

The Adversarial Attacks towards Human Action Recognition Models: A Comparison between unimodal and multimodal Models

Jizheng Li^{1,*}

¹ Victoria Shanghai Academy, Hongkong, China

*Corresponding author: ps20205020@student.vsa.edu.hk

Abstract:

Human-action recognition models are neural networks that analyse visual inputs and provide classification or text outputs. This technology has and will significantly impact society in security, education, healthcare, etc. However, Human-Action Recognition models, like other neural networks, are still susceptible to malicious adversarial attacks. Therefore, this paper proposes an experimental adversarial attack towards ResNet-18 using FGSM. First, ResNet-18 is finetuned using the UCF-101 dataset, and keyframes are selected from sample videos. The keyframes will be given to ResNet-18 for classification while FGSM will be implemented, and ResNet will do another classification of the attached sample. The classification results (Original and Attacked) are given to the language model (GPT-4o) through a prompt that provides the language model with a specific role (e.g. a smart home assistant), and this section is regarded as unimodal. The original and attacked frames will be sent directly instead of the labels in the multimodal section. Lastly, this paper proposes to observe the effects on textual responses generated based on a given prompt and the classification result and evaluate the impact of the attack through Cosine Similarities and Human Evaluation.

Keywords: FGSM; Adversarial Attacks; Human-Action Recognition; Deep Learning.

1. Introduction

With the advancements in artificial intelligence (AI), particularly the rapid progress in the Large Language Model (LLM), the Human-Action Recognition (HAR) technology, a subset of AI, plays a more and more critical role in enabling seamless human and machine interactions [1]. It has extensive application fields, including security, education, sports, healthcare, entertainment, etc. The HAR technology has redefined and enhanced user experiences by accurately interpreting human actions, such as gestures and movements. Therefore, it has become a key technology in various user cases when AI/LLMs are implemented, and this is particularly true when an increasing number of multimodal LLMs become publicly available.

However, adversarial attacks aiming towards neural networks and HAR technology could lead to inaccuracies or, even worse, false results, raising significant safety concerns. Therefore, it is critical to implement vigorous adversarial attack tests and assess the robustness of the models, ensuring the models can withstand possible malicious attacks. Furthermore, this requires considering adversarial attacks in the model development and training stages to lead to more robust and efficient models.

Adversarial attacks can be implemented through several

modalities, while the most common is through vision. There are two kinds of adversarial attacks: The White box describes when the attacker has the model's internal model architecture, parameters, and gradients such that attacks like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) can be implemented while the Black box indicates that the attacker has no information on the internal details of the model, therefore using the transferability from attack methods for similar models [1]. Recent advancements in HAR technology have been significantly driven by the developments of adversarial attacks, resulting in a substantial impact on the design and evaluation of machine learning models, particularly in video analysis. Liu et al. implement the Robust Visual Question Answering (RVQA) model and emphasize the critical need to address the vulnerabilities of existing HAR systems when subjected to adversarial perturbations [2]. The traditional HAR models have successfully incorporated deep learning techniques, e.g. convolutional neural networks (CNN). However, these models usually perform unsatisfactorily when subject to adversarial attacks, and even minor perturbations can compromise model predictions [3]. Meanwhile, the model accuracy can also be dramatically reduced by adversarial examples through the attack technique, the Fast Gradient Sign Method (FGSM), as

illustrated by Moosavian et al., revealing the importance of improving not only pre-training strategies but also the post-training evaluation [4]. Zhang et al. also highlight the gaps in current methodologies and propose leveraging advanced techniques, such as keyframe extraction and snippet selection, to enhance the overall performance of HAR models [5]. For real-world applications like security and healthcare, adversarial training strategies have been integrated into model development. Li et al. have achieved greater resilience of HAR systems [6]. Considering these challenges and progress so far, continual exploration into more robust evaluation metrics and training methodologies is urgently needed, underscoring the relevance of this research in advancing these technological capabilities, particularly from the angle of safeguarding HAR models

against adversarial attacks.

With the advancements in AI, particularly in Large Language Models (LLMs), Human-Action Recognition (HAR) technology is becoming increasingly significant. This subset of AI enhances seamless interactions between humans and machines, thus making it crucial to improve human-machine communication.

