A Deep Learning Visual Content Based Recommender System to Defend Adversarial Attacks

Research Project
MSc in Data Analytics

Komal -
Student ID: x20207034

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes
Dr. Musfira Jilani
Dr. Pramod Pathak

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Komal -
Student ID:	x20207034
Programme:	MSc in Data Analytics
Year:	2021-22
Module:	Research Project
Supervisor:	Dr. Paul Stynes, Dr. Musfira Jilani, Dr. Pramod Pathak
Submission Due Date:	15/08/2022
Project Title:	A Deep Learning Visual Content Based Recommender System to Defend Adversarial Attacks
Word Count:	3600
Page Count:	13

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Komal -
Date:	18th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Deep Learning Visual Content Based Recommender System to Defend Adversarial Attacks

Komal -
x20207034

Abstract

Recommender System is a type of machine learning technique which is used to produce significant recommendations to a group of users based on their preferences in the past. *Deep Neural Networks* have proven to be a wonderful fit while being used to deploy recommender systems, however the challenge with deep neural networks is that they are vulnerable to *Adversarial Attacks*, according to recent studies. The main goal of this research is to demonstrate adversarial attack defenses for visual content-based recommender system using deep learning. This study consisted of the examination of vulnerability of visual content-based recommender system against different targeted adversarial attacks using state-of-the-art white-box adversarial attack techniques and different *Adversarial Training* defense mechanisms to make our recommender system more robust against these executed attacks. A DeepFashion dataset used in this study which is a combination of 800,000 labelled images of clothes. For evaluation success rate metric was used in this research. Results of our experiments showed that from *Fast Gradient Sign Method, Projected Gradient Descent and Carlini & Wagner* methods, PGD with 128 iterations and CW attacks had the highest success rate. And traditional *Adversarial Training* defense method made system more robust compared to *Curriculum Adversarial Training* method. This proposed study helped in understanding the positive impact of defense mechanism on the adversarial attacked model and encouraged to train our recommender systems against these attacks in advance.

1 Introduction

The use of the internet has increased dramatically in recent years, so the volume of data. Therefore, it is essential to refine the information that can be useful to a person from the vast amount of data available. Recommender system (RS) is a machine learning system that make recommendations to the users depending on variety of parameters Aggarwal et al. (2016).

Visual content-based RSs are indeed incredibly successful in a variety of fields such as fashion, entertainment, food, etc. However recent studies have unveiled serious security concerns against adversarial attacks on these RSs Dalvi et al. (2004). Adversarial attack is a machine learning method to modify input images which are also called as adversarial examples Szegedy et al. (2013). These modified images are feed to machine learning model and force the model to predict incorrectly. Further, these adversarial attacks can be categorized based on adversarial aims, adversarial knowledge, and their scope Rathore et al. (2020). In this study, we are focusing on targeted item-to-item attacks using state-of-the-art white-box attacks. When the attacker has specific class and model to attack and attacker force model to predict attacker's target class instead of the correct class are targeted attack and when attacker is assumed to have access and all the vital information related to the model like its parameters, architecture and dataset then it is called white-box attack.

Many studies and research articles have been published on adversarial attacks against collaborative filtering (CF) based RSs, to show their vulnerability against adversarial attacks whereas there are only few studies on adversarial attacks against content-based RSs. The different ways in which an adversarial attack can affect the output of a content-based RS is shown in figure

1 Therefore, in this study we are presenting targeted adversarial attacks using state-of-the-art white-box attacks against visual content-based RS using DNNs. Also in the following phase, two defense strategies: Adversarial Training (AT) and Curriculum Adversarial Training (CAT) are used to demonstrate the impact of AT because adversarial training is a defense method against adversarial examples that works to improve a neural network's robustness by training it with adversarial examples which will significantly improve robustness of our trained model against actual attacks.

The aim of this research is to investigate vulnerability of a visual content-based CNN and kNN RS against targeted state-of-the-art white-box adversarial attacks. And to compare the impact of two different- AT and CAT defense methods on trained visual content-based RS.

