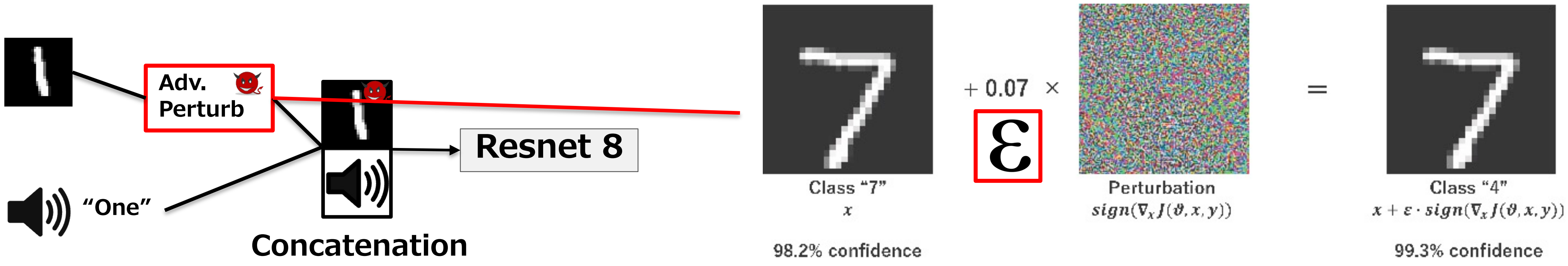
Attacks:
FGSM and PGD

For each Fusion Type:

- Apply Adv. to both modality
- Apply Adv. to image
- Apply Adv. to audio



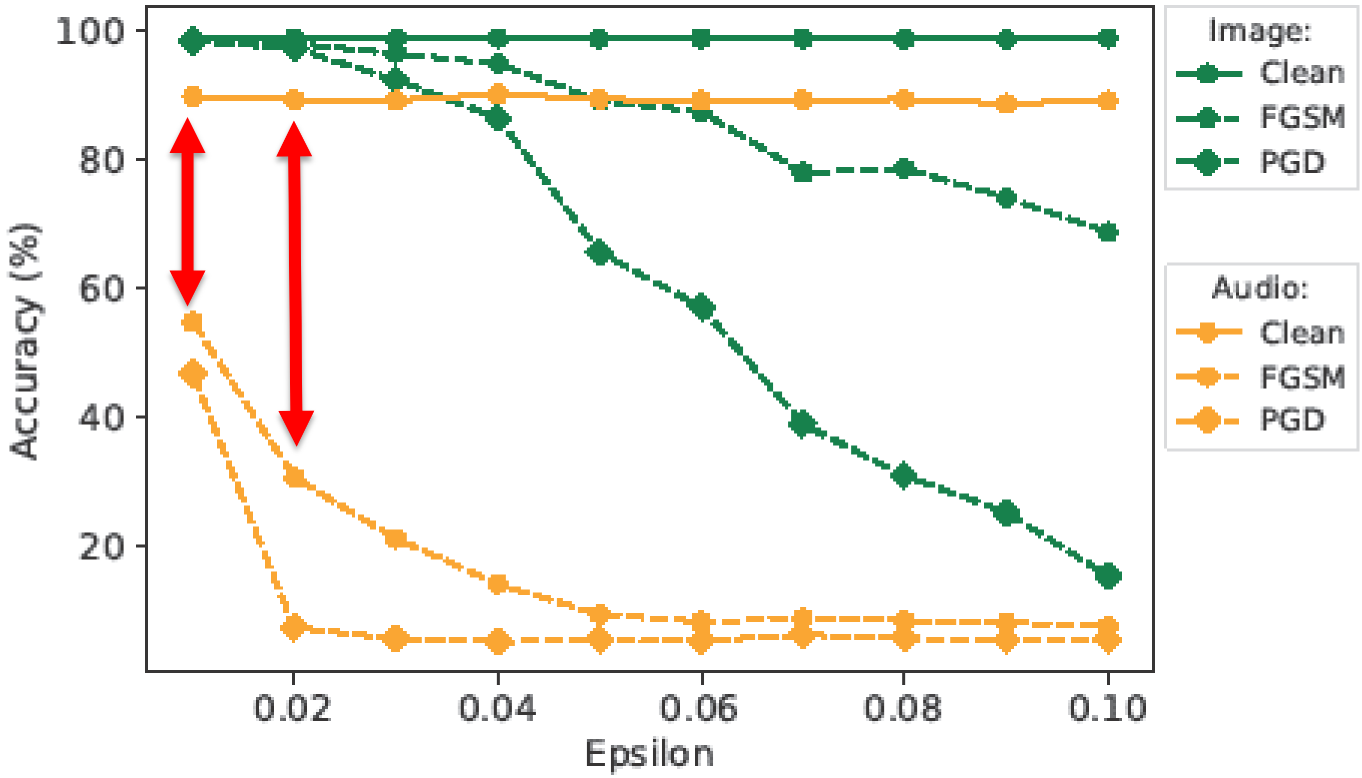
Case Study 1: Datasets and Attack Methods



- Image data: MNIST dataset (70000 digit images)
- Audio data: From Google Speech Commands (38908 utterances of digit)
  - Pre-processing by extracting the Mel Frequency Cepstral Coefficients (MFCC)
- Adv. Attacks: Fast Gradient Sign Method (FGSM), Projected Gradient Decent (PGD)
  - Explore epsilon values from 0.01 to 0.1

Case Study 1: Results & Analysis

Baseline (Single-modal model)

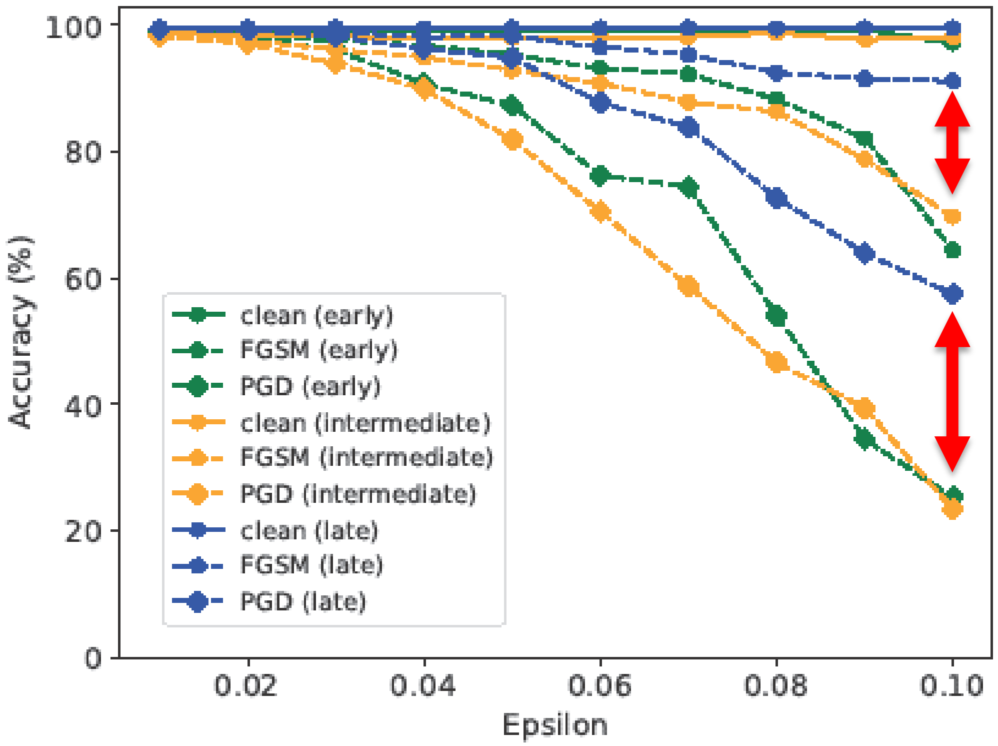


- Model trained on audio shows large accuracy degradation by FGSM and PGD
- Model trained on image shows much less degradation (at lower epsilon values)

Case Study 1: Results & Analysis

- Late fusion (Blue):
  - **Sustain** its accuracy for higher epsilon values
- Early (Green) and Intermediate (Yellow) fusion:
  - Accuracy is **degraded** more than late fusion

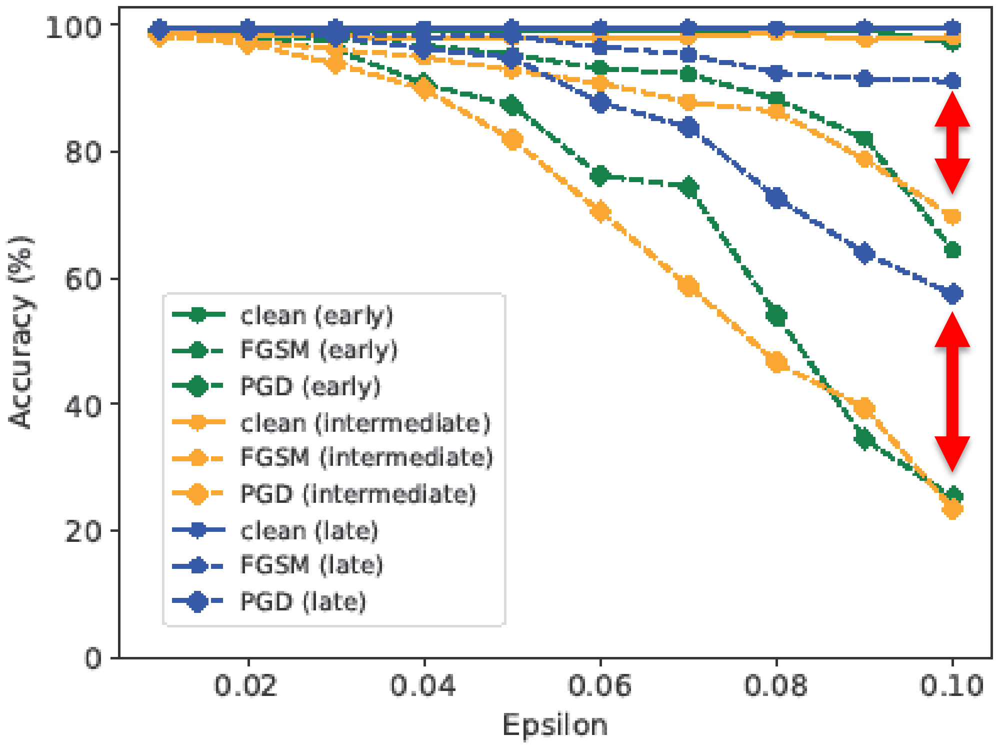
Attacks on Image Modality



Case Study 1: Results & Analysis

- Late fusion (Blue):
  - **Sustain** its accuracy for higher epsilon values
- Early (Green) and Intermediate (Yellow) fusion:
  - Accuracy is **degraded** more than late fusion

