Adversarial Learning

Introduction:

Adversarial learning is a defence mechanism against the exploitation of weakness in the machine learning model. Due to the complex nature of machine learning models, there are chances for leaving behind the weakness in it. This could provide an opportunity for the attackers to trick the model to incorrect predictions or giving away sensitive information (TDS, 2021). Different types of attacks include but not limited to 'poisoning' which is used to attack the training data, 'evasion' which focuses on the data the model uses and 'model stealing' which focuses on stealing the structure or training data of the model (TDS, 2021). Defending against such adversarial attacks require a model to be robust which can be achieved through adversarial learning.

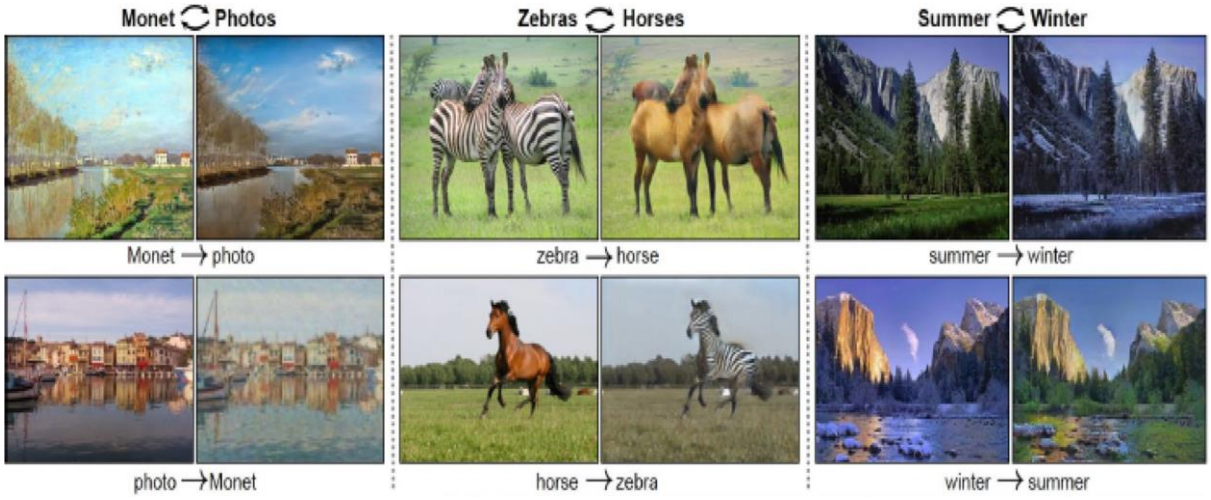
Applications:

The concept of adversarial learning can be applied across various domains such as healthcare, e-commerce, computer vision, etc. Generative Adversarial Networks (GANs) is one of the types of adversarial learning has numerous real-world applications in place with respect to texts, images, audios and videos. The concept of generative adversarial network (GAN) involves two neural networks 'discriminator' and 'generator' competing against each other. The later portion of the report provides detailed description of the GAN.

Image based Applications

Image to Image Translation:

Conditional adversarial network is a suitable approach to translate an input image to an output image in computer vision domain (Alqahtani and Kavakli-Thorne, 2019). The pix2pix model produced effective results for various problems in computer vision such as colouring of black and white images, generating maps from aerial photos, etc.



Text to Image Translation:

The architecture of generative adversarial network has the ability to produce images from text descriptions. For example, given a text caption of a bird with its features such as ‘white with some black on its head and wings and a long orange beak’, the GAN model will be able to create numerous images that matches with the given text description (Alqahtani and Kavakli-Thorne, 2019).

Anime Character Generation:

The expensive task of game development and animation production can be automated by GAN model. The functions of GAN can be used for transform photos of real-world scenes into cartoon style images (Alqahtani and Kavakli-Thorne, 2019).

Person Re-Identification:

The applications of GAN are not just limited to texts and images, it can be used across various methods which process audio and video inputs. For example, temporal GAN model has the ability to capture different aspects of the video. It can generate videos with facial expressions such as blinks and even sync of lip movements with the audio (Alqahtani and Kavakli-Thorne, 2019).

In addition to the applications mentioned above, GANs can be applied for other multiple image-based applications such as Image blending, Face-aging, high quality image generations to name a few.

