Relieve Adversarial Attacks Based on Multimodal Training

Hongjie Lai

Sydney Smart Technology College, Northeastern University at Qinhuangdao, 066000, Qinhuangdao, China

Abstract. This paper explores the role of multimodal training in mitigating the problems caused by adversarial attacks, building on the foundations of deep learning. Deep learning models have reached great success in many areas such as image recognition and natural language processing. But their robustness has always been a concern. However, the emergence of adversarial attacks has exposed shortages of neural networks, forcing people to confront their limitations and further increasing concerns about the security of deep learning models. Adversarial training is an effective defense mechanism that incorporates adversarial samples into the training data, enabling models to better detect and resist attacks. This paper first introduces the principles and types of adversarial attacks, as well as basic concepts and related methods, including Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), DeepFool, and Jacobian Saliency Map Attack (JSMA). The paper then focuses on analyzing the robustness of the multimodal model CLIP based on contrastive learning. Finally, the paper proposes whether audio data can be added to the training samples of the CLIP model to further improve its robustness, and raises related issues and bottlenecks.

1 Introduction

Deep learning has taken the field of artificial intelligence to a new level. It learns and processes complex data by mimicking the function and structure of the human brain's neural network. Deep learning models typically have many layers, each learning different features of the data. However, adversarial attacks have revealed weaknesses in neural network models. It makes small changes to the training data. Although these changes can't be discovered by human, it still could cause the model to make wrong predictions. This leads to the development of adversarial training, which is used to train models and improve their robustness. Traditional adversarial training methods usually train on single-modal data, such as images or text. But in our daily life data is often multimodal, such as images text and voice. Multi-modal training combines data from different modalities during training. For instance, in image recognition tasks, multi-modal training could combine the image information and the text information so that the model's ability of detecting subtle changes in images could be improved. In addition, multimodal training could improve the model's generalization

Corresponding author: 202319117@stu.neuq.edu.cn

learning ability, allowing it to maintain performance when it is facing with unknown or disturbed samples.

CLIP, one typical model, is a good example in nowadays. It uses contrastive learning to map image and text information into the same vector space. This makes related information closer and unrelated information further apart in that vector space. This multimodal training method is robust because attackers must attack both image and text samples to affect the model's predictions.

What's more, CLIP uses a amount of multimodal data while training, which increases its ability to generalize. This allows it to perform well in zero-shot and few-shot learning scenarios. Traditional image recognition models require a lot of manually labeled data, but CLIP trains on image-text pairs without the need for manual labeling. This reduces the cost and time of training the model, and enables it to learn more semantic information.

It can also learn the deep relationships between images and text, allowing it to accurately recognize and classify previously unknown images based on the text descriptions. This makes the model more flexible and practical for the applications in the real world. Given that it has excellent performance in zero-shot and few-shot tasks, it is used in various fields at the moment. For example, CLIP is applied to point cloud understanding, extending its 2D knowledge to the 3D point cloud domain [1], which could be used for autonomous driving, medical image analysis and industrial automation. CLIP-Vil is used in visual language tasks [2], showing significant improvements in visual question answering, image captioning and visual language navigation. And the result shows it's better than previous models.

This paper will introduce adversarial attacks and various adversarial training methods. It will also analyse the CLIP model and show how it can mitigate prediction errors caused by adversarial attacks. Finally, it will discuss the future applications and development of multimodal language learning, pointing out existing problems.

2 Principles and cases of adversarial attacks

2.1 Principles

A type of attack on machine learning models is called adversarial attacks. The key ideas is causing the model to make incorrect predictions by making small alters to the entered data. These changes are often difficult for humans to detect, sometimes being tiny variations at the pixel level. However, they will cause the model to make incorrect classifications. For example, in image classification tasks, an image has been added a noise and it is difficult for a human to detect. This can cause the model, which would normally correctly identify a cat, to incorrectly classify it as a dog.

Due to machine learning models are sensitive to the input data and their reliance on specific data distributions during training, adversarial attacks are found. Most machine learning models are trained in a data-driven manner, learning patterns and rules from the data. However, these models are often very sensitive to data outside their training distribution. Even small perturbations can lead to incorrect predictions.