Attacks on Image Modality



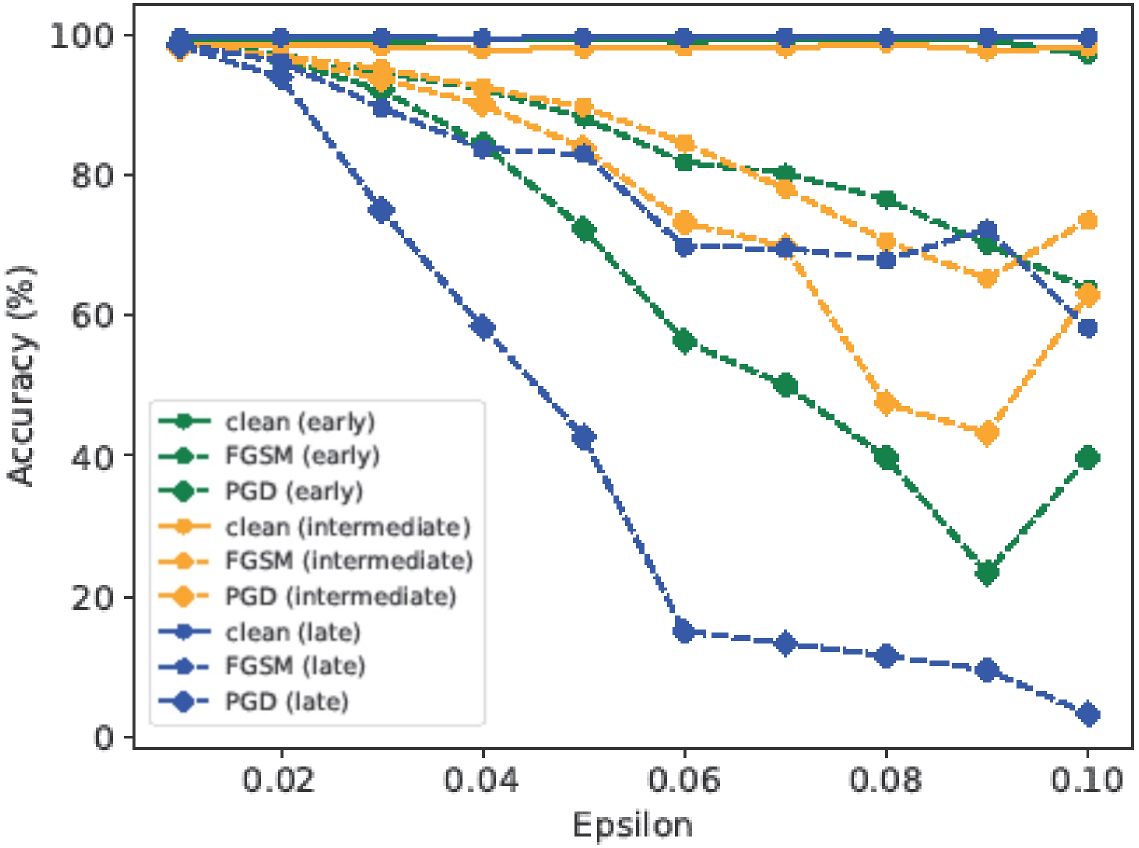
Observations:

- Late fusion appears more robust to adversarial attacks
- Previous research has shown early fusion can enhance accuracy (K. Gadzicki et al.)
- **Consider trade-off between accuracy and robustness based on fusion depth**

Case Study 1: Results & Analysis

- Late fusion (Blue) seems particularly weak against PGD attack on audio modality
- Intermediate fusion (Yellow) appears more robust against the PGD attack than the early and late fusion models.
- Fusion architecture may have some impact on model robustness to single-modal attack strategies.

Attacks on Audio Modality

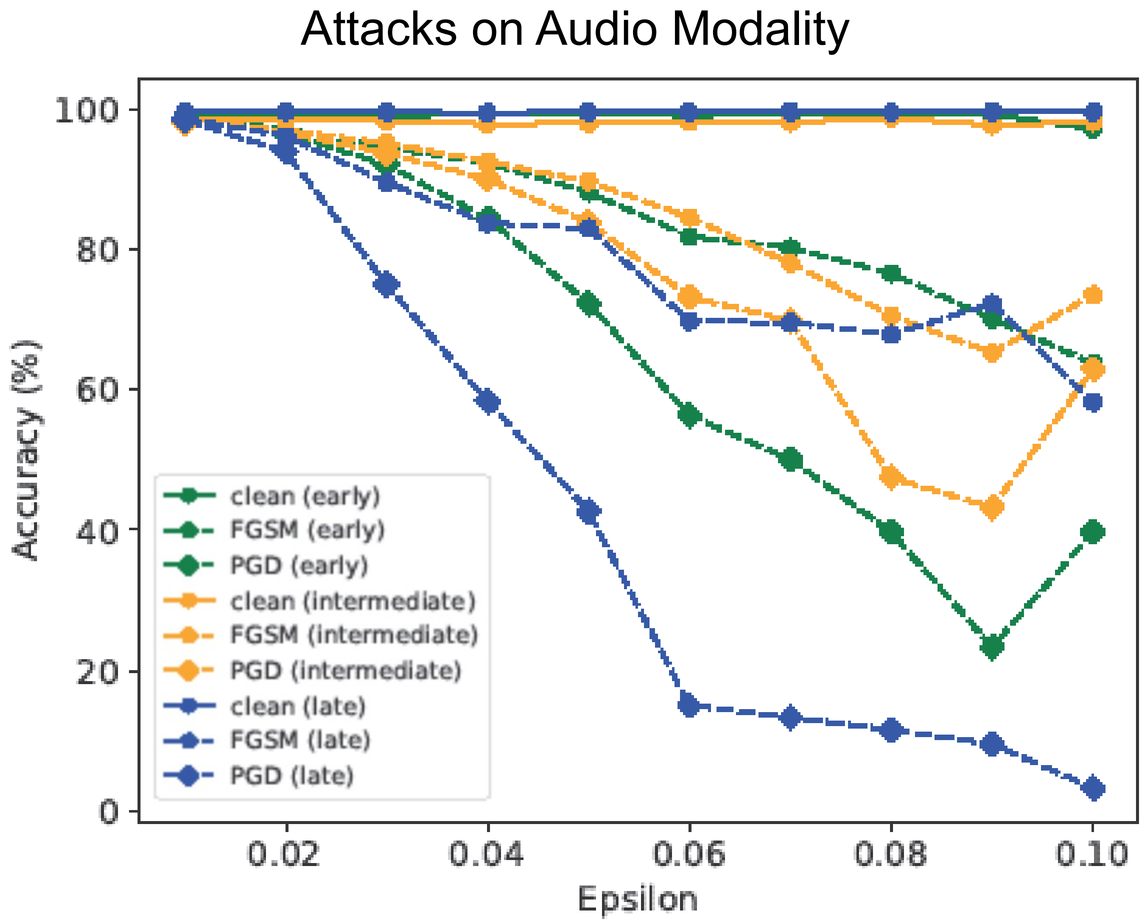


Case Study 1: Results & Analysis

- Late fusion (Blue) seems particularly weak against PGD attack on audio modality
- Intermediate fusion (Yellow) appears more robust against the PGD attack than the early and late fusion models.
- Fusion architecture may have some impact on model robustness to single-modal attack strategies.

This result also connects to the case study 2

- A susceptible modality can degrade robustness against adversarial attack

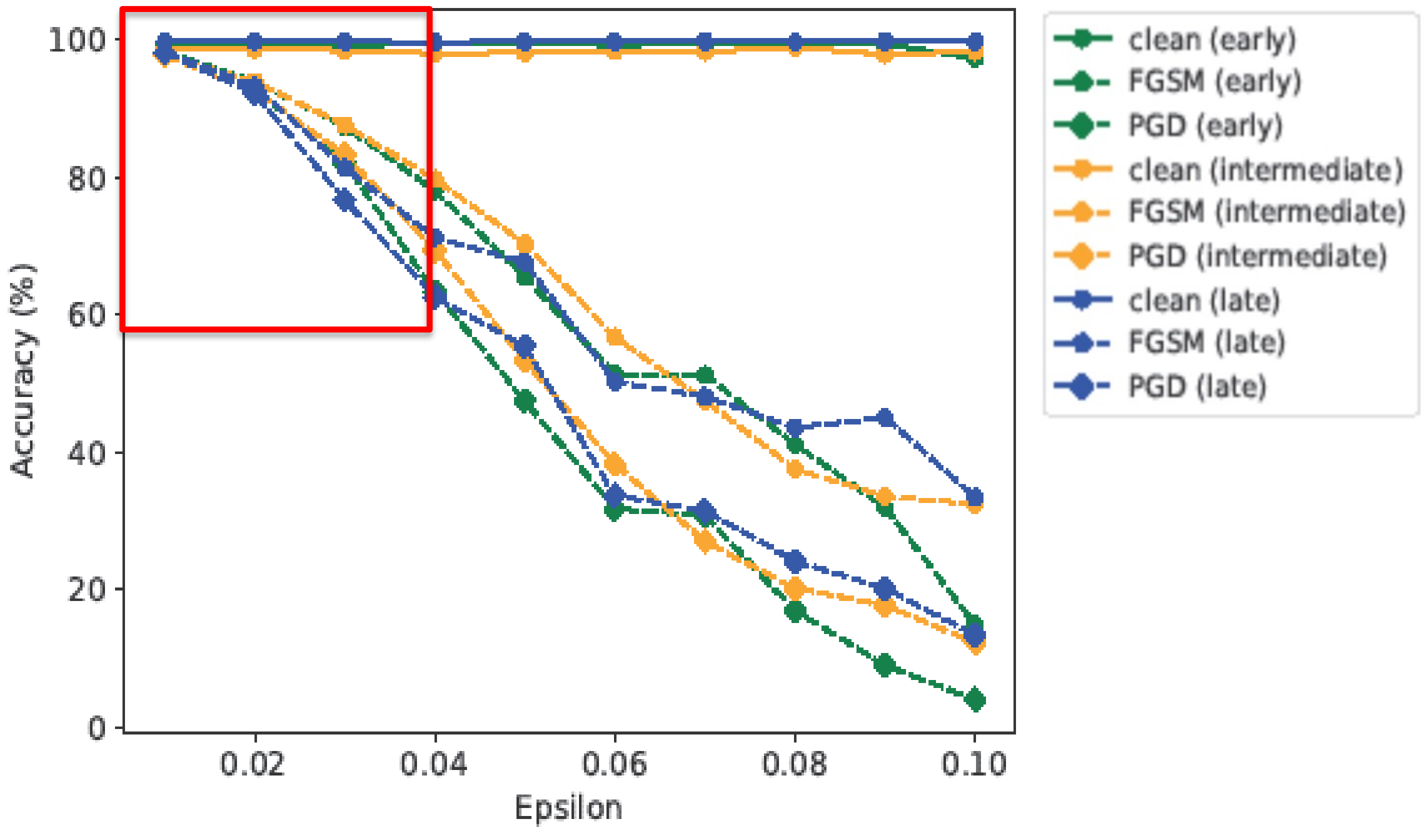


Case Study 1: Results & Analysis

- **Unsurprising result**: multi-modal attacks resulted in greater accuracy degradation because the multi-modal attacks could perturb both input modalities