2. Methodology

The overall structure of the study follows the flowchart in Figure 1. It is important to note that only samples that can be correctly classified by ResNet-18 and its perturbed frames leading to misclassification by ResNet-18 are collected for the study.

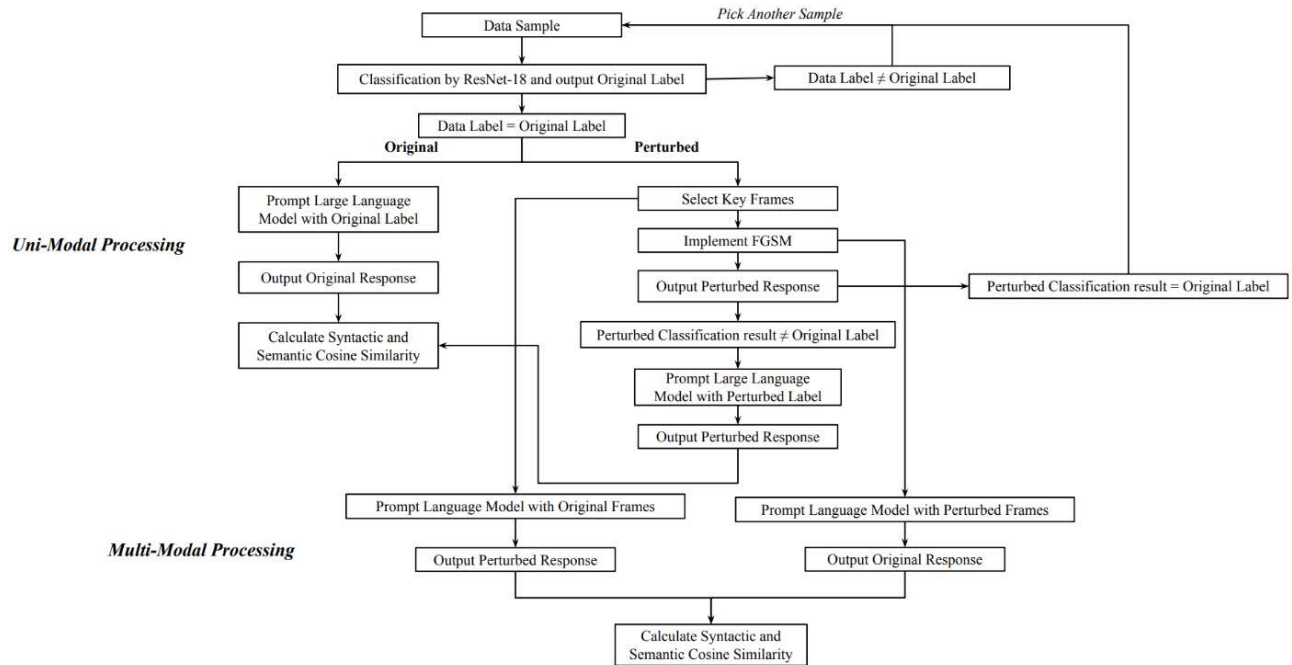


Fig. 1 A flow chart that describes the process of each data sample being collected as data.

The overall structure of the study follows the flowchart in Figure 1. It is important to note that only samples that can be correctly classified by ResNet-18 and its perturbed frames leading to misclassification by ResNet-18 are collected for the study. The classification ResNet-18 model has been fine-tuned using the UCF-101 dataset and will be discussed in the following sections. The unimodal processing indicates that the textual response from the language model is completed through a combination of original and perturbed classifications by ResNet-18 embedded into the prompt by ResNet-18. In contrast, the original and perturbed frames are sent with the prompt in multi-modal processing. Overall, the essential data collected are the textual responses that can be utilised as a reference for human evaluation and the Syntactic and Semantic Cosine

Similarity calculated between the original and attacked unimodal language model and multimodal language responses. Past studies have shown significant transferability through VLP (Vision Language Processing) models, with or without the same architecture. The transferability sets the foundation of this research, aiming to compare the effects of the attack against the unimodal and multimodal models [7].