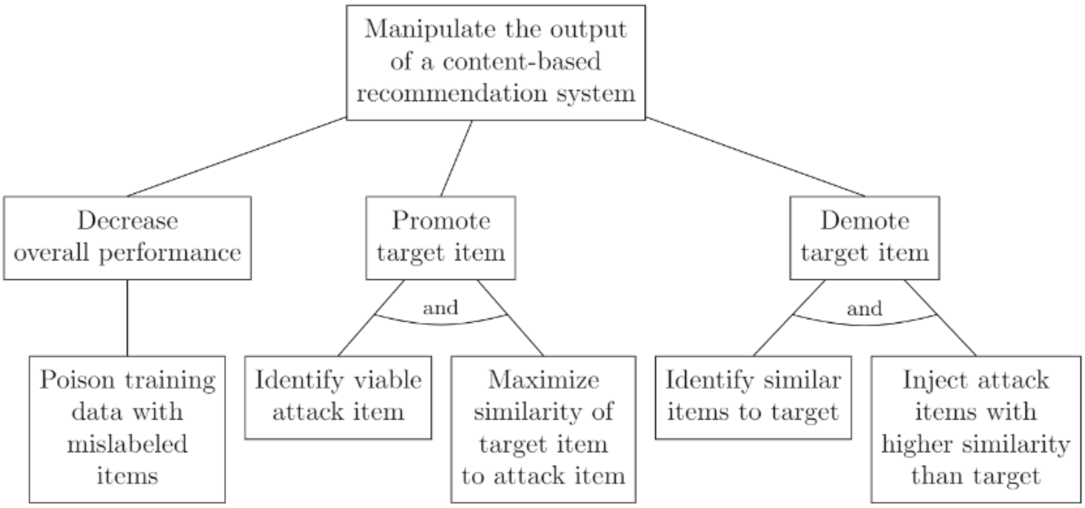

Manipulate the output of a content-based recommendation system
Decrease overall performance
Promote target item
Demote target item
and
and
Poison training data with mislabeled items
Identify viable attack item
Maximize similarity of target item to attack item
Identify similar items to target
Inject attack items with higher similarity than target

Figure 1: Different ways in which the output of a content based can be affected by an attack

2 Related Works

Recommender Systems are quite popular in various domains, and they were first presented by Bollacker et al. (1998). This model helped them to improve the recommendations of research articles. Melville and Sindhwani (2010) explored possibilities for RS in systems such as Amazon and Netflix and explained different categories of RSs such as Collaborative filtering (CF), Content-based filtering and Hybrid Aggarwal et al. (2016). Convolutional Neural Networks (CNNs) are very popular in image data processing which was first used by LeCun et al. (1989). They are very useful and widely implemented with RSs. However, RSs with DNNs are vulnerable to adversarial attacks which was first published by Dalvi et al. (2004). A useful RS for online shopping search engines was demonstrated with neural network by Chen et al. (2017). Sravani and Kurumeti (2021) build CF based movie RS and discovered that CNN based model was more accurate as compared to SVM based model. Sharma et al. (2021) examined the issues faced by RSs because of large amount of data available now and make RSs to give unideal recommendations. They build Semantic Personalized Recommendation System (SPRS) for semantic gap removal and this model proposed to recommend videos on the basis of user's previous actions.

Adversarial attacks are of different types and Ye (2021) developed Thundernna attack which was first order white box adversarial attack. He built a neural network model and performed Thundernna attack, FGSM and PGD attacks on the model. According to his results, Thundernna attack was more successful than FGSM but PGD had the highest success rate. Uchendu et al. (2021) used FGSM, PGD and EOT white-box attacks to investigate Bayesian Neural