Machine learning models could be fooled by attackers with carefully crafted input data, causing them to make incorrect predictions. White box attacks, black box attacks, and grey box attacks are three types of adversarial attacks. And they are sorted by the attackers' knowledge of the target model.

White box attacks appear when the attackers have full information about the target model, including its structure, parameters, training data, activation functions, and loss functions. The attacker has free access to internal information like weights, biases, and intermediate results of the model. White-box attacks tend to be more effective because attackers can generate

adversarial examples more accurately. FGSM, PGD, DeepFool, JSMA and C&W are common black box attack methods.

Black box attacks refer to scenarios where attackers have access only to the model's inputs and outputs, without any internal information. They need to observe the model's predictions in order to carry out their attacks. Black-box attacks tend to be more complex because attackers need sophisticated methods to create adversarial examples. Usually, gradient estimation-based attacks, transfer learning-based attacks, and evolutionary algorithm-based attacks are included by black box attack methods.

Grey box attacks fall between white box and black box attacks. Attackers could make use of some information about the model, such as its structure or partial parameters, but not the whole internal details. Grey box attacks typically use this partial information to generate adversarial examples more effectively. Common grey box attack methods include gradient estimation based attacks, which differ from black box attacks by using some model information for gradient estimation.

The types of attacks reflect the attackers' understanding of the target model. Different types of attacks require different methods. Studying adversarial attack types and methods improves our understanding of their principles and helps to develop more effective defense measures.

2.2 Cases of adversarial attacks

Given that the highly non-linear nature of deep learning models, but they can be considered linear models within a local domain. Adversaries exploit this local linearity by making minimal adjustments to the parameters in the samples, causing changes in the model's predictions. Using the model's gradient information, they identify the direction that has the greatest impact on the model and perturb the samples along that direction to generate adversarial examples.

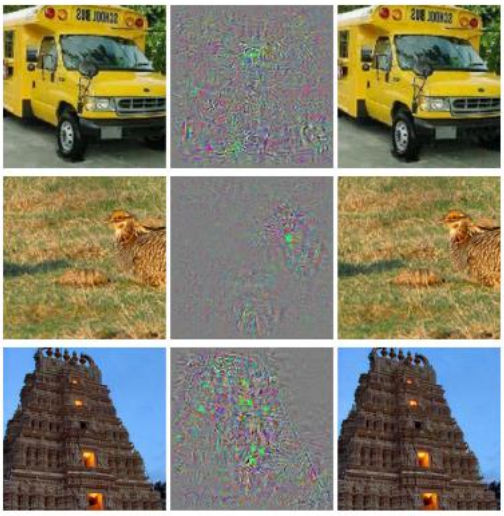

Fig. 1. Adversarial attack cases of AlexNet [3]

Fig. 1 shows adversarial examples generated by AlexNet. The leftmost image is the correctly predicted example, the middle column shows the maximum difference between the

correct and incorrectly predicted images, magnified by a factor of 10, and the rightmost image is the adversarial example.

The example is generated by adding a small amount of noise to the image. This noise is imperceptible to humans, but is enough to fool AlexNet into misidentifying the image. When presented with the adversarial sample, AlexNet predicts "ostrich, Struthio camelus". Clearly, these are neither camels nor ostriches in the correct prediction sample, indicating that the small perturbations led the model to make an incorrect prediction [4].

3 The overview of adversarial training and related methods

One of the simplest adversarial attack methods is FGSM. It is based on the gradient information of model output relative to the input. By adding a small perturbation along the gradient direction, adversarial examples are generated. The following function represents the implementation equation of FGSM:

$$x\_adv = x + \varepsilon * sign[\nabla\_x L(x, y)] \quad (1)$$

In this context, x represents the original sample, such as an image or a piece of text. x_adv is the generated adversarial example, created by adding perturbations to the original sample. ϵ is the perturbation size, a scalar that charges the intensity of the perturbation; larger ϵ results in stronger perturbations, making it easier for the model to misclassify the adversarial example. The term sign(∇_x L(x, y)) denotes the gradient's sign of the model output with respect to the input, indicating the direction of the perturbation. It is a vector where each element represents the sign of the gradient for the feature of corresponding data, either positive or negative [5].