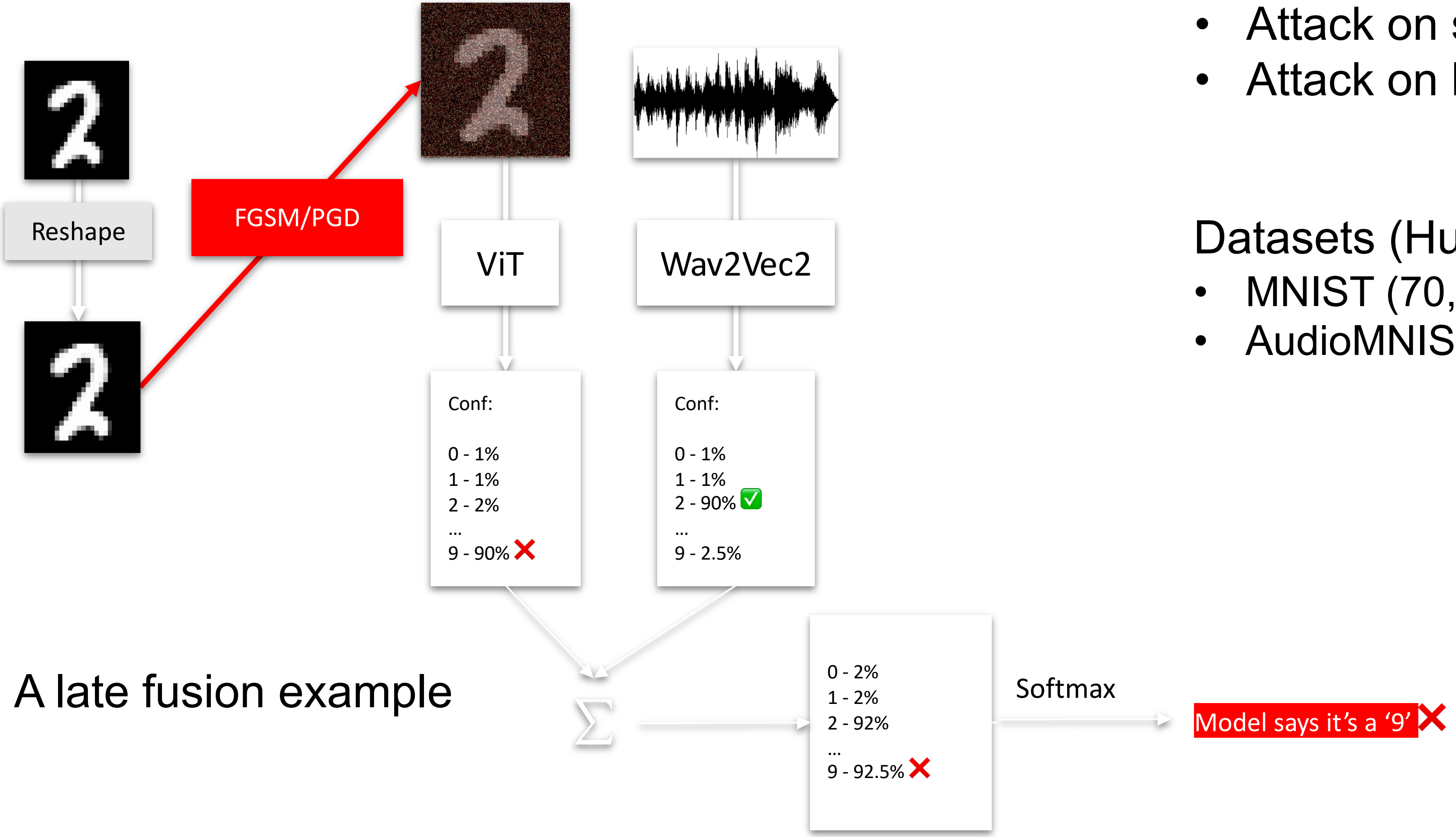
- Fusion still improves the robustness of the model when comparing to single-modal models (slide 22) at lower epsilon values

Attacks on Both Modalities (Multi-modal Attacks)



Case Study 1: Transformer-based Evaluations

Evaluations of Transformer-based architectures: early, mid, and late fusion models



A late fusion example

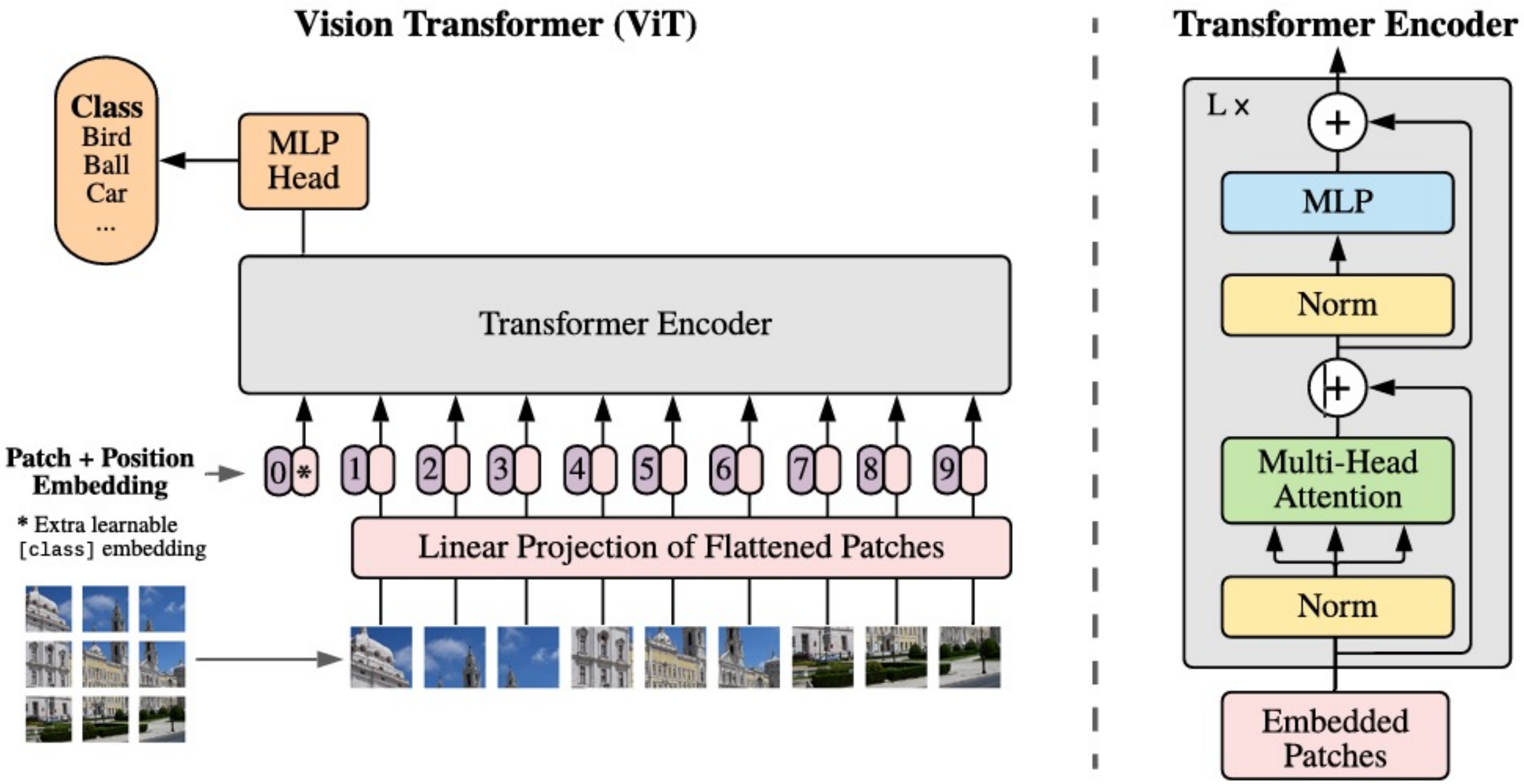
- Attack on single modality
- Attack on both modalities

Datasets (Hugging Face):

- MNIST (70,000 digit images)
- AudioMNIST (750 wav files)

Case Study 1: Vision Model

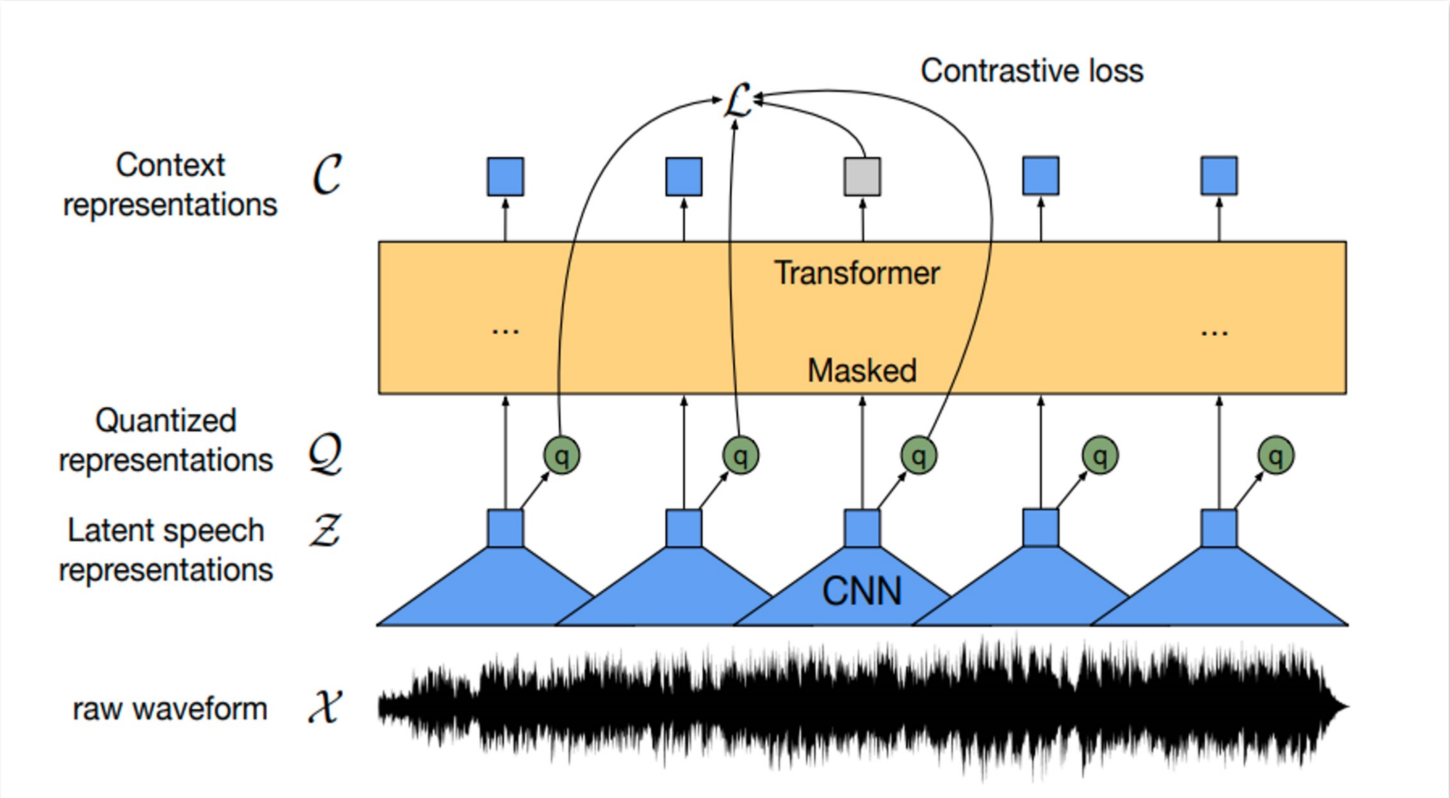
Image: Google ViT



Source : 2010.11929v2.pdf (arxiv.org), Google Research, ICLR 2021.