Audio-Video Based Application:

GANs can be applied to generate realistic videos directly from raw audio. Temporal GAN method can be used to generate videos which have lip movements that are in sync with audio and natural facial expressions such as eyebrow movements and blinks (Alqahtani and Kavakli-Thorne, 2019).

Challenges:

The primary challenge of adversarial learning lies in building the robust machine learning algorithms. Though great number of new algorithms have been published every year with respect to attacks and defences, there is still no finalized and optimized algorithm based on adversarial retraining (Silva and Najafirad 2020). The presence of certified defence mechanism is an alternative to generate robust models however it requires high computational cost. Even the certified defences can be broken by applying big disturbance to the model (Silva and Najafirad 2020).

GAN Training Challenges:

Mode Collapse: Mode collapse is a major issue in training the GAN models. It refers to the limitation of generating diversified samples even when trained on multi-model data (Saxena and Cao 2021). This could be caused if the discriminator gets caught up in the minimum trap and always rejecting all

instance of its input, then generator keeps creating same type of instances (Pavan Kumar and Jayagopal, 2020).

Instability: The issue of mode collapse eventually brings instability to the training of GAN models. The gradient vanishing problem resulting from mode collapse leads to generator's loss as discriminator can easily differentiate between real and fake samples (Saxena and Cao 2021). This leads to the instability between generator and discriminator.

Non-Convergence: Non-Convergence in training the GANs is caused due to the failure of generator and discriminator to reach an unbalanced state (Pavan Kumar and Jayagopal, 2020). This situation could be caused by irregularities in the model structure and training strategies.

GAN Performance Challenges:

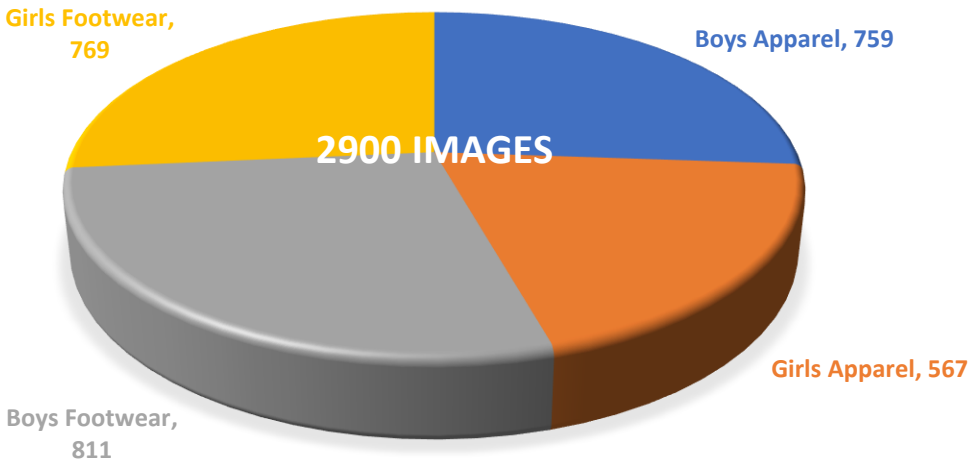
With respect to performance, GANs ability is questioned over creating fake videos called deep fakes. It is difficult to create a deepfake that shows blinking eyes as people in general do not take a picture with eyes closed (Pavan Kumar and Jayagopal, 2020). The requirement of similar skin tone and orientation of faces adds to the challenge of creating optimal deep fakes (Pavan Kumar and Jayagopal, 2020).

Problem Description:

Visual recommender systems (VRS) are shown to be effective in using the product images to make the recommendation. They are of use in different domains such as fashion, food and point of interest. However, they are vulnerable to different adversarial attacks which can harm the integrity of recommender systems. Both poisoning and evasion attacks can be carried out against these models. For example, adversarial attacks in computer vision for poisoning the dataset with adversarial product images can alter the training of the model and can push the target items towards higher recommendation (ResearchGate, 2021). As part of evasion attack, it is even easier for an attacker to alter the recommendation by uploading an adversarial version of product images on platforms such as Amazon, eBay, Pinterest, etc (ResearchGate, 2021). Hence it is crucial to make these models robust against the adversarial attacks through adversarial learning. The approach chosen to solve the problem is generative adversarial network which has the ability to create adversarial images.

Dataset:

The dataset of e-commerce fashion products has been chosen to solve the problem. It consists of images of fashion products for both boys and girls. The dataset has four categories of images in total namely Boys' Apparel, Girls' Apparel, Boys Footwear and Girls Footwear. The below chart represents the visualization of categorial data and the numeric figures with respect to each category.