Networks (BNNs) vulnerability. As a result of adding uncertainty measure and adversarial defense methods, they found their model was more robust in comparison of traditional neural network models. Dong et al. (2019) showed that face recognition models produced by using CNN are vulnerable to black-box adversarial attacks. Cohen et al. (2021) presented a visual RS to show the influence score and item's rank with the help of black-box testing. They showed that RSs are vulnerable to use the images which were provided by external source. Hirano et al. (2021) examined the vulnerability of DNNs against universal adversarial perturbation (UAP) and showed the positive impact of adversarial retraining. Żelasko et al. (2021) studied the vulnerability of speech recognition systems against targeted adversarial attacks. They used random smoothing and WaveGen to improve the robustness of the models. Rakin et al. (2021) published first adversarial weight attack of targeted scenario on DNNs. For untargeted attacks, Sadrizadeh et al. (2022) executed an untargeted white-box block-sparse adversarial attack on DNNs of text classifier. Wei et al. (2020) showed that the video recognition systems are vulnerable to heuristic untargeted black-box attacks and also their system reduced 28% queries to achieve human-invisible perturbations.

Pal et al. (2021) proposed hybrid adversarial training for the improvement of robustness of deep speaker recognition system. Their model achieved 3.29% adversarial accuracy for PGD attacks and 3.18% for CW attacks. Zhang et al. (2022) presented a framework named as TIKI-TAKA for deep learning-based Network Intrusion Detection Systems (NIDS) against adversarial attacks. They used model voting ensembling, ensembling adversarial training and query detection to make the model strong. Cai et al. (2018) presented curriculum adversarial training as defense mechanism against adversarial attacks. The model they used was the reproduction of Madry et al. (2017) models'. They could see the improvement of 25% to 35% in their model with the help of CAT. Sitawarin et al. (2021) studied the major problems in tradition adversarial training which are high clean accuracy and small generalization gap. They proposed curriculum-based formulation of adversarial training. And their model outperformed adversarial training by 23% and 3% normally. In recent study, Oh and Kumar (2022) presented that RSs are vulnerable to untargeted attacks and named it as CASPER. They discussed that how a small perturbation could lead to a great effect on the accuracy and performance of RSs.

We have thoroughly reviewed the related articles in this research area in order to get better understanding of our problem and importance of our research. We observed that a number of researches had been published when it comes to adversarial attacks on CF-based recommender systems, Contrary to the relatively small number of published studies on adversarial attacks against content-based recommender systems. Therefore, our aim is to present a deep learning visual content based recommeder system by using Tuinhof et al. (2018) method. We decide to develop our recommender system by using two stage method same as in the referenced article. CNN and kNN will be used to create our RS and then white-box attacks would be used to demonstrate the vulnerability of RS to adversarial attacks. Additionally, the effectiveness of defense techniques (adversary training techniques) in protecting RS from these actual threats will be illustrated.

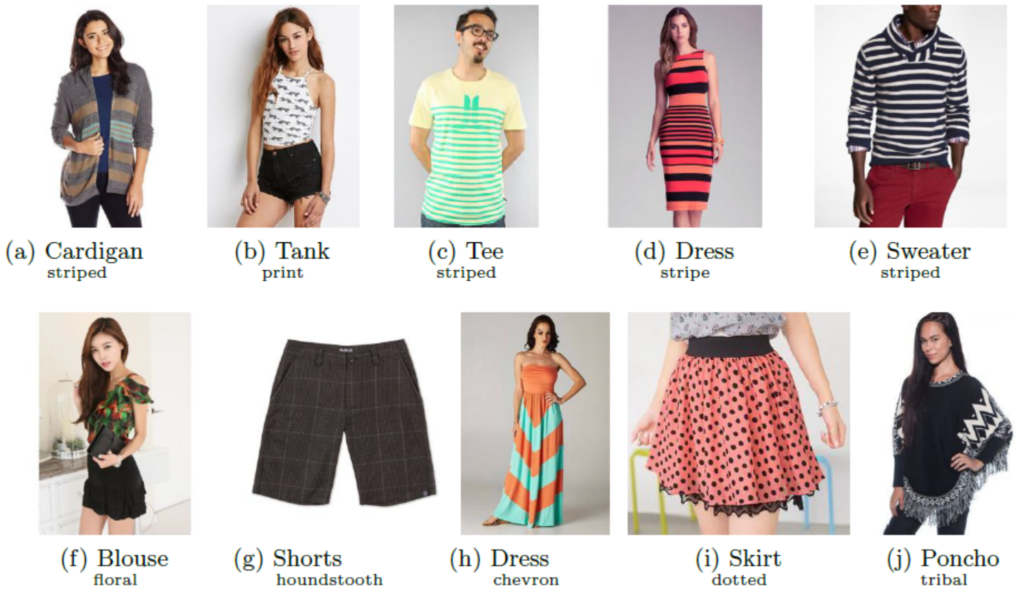
3 Methodology

The five steps of the study methodology include data collection, data pre-processing, data modelling, attacking, and training model, evaluation are discussed in this section.