There are 50 samples collected in total, equally divided into 2 groups, with each group containing 25 samples labelled as “BabyCrawling” and “Biking” respectively. The rationale behind this division is that they can be used in different settings of smart home assistant and road monitoring assistant.

2.1 Finetuning

The chosen dataset for this study is the UCF-101, which consists of 13320 short videos of 101 action categories. The realistic video extracts are selected from YouTube, unlike many other datasets being staged by actors. This benefits the study as the more realistic the samples are, the better they simulate our world and bring more accurate experimental results. However, this also increases the difficulty of action recognition due to the considerable variation in camera angle and motion, scale, background, lighting, etc. But overall, the fine-tuning is successful overall as it reaches 79.66% accuracy [8].

2.1.1 Data Transformation and Augmentation

Data transformation and augmentation methods have been utilized further to improve the generalising capabilities and robustness of the model. The methods are described in the following list: 1) Resizing frames from 320x240 to 224x224. 2) Random horizontal flipping. 3) Color Jittering. 4) Random rotations/rotational angles.

The size of the frames in the videos UCF-101 dataset (320x240) does not match the input size of ResNet-18 (224x224). Hence, the frames are resized, potentially improving the model's robustness. Random horizontal flipping is implemented so that the model is invariant to left-right orientation, improving its generalising capabilities. Colour jittering methods, including brightness, contrast, saturation, and hue modifications, have been utilised to randomly adjust all frames to simulate different camera qualities and lighting. Random rotations/rotational angles have been added to make the model less sensitive to object orientation. Lastly, 50 frames are extracted from each video during training, ensuring that all/nearly all the features from each category are represented within the samples.

2.1.2 Model Architecture

The base architecture for the Human-Action Classification task is a pre-trained ResNet-18 model. ResNet-18 is a Convolutional Neural Network (CNN) that is 18 layers deep and trained on ImageNet, a large-scale visual database. The model's fully connected layer was reshaped during fine-tuning to fit the 101 classes in the UCF-101 [8,9].

2.1.3 Optimization

The AdamW optimizer has been selected to boost the training process. The AdamW optimizer is set at an initial learning rate of 0.0001 and weight decay of 0.01; A cyclic learning rate scheduler was employed with the range of 0.0001 and 0.001. This substantially decreases the training time and avoids stopping at local minima [10].

Cross-entropy loss is also implemented across the 50 ep-

ochs of training. An accumulation step every 4 batches is selected to balance computational efficiency and gradient stability. The best validation accuracy of the model is selected to be the final fine-tuned version in the following adversarial attack [11].

2.1.4 Fine-tuning Results

After fine-tuning, the model's accuracy on the test set reached 79.66%. This is relatively high compared to results obtained by past studies on the same dataset: 3D ResNet-101 achieves 83.34%, and 3D ResNet-50 achieves 85.18% on the same UCF-101 dataset. Hence, the overall fine-tuning is successful as the model could make relatively accurate classifications [12].

2.2 Adversarial Attack

Fast Gradient Sign Method (FGSM) is the adversarial attack to be implemented in this study. FGSM is a white box, gradient-based attack, and it is implemented to add perturbations to the frames to mislead ResNet-18. The following formula describes FGSM:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \quad (1)$$

The x_{adv} indicates the perturbed image; x is the original image; ϵ which is the magnitude of the change towards the perturbed image; $\text{sign}()$ function returns the sign of the gradient, controlling the direction of the perturbation; $\nabla_x J(\theta, x, y)$ is the gradient of the loss function J for the input x , class label of y and parameters θ . During the attack, the loss is maximised by manipulating the gradients $\nabla_x J(\theta, x, y)$, which leads to misclassification.