The dataset also contains a CSV file along with the images. The CSV file has ten columns which includes various details of images. The ten columns hold the data such as product id which is unique to each image, gender, category, subcategory, product type, colour, usage, product title, image and the url of the image. The approach used to solve the problem however processed only the images from the dataset but not the data from the csv file. The data has been taken from the Kaggle platform as mentioned in the below url.

https://www.kaggle.com/datasets/vikashrajluhaniwal/fashion-images

Different Approaches:

Generative Adversarial Network:

GAN works on the aspect of min-max two-person zero-sum game. The players in the game are two neural networks called discriminator and generator. Input to the generator is a random noise while discriminator has two inputs 'real' and 'fake' sample. The purpose of discriminator is to determine whether a given sample is a real or fake while generator tries to fool the discriminator by generating fake samples (Alqahtani and Kavakli-Thorne, 2019). The output of discriminator will be between the values 0 and 1 with values close to zero indicates a fake sample.

Variation Autoencoders (VAE):

VAE is also a generative model which uses less dimensional latent variables to generate images (Everitt's blog, 2018). Similar to GAN, it works on two networks called encoder and decoder. The samples from the encoder gets converted to latent variables which is then used by the decoder to produce the final output.

GAN	VAE
The training procedure of GAN is relatively complex (Mirza and Iftekharul, 2019).	Compared to GAN, training the VAE model is easier
The output generated from the GAN used to have higher quality than the outputs from VAE (Mirza and Iftekharul, 2019).	Though the outputs can be closer to the original samples, it used to have blurs in the images (Mirza and Iftekharul, 2019).
The generated samples can only be judged real or fake however it can produce more realistic images (Everitt's blog, 2018).	It can do the direct comparison between generated and original samples (Everitt's blog, 2018).

Since the GAN model have some important upper hand than the VAE model especially with respect to the quality of the output, the approach of generative adversarial network has been taken to solve the adversarial problems in visual recommendation system.

Image Pre-processing:

Manual Steps:

Two folders one each to train and test the model have been created and then enough number of images from each category are placed into the folders. This has been carried out in such a way that the images in the train and test folders are different for each category to ensure appropriate evaluation of model performance.

Greyscale conversion:

This process converts the coloured images to black and white form. Since most images will not require recognition of colours to get processed, converting to greyscale will not impact the results. It reduces the computation complexity by lessening the number of pixels in an image (Section 2021).

Image Normalization:

In order to apply same algorithms over different data formats, normalization technique can be applied. This process converts the pixel values of an image to a predefined range. The normalization range applied to the dataset was (-1, 1). It helps to achieve a fairness across the images by allowing equal contribution from all images through re-scaling the pixels to the equal range (Section 2021). Since high pixel images could require low training rate and low pixel images could require high learning rate, normalization of pixels brings uniform learning rate across the images (Section 2021).

Feature extraction:

Since the approach used was conventional or unconditional GAN, no data other than images were used. The attributes in the csv file such as product id, gender,

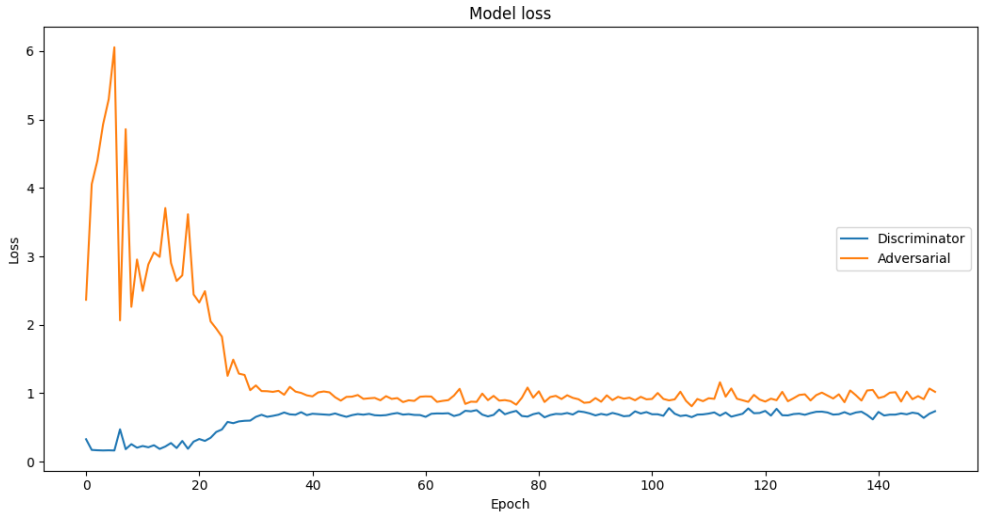
url of the image, etc. are ignored. Only the image and its category are used to train and test the model.