Case Study 1: Audio Model

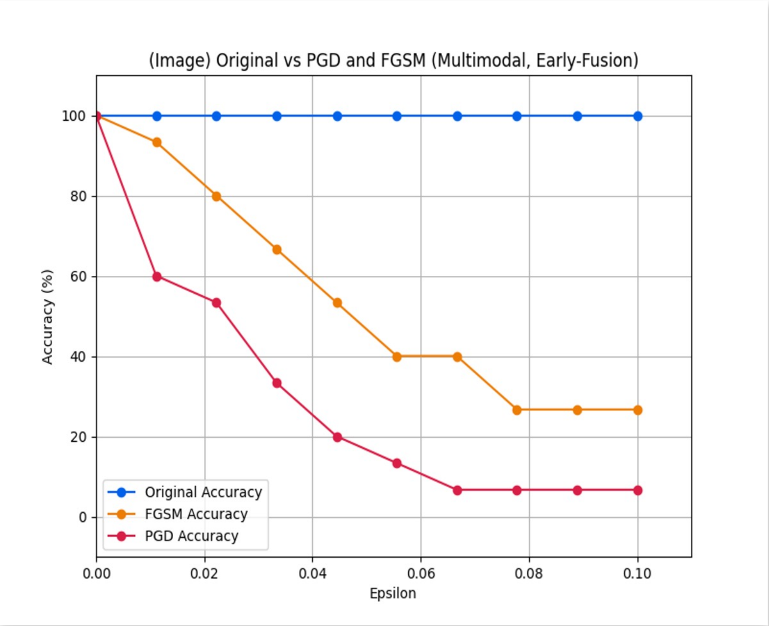
Audio: Wav2Vec2



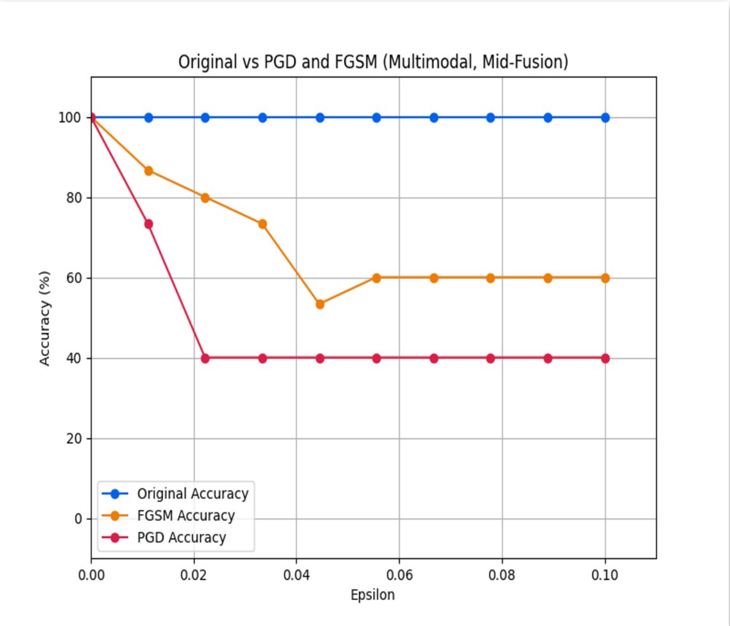
A. Baevski, et al, “wav2vec 2.0: A framework for self-supervised learning of speech representation, NeurIPS 2020.