Additionally, FGSM is also known for its simplicity has the benefits of being computationally efficient and is an effective method to attack visual neural networks.[13]

2.3 Large Language Model

GPT-4o by OpenAI is used in this study. GPT-4o was released on May 13, 2024. It is regarded as the state of the art for its language-generating capabilities, and it can integrate text, audio, image, and video inputs while generating text, audio, and image outputs. However, OpenAI has yet to disclose much information about GPT-4o, but GPT-4 and other GPT models generally share similar characteristics and training processes. GPT-4 is a transformer-style model that does token prediction in a document; it was also fine-tuned using Reinforcement Learning from Human Feedback (RLHF). The Transformer architecture includes an encoder-decoder structure and incorporates a self-attention mechanism, multi-head attention, Scaled Dot-Product Attention, Feed Forward layers, Normalisation layers, etc. GPT-4 shares similarities with the transformer architecture, consisting of billions/trillions of parameters, while GPT-4o is believed to have even more parameters than GPT-4. Testings were implemented upon GPT-4o through a couple of measures, and they are dis-

played in Tables 1 and 2 [14-16].

Textual testing results are obtained through the following testing methods in Table 1: MMLU: Measuring Massive Multitask Language Understanding, GPQA: A Graduate-Level Google-Proof Q&A Benchmark, MATH: Measuring Mathematical Problem Solving With the MATH

Dataset, DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs, MGSM: Multilingual Grade School Math Benchmark (MGSM), Language Models are Multilingual Chain-of-Thought Reasoners, and HumanEval: Evaluating Large Language Models Trained on Code.

Table 1. Textual testing results

	Prompt	MMLU (%)	GPQA (%)	MATH (%)	DROP (F1,3-shot) (%)	MGSM (%)	HumanEval (%)
Gpt-4o-2024-05-13	Assitant	88.7	53.1	75.9	79.8	90.0	90.2
Gpt-4o-2024-05-13	ChatGPT	88.7	53.6	76.6	83.4	90.5	90.2

Table 2. Visual Testing results from multiple visual testing sources

Eval Sets	MMMU (%) (val)	MathVista (%) (testmini)	AI2D (%) (test)	ChartQA (%) (test)	DocVQA (%) (test)	ActivityNet (%) (test)	EgoSchema (%) (test)
GPT-4o	69.1	63.8	94.2	85.7	92.8	61.9	72.2

Visual Testing results from multiple visual testing sources are shown in Table 2. MMMU: Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark, MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts, AI2D: A benchmark dataset for evaluating Multimodal’s Understanding under the Scientific Context, ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning, DocVQA: A Dataset for Vqa on Document Images, ActivityNet: Benchmark for evaluating complex Human Activity, and EgoSchema: A very long-form video

question-answering dataset.

The overall textual and visual testing results of GPT-4o are very high and reflect an accurate and capable model of handling general tasks and visual inputs. Hence, GPT-4o is a suitable choice for a language model to handle both unimodal and multimodal inputs at a constant high quality, ensuring the fairness of the experiment.

3. Results

3.1 General Analysis

Table 3. Overall, 50 samples were collected, and the results of their Cosine Syntactic and Semantic Similarities

		Mean	Variance	Standard Deviation
Unimodal	Syntactic Similarity	0.7019	0.0123	0.111
	Semantic Similarity	0.9543	0.0002	0.01377
Multimodal	Syntactic Similarity	0.8517	0.003	0.0552
	Semantic Similarity	0.9595	0.0005	0.0226

Overall, 50 samples were collected, and the results of their Cosine Syntactic and Semantic Similarities are shown in Table 3. Table 3 represents all the samples from the angle of the mean, variance and standard deviation of the syntactic and semantic similarity between the original and attacked language model responses for both unimodal and multimodal. From the results, it is evident that the mean semantic similarity for both unimodal and multimodal is approximately the same at around 95%, suggesting that the adversarial attack did not significantly impact the meanings of the paragraph when calculated using the cosine semantic similarity. Other than that, the unimodal responses’ semantic similarity seems more closely dis-

tributed to the mean as it has less variance and standard deviation than the multimodal responses’ semantic similarity. However, overall, there is no significant difference in semantics between the original and perturbed textual response when looking through the cosine similarities. However, there is a difference of 15% between the unimodal and multimodal syntactic similarity. This indicates that multimodal processing produces more consistent syntactically similar responses as it is further supported when multimodal’s syntactic similarity has less variance and standard deviation than unimodal. This divergence in syntactic similarities is more notable than in semantics, and it occurs in the 50 samples, while more data can be

collected in future studies to verify this pattern.