For the first step, data collection consists of acquiring and measuring data. In this study, we used publicly accessible a large-scale clothing dataset 'DeepFashion' [1]. A few of the images from the DeepFashion dataset are shown in the figure 2.

The second step, data preparation and data pre-processing involve techniques to be performed on data in order to ensure that the data is clean for further steps. For data preparation,

[1] https://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html



(a) Cardigan
striped

(b) Tank
print

(c) Tee
striped

(d) Dress
stripe

(e) Sweater
striped

(f) Blouse
floral

(g) Shorts
houndstooth

(h) Dress
chevron

(i) Skirt
dotted

(j) Poncho
tribal

Figure 2: Some randomly sampled images from DeepFashion dataset

we used Google Drive to import the data and Google Colab to execute the scripts to retrieve the images systematically. After data preparation, data pre-processing techniques were performed to remove unlabelled data in texture, duplicate images, noise, and irrelevant data.

The third step, data modelling involves training and testing of model. For visual content-based recommeder system model, we reproduced Tuinhof et al. (2018) two-staged model approach. In first stage, CNN was used and trained for category and texture prediction of the clothes at the same time. The use of CNN involved input image's features extraction and return of feature vectors. These vectors were used to obtain an embedding matrix for utilizing in kNN ranking algorithm. In second stage, kNN found similar nearest neighbor and for similarity measures cosine distance was used.

The fourth step consists of adversarial attacks and defenses execution. Three different white-box adversarial attacks- Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) and the last attack Carilini & Wagner (CW) executed on the trained RS. And after executing attacks, for defense methods recommender system was trained against adversarial attacks by using Adversarial Training (AT) and Curriculum Adversarial Training (CAT).

The last step, evaluation consists of measuring the performance of a deep learning visual content-based recommender system based on success rate metric. If the adversarial object's recommendation rank falls below the $rank_{min}$ minimal threshold $\{rank(F(A+\delta), F(T))\}$ among the kNNs for the target, the attack would be considered as successful. Consequently, the equation for n number of attack tuples:

$$\text{success rate} = \frac{1}{n}\sum_{i=0}^{n} 1\{rank(F(A_i+\delta_i), F(T_i)) \leq rank_{min}\} \tag{1}$$

The performance of undefended RS and adversarial trained RS was compared according to success metric in the end. The methodology described above has been summarized and illustrated in the figure 3.