Case Study 1: Results & Analysis

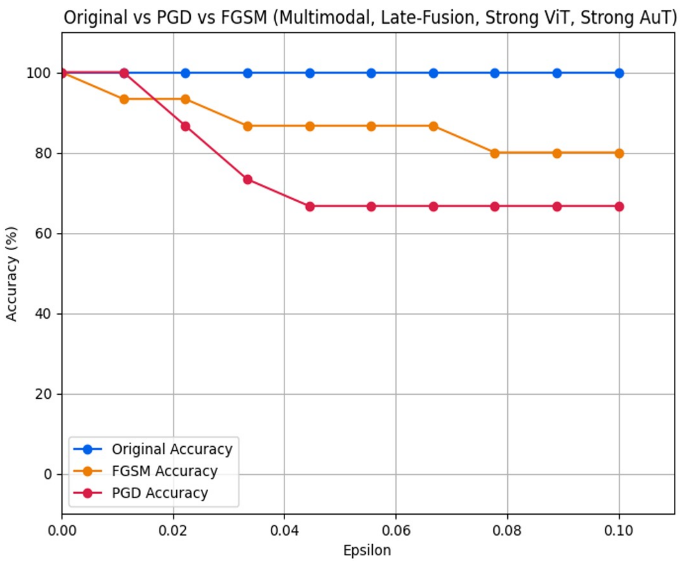
Attacks on Image Modality



Early



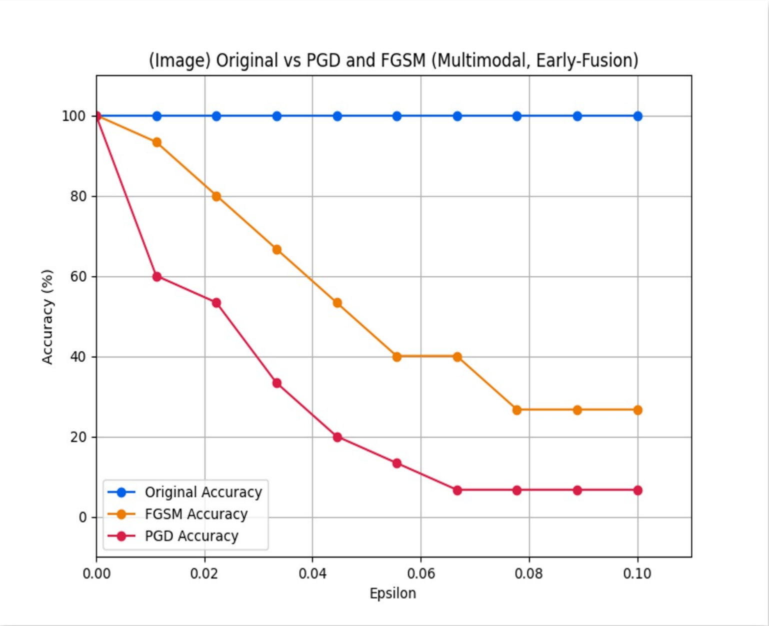
Mid



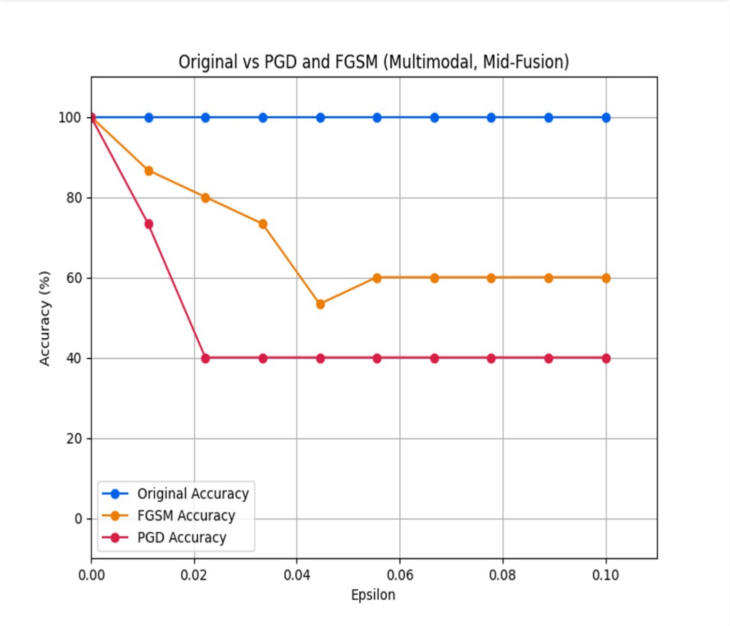
Late

Case Study 1: Results & Analysis

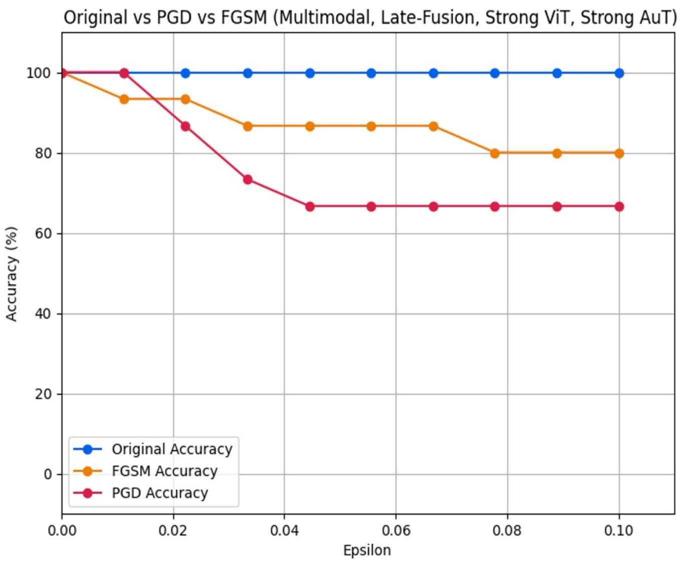
Attacks on Image Modality



Early



Mid

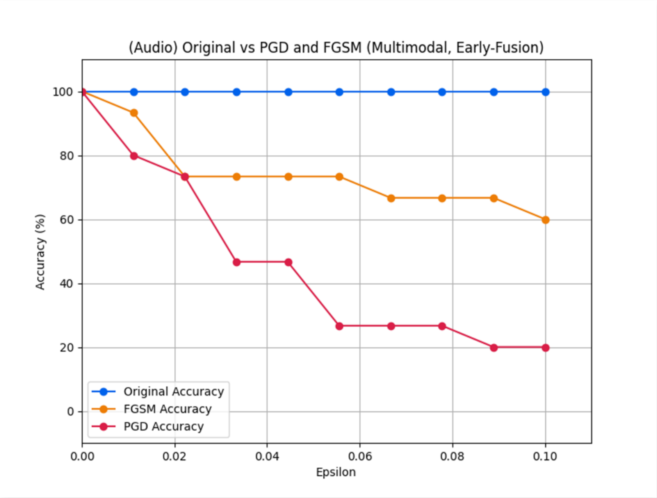


Late

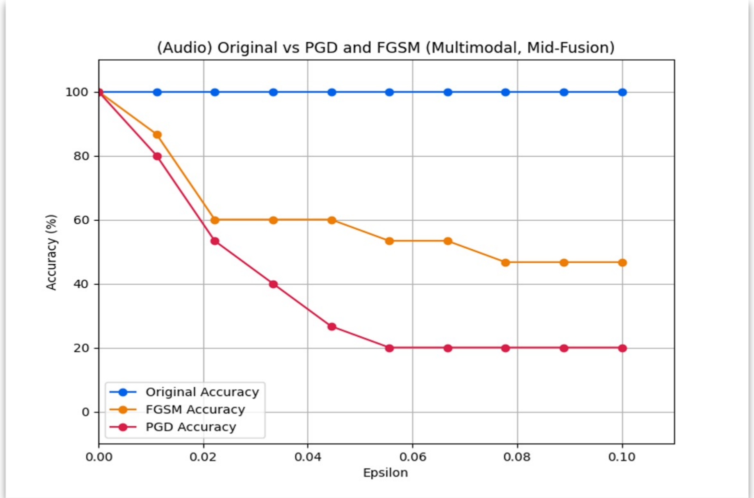
- Similar to the CNN architectures, late fusion is better than early or mid fusion for attack on image modality

Case Study 1: Results & Analysis

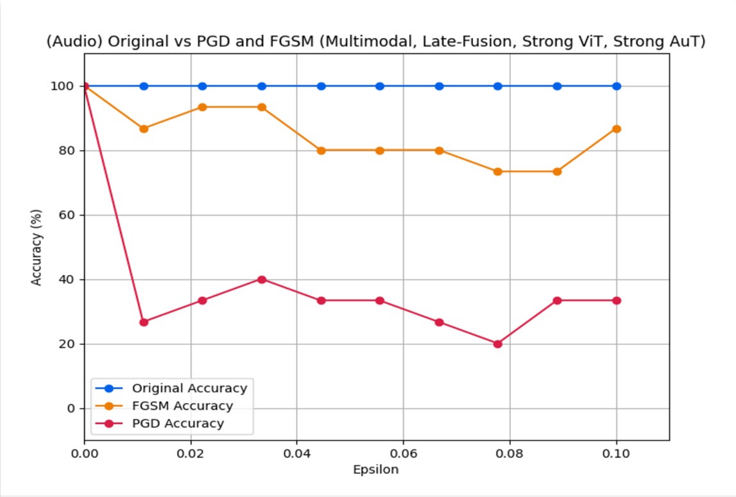
Attacks on Audio Modality



Early



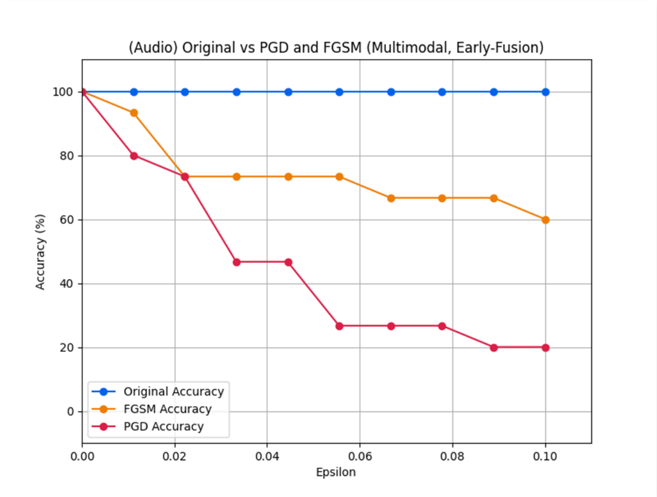
Mid



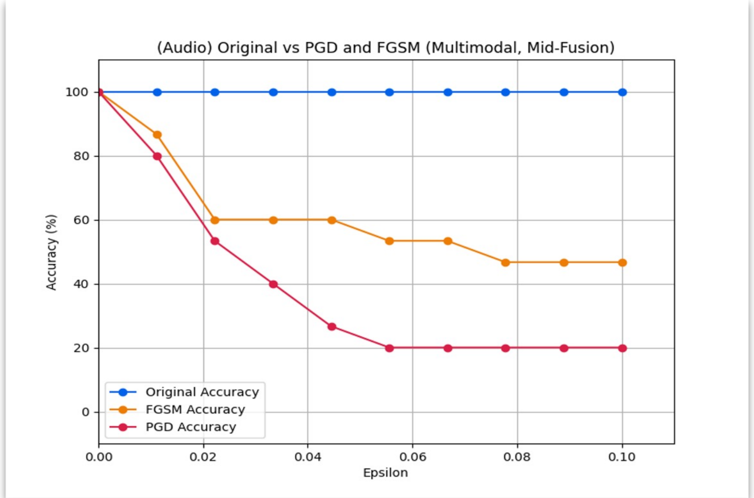
Late

Case Study 1: Results & Analysis

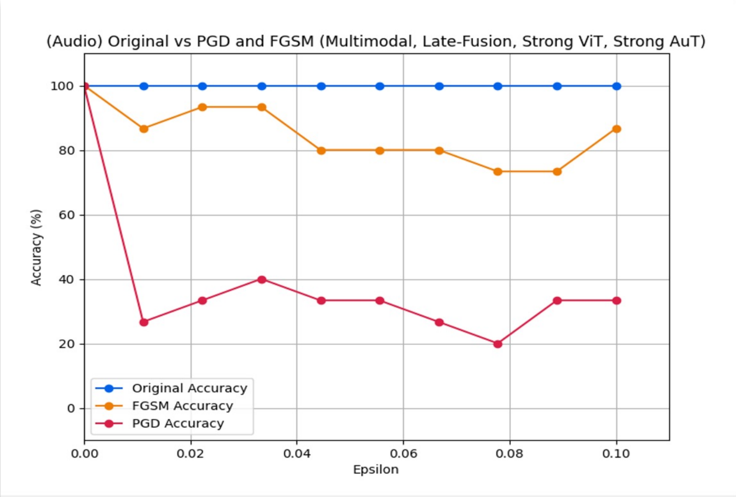
Attacks on Audio Modality



Early



Mid

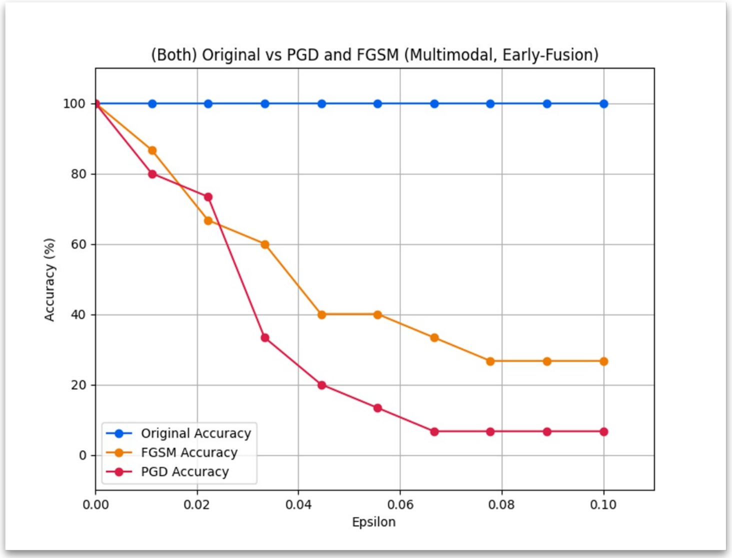


Late

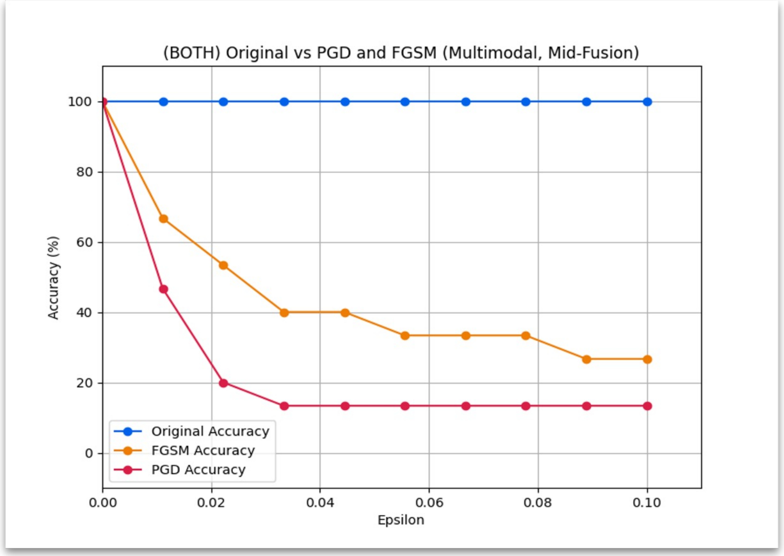
- For audio attacks, late fusion is slightly better than early or mid fusion strategies.
- Similar to the CNN experiments, audio modality seems more susceptible to attacks comparing to image modality, at least for mid and late fusion architectures.