3.2 Scenarios and Human Evaluation

“BabyCrawling”	
Uni-Model Prompt	As a smart home security assistant, the analysis results of the baby's actions from the video footage are {original_label}, can you determine if the baby is in any dangerous situations or engaging in potentially hazardous activities that may require assistance?
Uni-Model Response (Language Model Response)	Based on the analysis results indicating that the baby's actions are categorized as BabyCrawling, it is essential to ensure the environment is safe to prevent any potential hazards. Crawling babies are naturally curious and may encounter dangerous situations such as accessing electrical outlets, reaching for small objects that could pose a choking hazard, or approaching stairs without safety gates.....
Uni-Model Response (Language Model Response)	Based on the analysis results indicating "CuttingInKitchen," it seems that the baby may be engaging in a potentially hazardous activity, as the kitchen is filled with various sharp objects and appliances that could pose a significant risk to a child's safety. Immediate attention and intervention are recommended to ensure the baby's well-being. It is crucial to remove any sharp objects, such as knives, scissors, or other cutting tools, from the baby's reach and to ensure that the child is supervised at all times while in the kitchen.....
“Biking”	
Multi-Model Prompt	As a smart home security assistant, could you please analyze the frames extracted from the video footage, identify the baby's actions, and determine if the baby is in any dangerous situations or engaging in potentially hazardous activities that may require assistance?
Multi-Model Response (Language Model Response)	As your smart home security assistant, upon analyzing the provided video frames, I can identify that the baby appears to be crawling or moving around. There is no immediate indication of the baby being in a dangerous situation or engaging in potentially hazardous activities. However, to ensure the baby's safety, it is important to keep the environment baby-proofed, which includes securing any sharp objects.....
Multi-Model Response (Language Model Response)	As your smart home security assistant, I have analyzed the provided images. The baby appears to be crawling on the floor and exploring their surroundings. In some frames, the baby is interacting with objects, possibly reaching out to touch or grab them. While these activities are normal for a baby, it is essential to ensure that the environment is safe and free of potential hazards.....
Uni-Model Prompt	As your road safety monitoring assistant, the results of the traffic footage is {original_label}, can you determine if the cyclist is safely navigating a roadway designated for vehicles, including bicycles, and identify any prohibited actions or potential dangers present?
Uni-Model Response (Language Model Response)	As your road safety monitoring assistant, based on the analysis of the traffic footage involving biking, I can determine potential risks and identify any prohibited actions or dangers. This includes assessing whether the cyclist is safely navigating the roadway, adhering to designated bike lanes, following traffic signals, and avoiding any hazardous manoeuvres that could lead to accidents or conflicts with vehicles and pedestrians.....
Uni-Model Response (Language Model Response)	As your road safety monitoring assistant, based on the analysis of the traffic footage involving handstand walking, I can identify potential risks and determine if any prohibited actions or dangers are present. This includes evaluating the safety of the activity in the given environment, checking for compliance with local regulations, and identifying any hazardous conditions, such as the proximity to vehicles or obstacles, that could lead to accidents or injuries.....
Multi-Model Prompt	As your road safety monitoring assistant, can you evaluate the frames extracted from the traffic footage to determine if the cyclist is safely navigating a roadway designated for vehicles, including bicycles, and identify any prohibited actions or potential dangers present?
Multi-Model Response (Language Model Response)	Based on the provided traffic footage frames, it appears that the cyclist is navigating a roadway that accommodates both vehicles and bicycles. The cyclist is seen riding within the designated bicycle lanes, which is a safe practice. However, there are some potential dangers and prohibited actions observed. In one frame, the cyclist appears to be riding very close to the curb, which could pose a risk if they need to manoeuvre suddenly. Additionally, in another frame, the cyclist seems to be
Multi-Model Response (Language Model Response)	Upon reviewing the frames from the provided footage, it appears the cyclist is navigating the roadway designated for vehicles, including bicycles. The cyclist seems to be following the designated lanes and is positioned correctly within the roadway. However, there are a few potential dangers and prohibited actions observed. In some frames, the cyclist is seen riding very close to other vehicles, which increases the risk of accidents or collisions, especially if any sudden manoeuvres are made by other road users.....