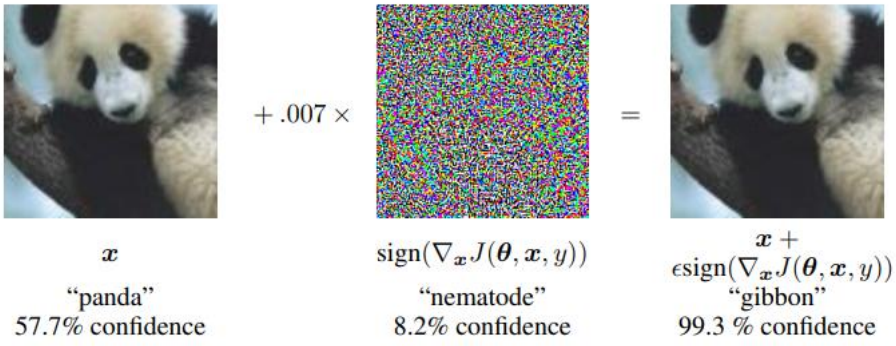

Fig. 2. Adversarial attack's case of FGSM [5]

Fig. 2 is an example of FGSM attacks. The model has 57.7% confidence that the input x is "panda". After applying a small perturbation, the model predicts "gibbon" with 99.3% confidence, which clearly does not match the information in the image.

PGD is an iterative attack method that calculates the gradient at each iteration and updates the example along the direction of the gradient. It projects the example back into the neighborhood of the original sample to ensure that the perturbation size does not exceed a preset value, thus keeping it sufficiently small. Compared to FGSM, PGD has a stronger attack effect and can generate more effective examples. The intensity of the attack can be charged by adjusting the number of iterations and the perturbation size. However, due to the multiple iterations, it requires significant computational resources, which reduces its attack efficiency. The following function represents the implementation equation of PGD:

$$x\_adv = \text{clip}(x + \alpha * \text{sign}(\nabla\_\text{x}\, \text{L}(x\_adv, \text{y})), \text{x} - \varepsilon, \text{x} + \varepsilon) \quad (2)$$

x, x_adv and ϵ have the same definitions as in FGSM，

$$\text{clip}(x + \alpha * \text{sign}(\nabla\_\text{x}\, \text{L}(x\_adv, \text{y})), \text{x} - \varepsilon, \text{x} + \varepsilon) \quad (3)$$

represents the projection operation [6]， Project the

$$x + \alpha * \text{sign}(\nabla\_\text{x}\, \text{L}(\text{x\_adv}, \text{y})) \quad (4)$$

into the neighbourhood of the foregone sample x to ensure that the perturbation is not too large. PGD is also applied to large language models. In the paper "Attacking Large Language Models with Projected Gradient Descent" [7], it is noted that by continuously relaxing the input prompts, discrete token sequences are transformed into continuous probability distributions, allowing for optimization using gradient descent. After each gradient update, they use a simplex projection to keep the representations within the valid probability distribution range, and employ entropy projection to control errors introduced by the relaxation. This method not only achieves comparable attack success rates to traditional discrete optimization methods but also significantly improves efficiency.

DeepFool is a linearisation-based attack method suitable for both binary and multi-class classifiers [8]. It works by finding a point on the model's decision boundary and then calculating the distance from that point to the original sample. It creates adversarial examples through adding a small disturbance in the direction of the line connecting the point and the original sample. High efficiency, strong interpretability and good generality are the advantages of this method. However, it also has some limitations, including targeting, ease of model defense, and high algorithmic complexity.

DeepFool could be applied to deep neural network fingerprinting [9]. A set of data points close to the the goal model’s classification boundary are extracted and their predictions about the goal model are recorded. These data points and their predictions form the fingerprint of the target model. The fingerprint data points are input into the model being verified and its predicted results are compared with the target model's fingerprint labels to certify whether a model is a copy or modified version of the target model. A high matching rate means that the model being validated is likely to be a copy or modified version of the goal model.

JSMA is an attack method based on gradient increase, which uses saliency maps to identify the features in the input image that have the largest impact on classification results. JSMA generates adversarial examples by iteratively changing these features and then makes the model erroneously classify the input image.