Reflective Conclusion:

Even though the concept is about machine learning, I believe that I have learned more than the machine. Learning starts with understanding the basics in adversarial learning and its impacts. The mindset at the beginning was whether adversarial attacks on the machine learning models will result in any serious impact? But it got changed after reading the consequences of adversarial attacks in a traffic signal board which is being used by auto driving cars. While the new machine learning models getting developed it is important that they are also getting trained to different adversarial attacks. The various types of adversarial attacks and the different methods to combat against are some of the key takeaways from this work.

Result:

The model below is for random samples of 100 each for training and testing the model.



The model is able to slightly increase the discriminator's loss as evidenced by its progress through the epochs. However, the model can be trained appropriately to adversarial attacks only with enough number of training and testing data. Though the number of images used was 2900 which satisfies the criteria of the coursework, a greater number of images would have been better in training and evaluating the model. This could have been accomplished by either opting for dataset with higher images or by using the augmentation technique to expand the existing dataset which I have not used. As part of the future work, one of the extensions of GAN model called conditional GAN can be used to improve the performance of the model. As the approach of conditional GAN will use additional data to generate the images, possibility of achieving the better outcome

is highly likely. Any of the data from the CSV file such as product ID, gender, apparel type, etc. may be used for the new approach.

References

Alqahtani, H. and Kavakli-Thorne, M. (2019). Applications of Generative Adversarial Networks (GANs): An Updated Review, *Archives of computational methods in engineering*, 28(2), p.525-552.

Everitt's blog (2018). A comparison between VAE and GAN. Available from:

https://everitt257.github.io/blog/2018/07/05/VAE_GAN.html [Accessed May 1 2022]

GitHub (2017). Kroosen. Available from:

https://github.com/kroosen/GAN-in-keras-on-mnist/blob/master/GAN-keras-mnist-MLP.ipynb [Accessed April 25 2022]

GitHub (2019). Garima13a. Available from:

https://github.com/Garima13a/MNIST_GAN/blob/master/MNIST_GAN_Solution.ipynb [Accessed April 26 2022]

GitHub (2018). Divyanshj16. Available from:

https://github.com/divyanshj16/GANs/blob/master/GANs-TensorFlow.ipynb# [Accessed April 25 2022]

Medium (2020). The Ethical Concerns of GANs. Available from:

https://medium.datadriveninvestor.com/the-ethical-concerns-of-gans-9ef6b88f79db [Accessed April 28 2022]

Mirza, S. and Iftekharul, M. (2019). A Comparative Study on Variational Autoencoders and Generative Adversarial Networks, *International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, p.1-5

Pavan Kumar, M, R. and Jayagopal, P. (2020). Generative adversarial networks: a survey on applications and challenges, *international journal of multimedia information retrieval*, 10(1), p.1-24

ResearchGate (2021). Adversarial Attacks against Visual Recommendation: an Investigation on the Influence of Item's Popularity. Available from:

https://www.researchgate.net/publication/354117200_Adversarial_Attacks_against_Visual_Recommendation_an_Investigation_on_the_Influence_of_Items%27_Popularity [Accessed April 26 2022]

Saxena, D. and Cao, J. (2021). Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions, *ACM computing surveys*, 54(3), p.1-42

Silva, S., and Najafirad, P. (2020). Opportunities and Challenges in Deep Learning Adversarial Robustness: A Survey.

Section (2021). Getting Started with Image Preprocessing in Python. Available from:

https://www.section.io/engineering-education/image-preprocessing-in-python/#:~:text=There%20are%20several%20techniques%20used%20to%20preprocess%20image,resizing%2C%20converting%20images%20to%20grayscale%2C%20and%20image%20augmentation [Accessed April 29 2022]

TDS (2021) What is Adversarial Machine Learning? Available from:

https://towardsdatascience.com/what-is-adversarial-machine-learning-dbe7110433d6 [Accessed April 13 2022]