Figure. 2. The differences between the unimodal and multimodal prompts are marked in red. For the unimodal section, the labels as results of ResNet-18’s classification are part of the prompt whereas only the question and the frames are sent as the prompt in the multimodal section.

The 50 collected samples constitute two original labels: “BabyCrawling” and “Biking”. These labels have particular implications in smart home and traffic management and are significant in our daily lives. Using cosine similarity effectively quantifies the similarity between the original and perturbed results. To verify the similarity results from 4.1, this section performs a human evaluation of the responses.

From Figure 2 we can see a sample of each “BabyCrawling” and “Biking”. Inside the prompts, the language model has been given the role of a smart home security assistant to monitor the baby’s actions when it is reviewing samples of “BabyCrawling” and the role of a road safety monitoring assistant to ensure no prohibited actions and dangers are present when it is reviewing samples of “Biking”. The prompts of the unimodal response and the multimodal response are similar despite minor modifications to fit the inputs.

Comparing the original and perturbed unimodal response for both examples, the resulting textual response from the language model suggests that the attack was very effective and caused severe hazards. For “BabyCrawling”, the original response suggests that “it is important to ensure that the environment is safe to prevent any potential hazards”, whereas, in the attacked response, the attacked label is “CuttinginKitchen”, which the model regards as “potentially hazardous”, and requires “Immediate attention and intervention”. This suggests that in unimodal responses, the adversarial attack poses potential hazards and threats to the smart home system and misleads the system into taking wrong actions. The “Biking” example further supports this, as the original label of “Biking” is converted to “Handstandwalking”, and the language model’s response changed from “determine potential risks within the action of Biking” to a “prohibited and hazardous activity”. Therefore, the semantic similarities fail to accurately represent the language model’s responses and the hazards underlying the adversarial attack.

In multimodal attacks, the adversarial attack could be more efficient. The classification results match the original label for both the original and perturbed responses, indicating that the adversarial attack did not impact human-action recognition. As a result, the model delves into the specific details of the action: In “BabyCrawling”, the original response describes that “the baby is crawling and moving around” whereas, in the perturbed response, the language model suggests that the baby is also “interacting with objects”, which creates a relatively low level of risk compared to the unimodal attack; In “Biking”, the original response describes that the “cyclist is riding very close to the curb” and “approaching intersections without signaling” whereas in perturbed frames, the cyclist was very

close to other vehicles and overtaking a vehicle from the right, and the “Biking” sample describes more significant risks underlying in the system that may cause more severe results.

In conclusion, the observations from the human evaluations suggest that the adversarial attack can cause severe damage to unimodal models under certain contexts while causing less damage to multimodal models.

4. Discussion

This study investigates the effects of a visual adversarial attack towards the textual responses from two unimodal models connected and one single multimodal model. The study used cosine similarities and human evaluation methods to evaluate the difference in the generated text between the original and attacked textual responses. The conclusion that the adversarial attack was more effective towards the unimodal model overall has been drawn.

However, there are still areas of improvement that can further enhance the experiment’s effectiveness. In quantifying the effects of the attack, the Syntactic and Semantic Cosine similarities calculated do not represent the divergence between the sentences well. They need to be more accurate in evaluating the results without section 4.2, Scenarios and Human Evaluation. Hence, a more accurate similarity calculation can be implemented for more accurate representations of the results, which is needed to further evaluate a more significant number of samples and validate the conclusions.

Other than that, a more effective attack method can be selected that would cause more severe misclassification and misleading results towards the multimodal model GPT-4o, which is relatively robust. FGSM was selected for its simplicity and effectiveness in neural networks. However, the transferability of successful adversarial attack samples that is successful on ResNet-18 causes very little harm towards GPT-4o. Therefore, a different attack approach can be selected to have substantial impacts, and both ResNet-18 and GPT-4o can be utilized and experimented upon.

From the angle of developers, it is expected to analyze and process several modalities. No doubt, connecting several unimodal models and processing and passing on information from the previous to the next requires less cost than a multimodal model. However, the robustness of the models is also worth considering, as harmful results may be caused by malicious attacks, as discussed in section 4.2.