DE-JSMA [10] is a variant of JSMA specifically designed for Synthetic Aperture Radar Automatic Target Recognition (SAR-ATR) systems. It is a novel sparse adversarial attack algorithm that addresses the challenges posed by the sparsity of SAR images.

4 Multimodal models

Multimodal language models are designed to understand and process different types of data, including text and images. They integrate information from different modalities to achieve a richer understanding and perform more complex tasks. There are many methods of mutimodal learning, such as Visual Semantic Embedding (VSE), Self-Supervised Contrastive Learning (SCAN), Multi-Modal BERT (MM-BERT) and so on. VSE++ [11] is a simple and effective method for improving the performance of visual-semantic embedding models. It builds upon VSE by introducing the concept of hard negative samples and using a triplet loss function to train the model. SCAN is a novel approach to image classification without relying on labeled data. The core idea is to leverage the semantic similarity between

images to learn a discriminative representation [12]. However, a typical example of such a model is CLIP. Fig. 3 shows the architecture of the CLIP model:

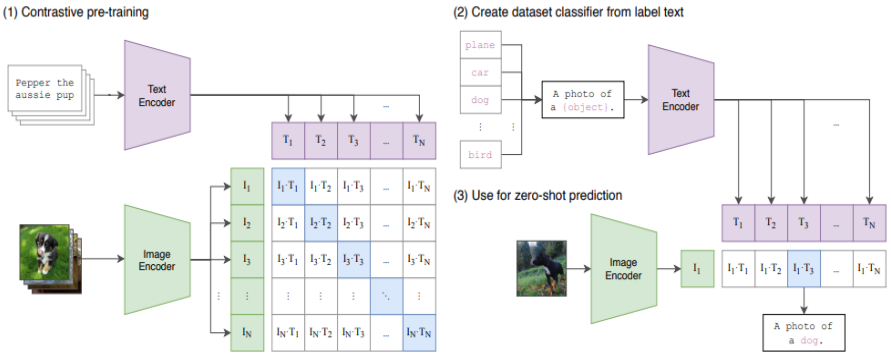

Fig. 3. CLIP Model Architecture [13]

Fig. 3 shows that CLIP learns image encoders and text encoders through contrastive pre-training on a data set of image-text pairs. By maximizing the similarity between relevant image and text vectors and minimizing the similarity between irrelevant vectors, it vectorizes the image and text information into a shared vector space. This allows the model to predict which image embedding matches which text embedding.

The model can then create a data set classifier from the labeled text, producing text embeddings for each label that represent the meanings associated with those labels.

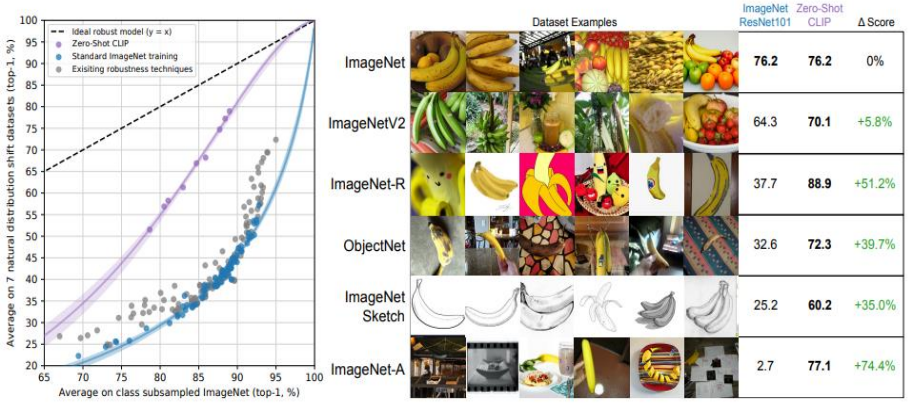

Fig. 4. Performance of different image classification models on various data sets [13]

The average top-1 accuracy of different models on the ImageNet dataset (x-axis) and their average accuracy on seven naturally distributed variation datasets (y-axis) is shown in the left scatter plot in Fig. 4. The black dashed line represents the ideal case where the accuracy of the model is the same for both sets of data. The blue points represent the performance of standard ImageNet training, the grey points represent existing robustness techniques, and the purple points represent CLIP, a zero-shot image classification model. This shows that CLIP achieves accuracy levels very close to the ideal case on both datasets, demonstrating strong generalization capabilities. It adapts well to different data distributions, maintaining high accuracy even for distributions not seen during training.