Case Study 1: Results & Analysis

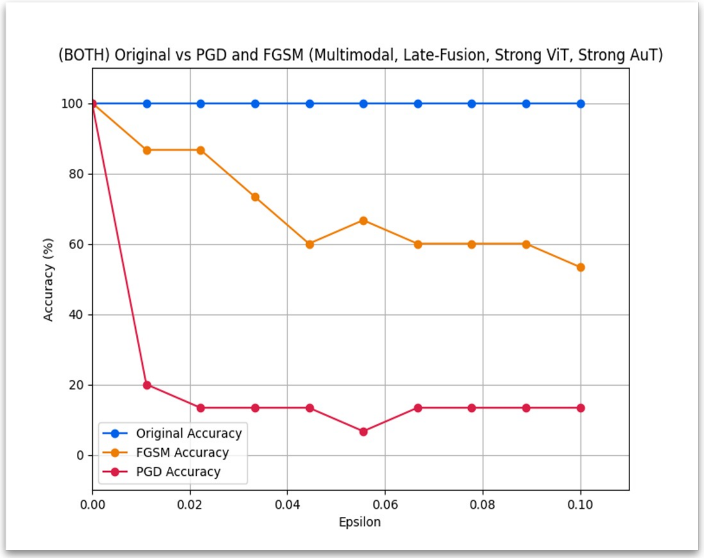
Attacks on Both Modalities



Early



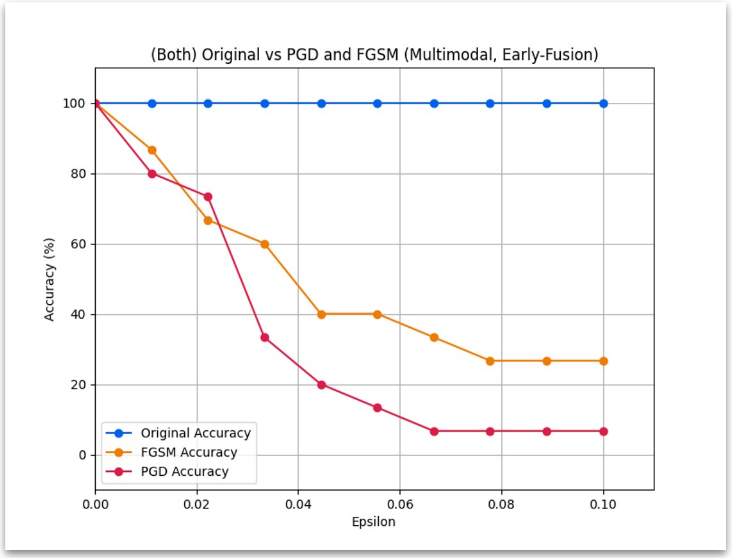
Mid

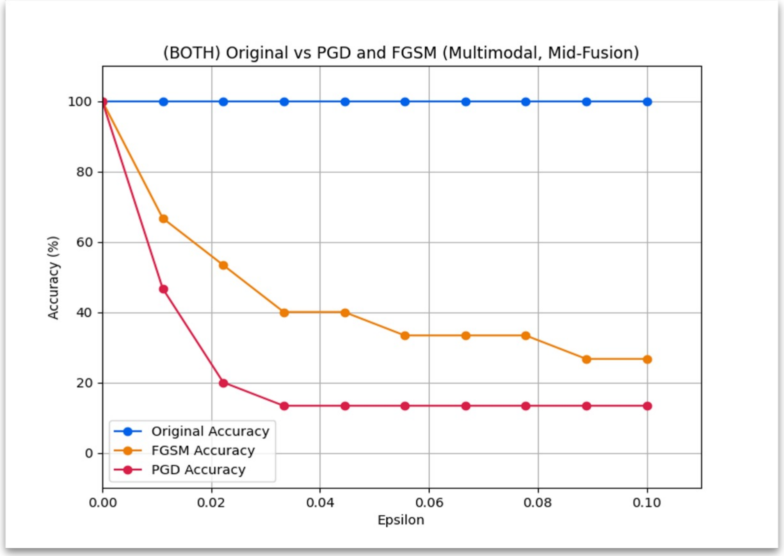


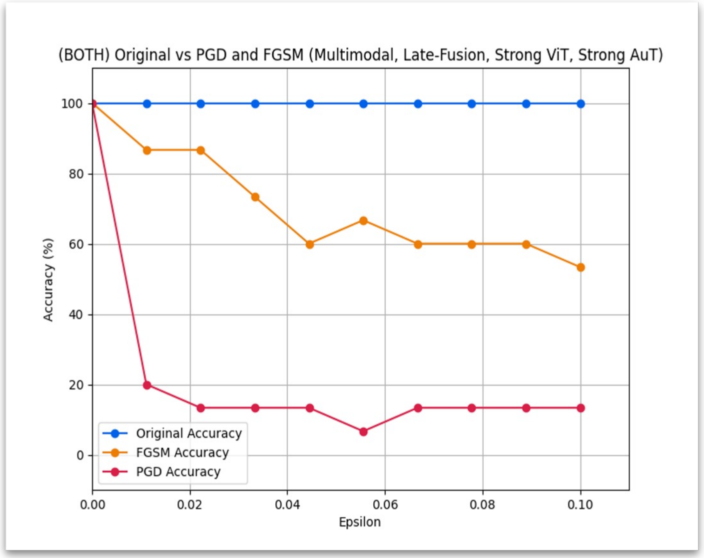
Late

Case Study 1: Results & Analysis

Attacks on Both Modalities







Early

Mid

Late

- For multimodal attacks, late fusion is still better than early or mid fusion, particularly for FGSM attacks.
- Again, multi-modal attacks resulted in greater accuracy degradation because the multi-modal attacks could perturb both input modalities.

Case Study 1: Results & Analysis

Transformer-based models

- In this experiment, late fusion appears more robust to adversarial attacks on single modality (image or audio).
- When compared to image-only or audio-only attack, multi-modal attack seem to result in greater accuracy degradation. This is consistent with earlier findings that multi-modal attacks perturb both input modalities.
- Again, need to consider trade-off between accuracy and robustness based on fusion depth.

Research Questions

- Question 1: Does fusion depth in a ML model impact robustness, particularly to single-modal attacks?
- Question 2: Can the inclusion of data modalities that are more vulnerable to perturbation make a model less robust to adversarial attacks?
- Question 3: Does the impact of quantization on model robustness differ by data modality?

Case Study 2: Overview

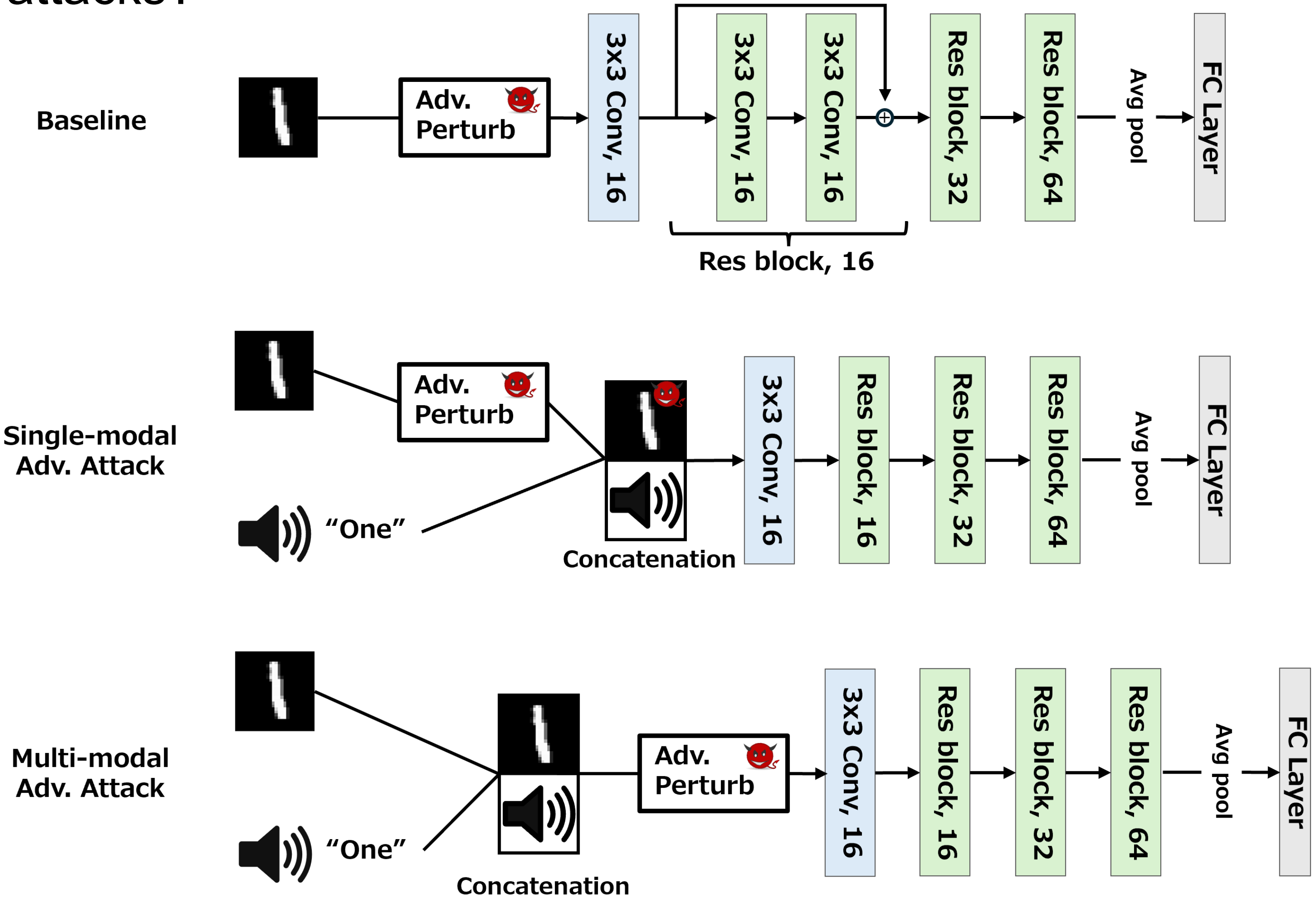
Question 2: Can the inclusion of data modalities that are more vulnerable to perturbation make a model less robust to adversarial attacks?