5. Conclusion

This paper discusses the effect of implementing FGSM on human action recognition models, compares the effects

of the attack between unimodal and multimodal models, and evaluates the results through cosine similarities and human evaluation. In the findings, the multimodal model exhibited more robustness than the unimodal model, as severe misclassification occurred in the unimodal model. At the same time, only minor details were misinterpreted by the multimodal model. This also suggests limited transferability from ResNet-18 to GPT-4o using FGSM. Despite that, there are still underlying threats and hazards posed by Human Action Recognition models that could cause significant risk, especially in smart home systems and traffic management, as pointed out in the human evaluation section. Therefore, this paper proposes that future studies could further evaluate the robustness of unimodal and multimodal models through more sophisticated and efficient attacks and develop potential adversarial defence mechanisms. In AI research, the security and safety of the models remain crucial, requiring defence mechanisms to evolve alongside newly emerging attacking methods. Therefore, this paper proposes that future studies could further evaluate the robustness of unimodal and multimodal models through more sophisticated and efficient attacks and develop potential adversarial defence mechanisms.

References

- [1] Haldar, S. (2020, April 9). Gradient-based Adversarial Attacks : An Introduction. Medium. <https://medium.com/swlh/gradient-based-adversarial-attacks-an-introduction-526238660dc9>
- [2] Liu, J., Wang, X., & Li, Y. (2021). Understanding the limitations of traditional action recognition methods. *Journal of Visual Communication and Image Representation*, 79, 103107.
- [3] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- [4] Moosavian, M. A., & Sadeghi, M. (2020). A comprehensive review of the recent advances in HAR. *Computer Vision and Image Understanding*, 195, 102924.
- [5] Zhang, H., Wu, Y., & Zhang, D. (2021). Adversarial training for deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 21-26.
- [6] Li, Y., Cheng, S., Wang, Q., & Li, Y. (2022). Keyframe extraction methods for video summarization: a survey. *ACM Computing Surveys*, 54(5), 1-36.
- [7] Gao, S., Jia, X., Ren, X., Tsang, I., & Guo, Q. (2024). Boosting Transferability in Vision-Language Attacks via Diversification along the Intersection Region of Adversarial Trajectory. In *ArXiv*. Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, Qing Guo. <https://arxiv.org/pdf/2403.12445>
- [8] UCF101 - Action Recognition Data Set. (2013, October 17). UCF Centre for Research in Computer Vision. <https://www.crcv.ucf.edu/data/UCF101.php>
- [9] resnet18 — Torchvision main documentation. (n.d.). Retrieved September 3, 2024, from <https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html>
- [10] Loshchilov, I., & Hutter, F. (2019). DECOUPLED WEIGHT DECAY REGULARIZATION. In *arXiv*. [arXiv. https://arxiv.org/pdf/1711.05101](https://arxiv.org/pdf/1711.05101)
- [11] [Pykes, K. (2024, August). Cross-Entropy Loss Function in Machine Learning: Enhancing Model Accuracy. DataCamp. <https://www.datacamp.com/tutorial/the-cross-entropy-loss-function-in-machine-learning>
- [12] HUMAN ACTIVITY RECOGNITION MODELS USING DEEP RESIDUAL NETWORKS. (n.d.). https://idrlib.iitbhu.ac.in/xmlui/bitstream/handle/123456789/1480/Chapter_5.pdf?isAllowed=y&sequence=14
- [13] Adversarial Example Generation — PyTorch Tutorials 2.4.0+cu121 documentation. (n.d.). Retrieved September 3, 2024, from https://pytorch.org/tutorials/beginner/fgsm_tutorial.html
- [14] ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. (n.d.). In ScienceDirect. Partha Pratim Ray. <https://www.sciencedirect.com/science/article/pii/S266734522300024X>
- [15] OpenAI. (2023). GPT-4 Technical Report. OpenAI. <https://cdn.openai.com/papers/gpt-4.pdf>
- [16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). Attention is All You Need. In *arXiv*. [arXiv. https://arxiv.org/pdf/1706.03762](https://arxiv.org/pdf/1706.03762)