On the right-hand hand of Fig. 4, a table contrast the performance of ResNet101 and CLIP on five different datasets, each representing a different variation of the data distribution. The results show that while ResNet101 performs well on ImageNet, its performance drops

significantly on other datasets, indicating a high sensitivity to changes in data distribution. In contrast, CLIP outperforms ResNet101 on all datasets, particularly on datasets with significant distribution variation, with accuracy improvements of over 10%. This highlights CLIP's ability to better understand the semantic information of images and its superior performance in zero-shot and few-shot learning scenarios compared to other models.

The usages of contrastive and multimodal learning, enables CLIP to understand the information from multiple sides, which giving the model stronger generalization capabilities and higher robustness. This effectively reduces the influence of single adversarial attacks.

5 Conclusion

This paper is based on the deep learning and probes the role of multimodal training in mitigating adversarial attacks. In image recognition and natural language processing, deep learning models have achieved great success, but their vulnerability to the attacks remains a persistent concern. The emergence of adversarial attacks exposes the weaknesses of neural network models and increases concerns about their security.

Adversarial training has occurred as an effective defense. It improves the robustness of deep learning models by mixed adversarial samples into the training data, thus the model's ability to recognize and resist attacks could be enhanced.

This paper firstly introduces the principles and three kinds of attacks methods, including white box attacks, black box attacks and grey box attacks. It also describes common attack methods such as FGSM, PGD, DeepFool and JSMA. These methods introduce small changes to the entered data that may not be found by humans, but could drive the model to fail to predict or to predict wrong.

And then this paper focuses on analyzing the robustness of the multimodal model CLIP, which is based on contrastive learning. It maps image and text information into the same vector space, then uses contrastive learning to ensure that related information is placed closer, while unrelated information is placed further apart. This multimodal training approach could improve the robustness of the model, because attackers must target both image and text samples at the same time to affect the model's predictions. In addition, by training CLIP with a high amount of multimodal data, its generalization ability is increased, enabling good performance even in zero-shot and low-shot learning scenarios.

Lastly, this paper proposes to include audio data in the training samples of the CLIP model to further improve its robustness. This idea is based on the complementarity of multimodal information. By combining audio with image and text data, the model can become more sensitive to noise and disturbances, improving both robustness and generalization. However, the challenge is whether the information carried by the vectors after processing by the image and text encoders is at the same level. There is currently no mature method for comparing the levels of the two different types of vectorization, which is a significant challenge for the future. The addition of a third dimension of information undoubtedly increases the difficulty of meeting this challenge.

References

1. R. R. Zhang, et al. Pointclip: Point cloud understanding by clip. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2022).
2. S. Shen, et al. How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383 (2021).
3. C. Szegedy, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013).

4. K. Alex, S. Ilya, and H. Geoff. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pages 1106 – 1114, (2012).
5. I. J. Goodfellow, S. Jonathon, and S. Christian. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).
6. A. Madry, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017).
7. S. Geisler, et al. Attacking large language models with projected gradient descent. arXiv preprint arXiv:2402.09154 (2024).
8. D. Moosavi, M. Seyed, F. Alhussein, and F. Pascal. Deepfool: a simple and accurate method to fool deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition. (2016).
9. S. Wang, and C. H. Chang. Fingerprinting deep neural networks-a deepfool approach. 2021 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, (2021).
10. J. I. N. Xiaying, L. I. Yang, and P. A. N. Quan. DE-JSMA: a sparse adversarial attack algorithm for SAR-ATR models. Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University 41(6): 1170-1178 (2023).
11. F. Faghri, D. J. Fleet, J. R. Kiros, et al. Vse++: Improving visual-semantic embeddings with hard negatives[J]. arXiv preprint arXiv:1707.05612, (2017).
12. W. Van Gansbeke, et al. Scan: Learning to classify images without labels. European conference on computer vision. Cham: Springer International Publishing, (2020).
13. A. Radford, et al. Learning transferable visual models from natural language supervision. International conference on machine learning. PMLR, (2021).