Modality: Audio (susceptible), Image

Attacks: FGSM and PGD

Fusion Types: Early, Intermediate, Late Fusion

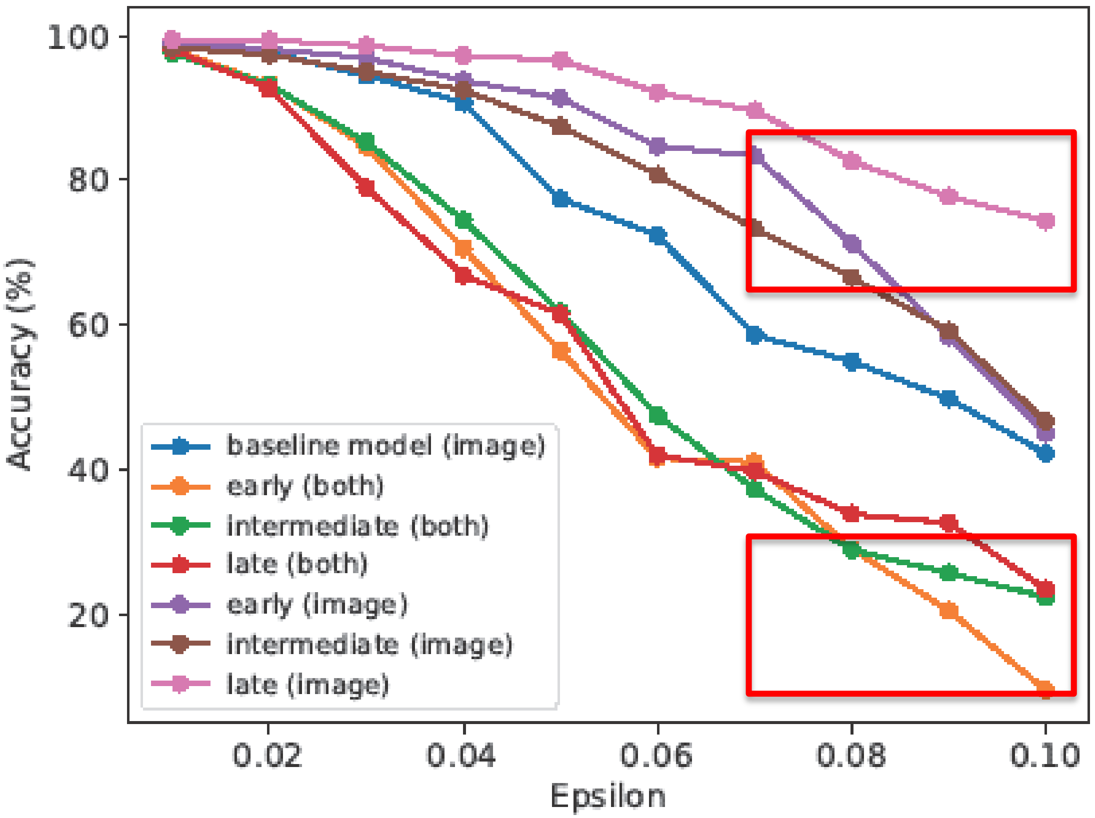
Evaluation:
Compare single and multi modal attack results



Case Study 2: Results & Analysis

- Attacking only image modality (Purples): Accuracy is **higher** than baseline (Blue) as adding audio helps improve robustness

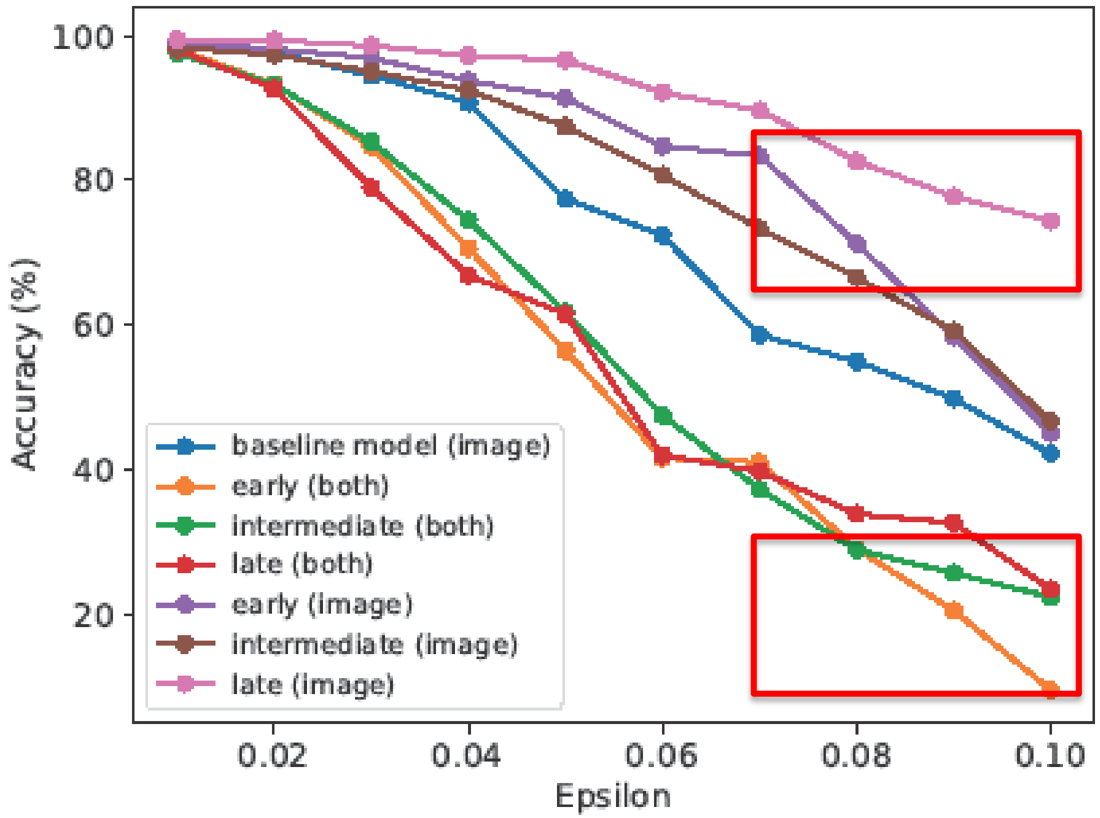
- Attacking on both modalities (Red, Green, Yellow): Accuracy is **lower** than baseline as audio is more suspectable to adversarial attacks



Case Study 2: Results & Analysis

- Attacking only image modality (Purples): Accuracy is **higher** than baseline (Blue) as adding audio helps improve robustness

- Attacking on both modalities (Red, Green, Yellow): Accuracy is **lower** than baseline as audio is more suspectable to adversarial attacks



Observations:

- A new susceptible modality can degrade resistance to multi-modal adversarial attacks
- **A counterexample to the conventional view that fusion inherently improves robustness**

Research Questions

- Question 1: Does fusion depth in a ML model impact robustness, particularly to single-modal attacks?
- Question 2: Can the inclusion of data modalities that are more vulnerable to perturbation make a model less robust to adversarial attacks?
- Question 3: Does the impact of quantization on model robustness differ by data modality?

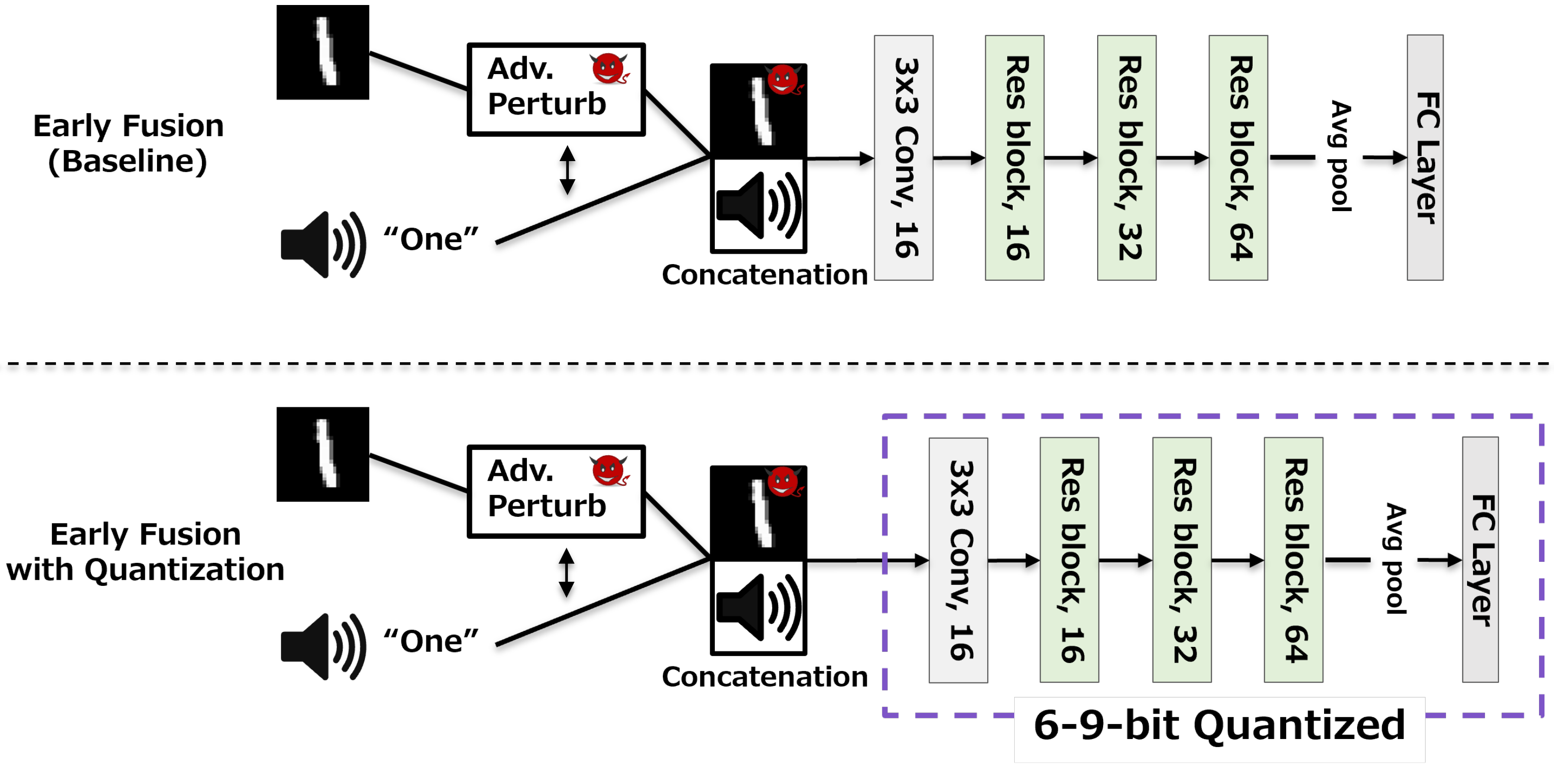
Case Study 3: Overview

Question 3: Does the impact of quantization on model robustness differ by data modality?

Modalities: Audio, Image

Attacks:
FGSM and PGD
(Single-modal attack)

Fusion Type: Early Fusion

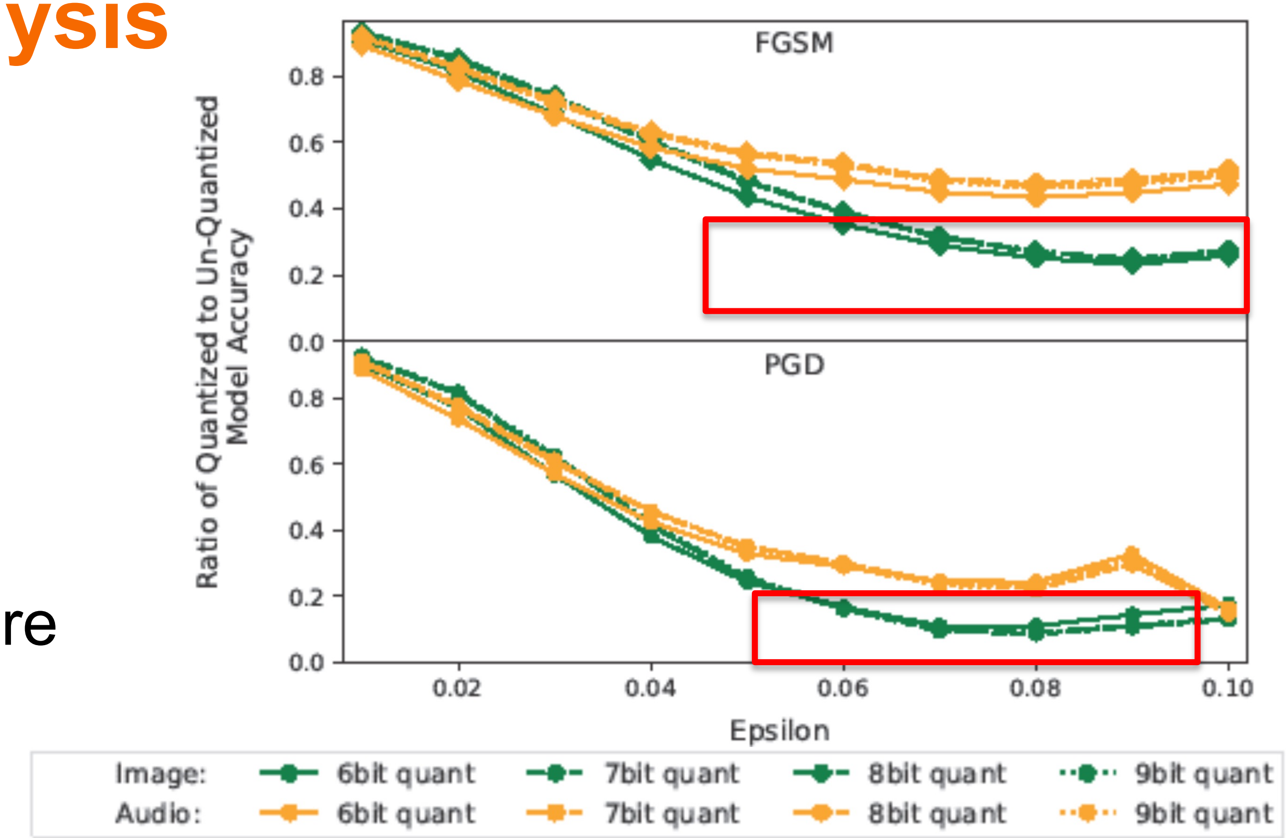


Quantization Technique: Quantization with min-max scaling (for each layer)

Evaluation: Compare Adv. Attacks on quantized and un-quantized early fusion models

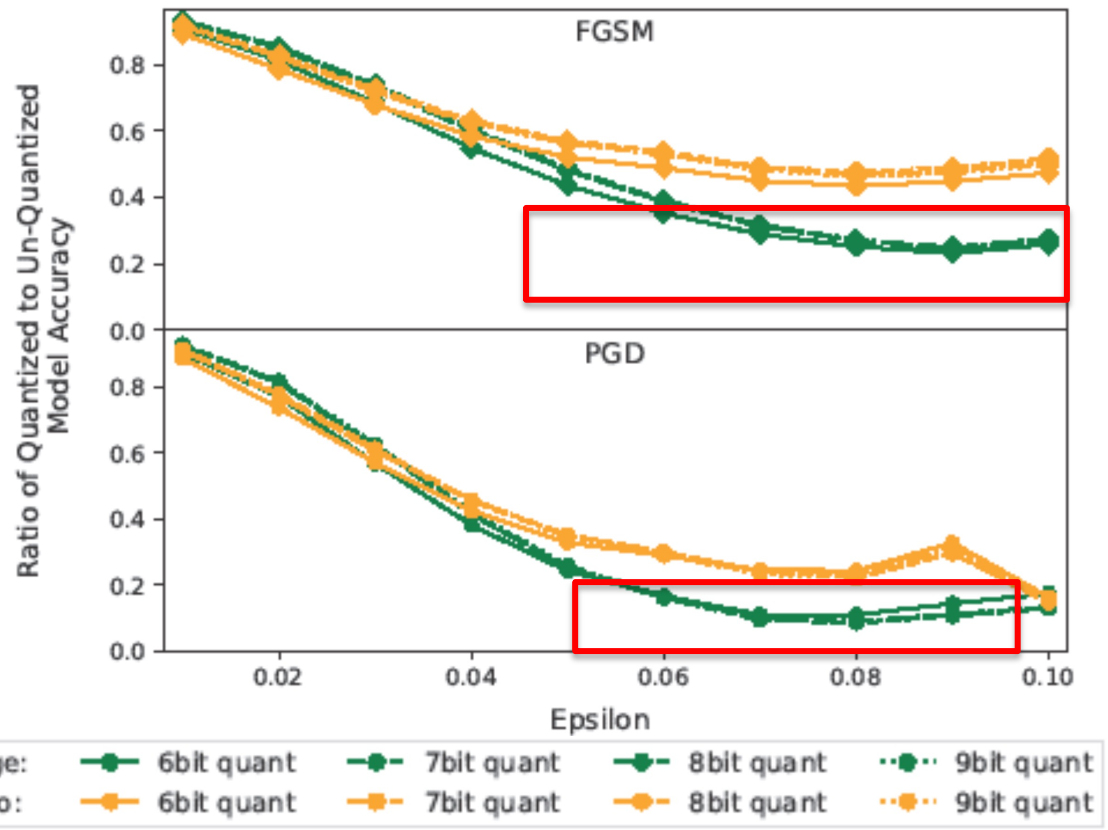
Case Study 3: Results & Analysis

- Attacks on audio modality (Yellow):
- Attacks on image modality (Green) :
  - Quantization reduced adversarial robustness in the **image** modality more



Case Study 3: Results & Analysis

- Attacks on audio modality (Yellow):
- Attacks on image modality (Green) :
  - Quantization reduced adversarial robustness in the **image** modality more



Observations:

- Quantization impacts model robustness differently across data modalities
- **Modality-dependent quantization algorithms could benefit multimodal ML applications**

Key Takeaways

Key Takeaways

Case study 1: Fusion strategy impacts adversarial robustness to single-modal attacks and this result appears to differ by data modality

⇨ The available modalities may be relevant when selecting a fusion strategy

Key Takeaways

Case study 1: Fusion strategy impacts adversarial robustness to single-modal attacks and this result appears to differ by data modality

⇨ The available modalities may be relevant when selecting a fusion strategy

Case study 2: The robustness of multi-modal models against multi-modal adversarial attacks is limited by the more vulnerable to attack modality

⇨ A counterexample to the view that fusion inherently improves robustness

Key Takeaways

Case study 1: Fusion strategy impacts adversarial robustness to single-modal attacks and this result appears to differ by data modality

⇨ The available modalities may be relevant when selecting a fusion strategy

Case study 2: The robustness of multi-modal models against multi-modal adversarial attacks is limited by the more vulnerable to attack modality

⇨ A counterexample to the view that fusion inherently improves robustness

Case study 3: Robustness to adversarial perturbations differs not only by data modality, but also by the level of quantization applied to the modality

⇨ Quantization in multimodal ML apps should consider quantization by modality

Future Work

Future Work

- Robustness of ML/DL architectures against adversarial attacks on a broader range of modalities.

Future Work

- Robustness of ML/DL architectures against adversarial attacks on a broader range of modalities.
- Implications of fusion architectures (depth) against different attack strategies (single-modal, multi-modal) for advanced ML/DL models and applications.

Future Work

- Robustness of ML/DL architectures against adversarial attacks on a broader range of modalities.
- Implications of fusion architectures (depth) against different attack strategies (single-modal, multi-modal) for advanced ML/DL models and applications.
- Adversarial robustness of candidate modalities considering the relative difficulty of performing adversarial perturbation to a candidate data modality.

Future Work

- Robustness of ML/DL architectures against adversarial attacks on a broader range of modalities.
- Implications of fusion architectures (depth) against different attack strategies (single-modal, multi-modal) for advanced ML/DL models and applications.
- Adversarial robustness of candidate modalities considering the relative difficulty of performing adversarial perturbation to a candidate data modality.
- Modality-dependent quantization algorithms and strategy.

Future Work

- Robustness of ML/DL architectures against adversarial attacks on a broader range of modalities.
- Implications of fusion architectures (depth) against different attack strategies (single-modal, multi-modal) for advanced ML/DL models and applications.
- Adversarial robustness of candidate modalities considering the relative difficulty of performing adversarial perturbation to a candidate data modality.
- Modality-dependent quantization algorithms and strategy.
- Mitigation techniques, e.g., data augmentation, regularization, adversarial training.

Future Work

- Robustness of ML/DL architectures against adversarial attacks on a broader range of modalities.
- Implications of fusion architectures (depth) against different attack strategies (single-modal, multi-modal) for advanced ML/DL models and applications.
- Adversarial robustness of candidate modalities considering the relative difficulty of performing adversarial perturbation to a candidate data modality.
- Modality-dependent quantization algorithms and strategy.
- Mitigation techniques, e.g., data augmentation, regularization, adversarial training.
- Digital-space attacks vs. physical-world attacks.

Acknowledgement

- Katsuaki Nakano
- Michael Zuzak
- Renaaron Ellis
- Cory Merkel
- Dongfang Liu

References

- K. Nakano, M. Zuzak, C. Merkel, and A. Loui, “Trustworthy and robust machine learning for multimedia: challenges and perspectives,” *Proc. IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR’24),* San Jose, CA, August 7-9, 2024.
- Z. Cheng, J. Liang, H. Choi, G. Tao, Z. Cao, D. Liu, and X. Zhang, “Physical Attack on Monocular Depth Estimation with Optimal Adversarial Patches,” *ECCV 2022*.
- Z. Cheng, H. Choi, S. Feng, J. Liang, G. Tao, D. Liu, M. Zuzak, and X. Zhang, “Fusion Is Not Enough: Single Modal Attacks on Fusion Models for 3D Object Detection,” *ICLR 2024*.
- K. Gadzicki, R. Khamsehashari, and C. Zetzsche, “Early vs late fusion in multimodal convolutional neural networks,” *Proc. IEEE 23rd international conference on information fusion (FUSION), 2020.*
- C. Qi, L. Yi, H. Su, L. Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space,” *arXiv:1706.02413v1,* June 2017.
- B. Wells, and A. Loui, "Comparing Modality performance for Deep Video Summarization", *IEEE Western NY Image and Signal Processing Workshop*, Rochester NY, Oct. 4, 2019.
- J. Abru, C. Cassidy, J. Kubeck, J. Laos, M. McGarvey, R. Ptucha, and A. Loui, “The advancement of autonomous vehicle navigation,” *Proc. 2020 ASEE Annual Conference*, Rochester, NY, Apr. 3-4, 2020.