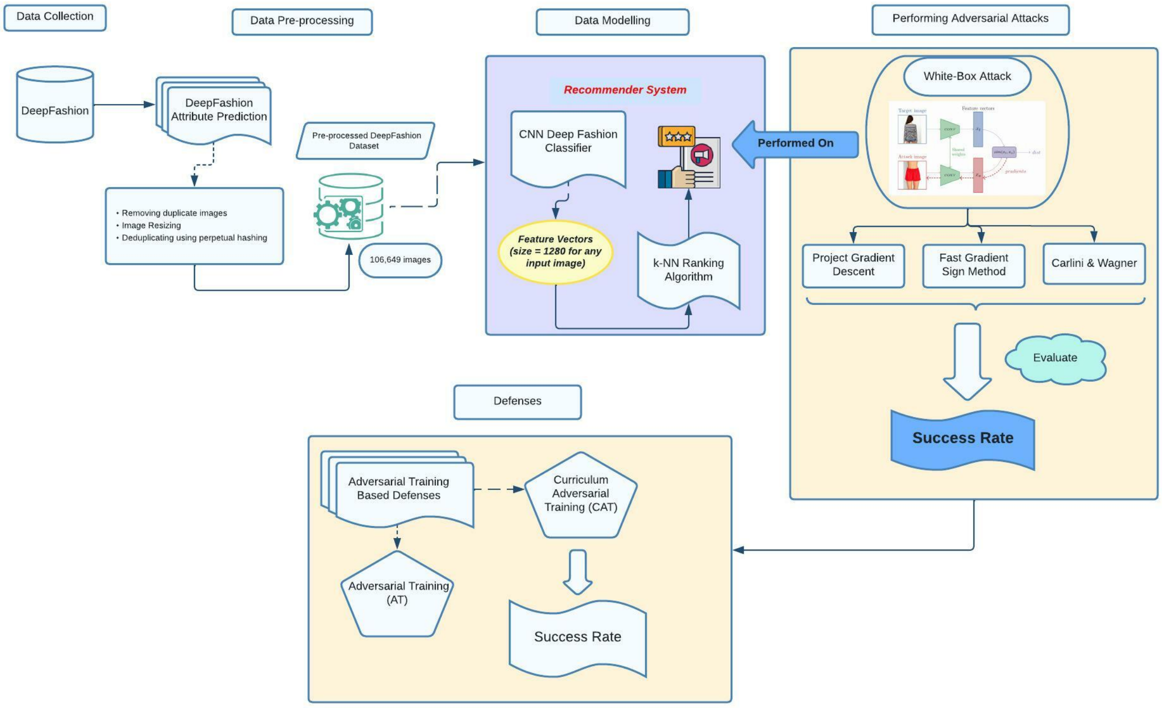

Data Collection
Data Pre-processing
Data Modelling
Performing Adversarial Attacks
DeepFashion
DeepFashion Attribute Prediction
Pre-processed DeepFashion Dataset
• Removing duplicate images
• Image Resizing
• Deduplicating using perpetual hashing
106,649 images
Recommender System
CNN Deep Fashion Classifier
Feature Vectors (size = 1280 for any input image)
k-NN Ranking Algorithm
Performed On
White-Box Attack
Project Gradient Descent
Fast Gradient Sign Method
Carlini & Wagner
Evaluate
Success Rate
Defenses
Adversarial Training Based Defenses
Curriculum Adversarial Training (CAT)
Adversarial Training (AT)
Success Rate

Figure 3: Methodology followed for the research

4 Design Specification

A deep learning visual content-based RS model was the combination of CNN and kNN methods in architecture of deep learning framework as shown in figure 4. Deep learning-based RS model is explained in section 4.1, the adversarial attacks and defenses are also explained in the following sections:

4.1 Deep learning visual content-based Recommender System

The use of CNN was for the classification of images, to identify category and texture of clothes, and for feature extraction. MobileNetV2 acted as the backbone of our CNN's architecture. We used the pre-trained network's weights from ImageNet as an initialization for the weights which is also known as transfer learning. MobileNetV2 trained on Adam optimizer with 32 batch size with learning rate 0.001. For measuring similarity cosine distance was used and to improve the performance of kNN approximate nearest neighbor index was used.

4.2 Adversarial attacks

Three different white box attacks were used in this study which were FGSM, PGD and CW. For FGSM adversarial method, the cosine distance was minimized between image embedding identified by CNN's feature extractor. FGSM was targeted item-to-item attack performed on recommender system. PGD attack is an iterative method based on *Basic Iterative Method(BIM)*, PGD also initialized an uniform random perturbations to the clothes pictures then iterations executed to search adversarial examples. CW used three attacks for decreasing similarity metrics L_0, L_1 and L_∞. CW is one of the strongest and very accurate approach for producing perturbations. However, its computing performance is slow.

4.3 Defense methods

To make our system robust we need to prevent our model form such attacks. There are two defense methods discussed in this study which are Adversarial training and Curriculum Adversarial Training. During adversarial training, to decrease the effectiveness of the model CNN was paired with two potent attacks for inaccurate categorization. In order for the model to be able to learn the information needed for stronger decisional limits. CAT technique is a method of AT's class intend to increase accuracy for considerably more difficult tasks on both clean and adversarial data and executed on model same as AT. These methods trained the model using adversarial examples and made the system more robust.

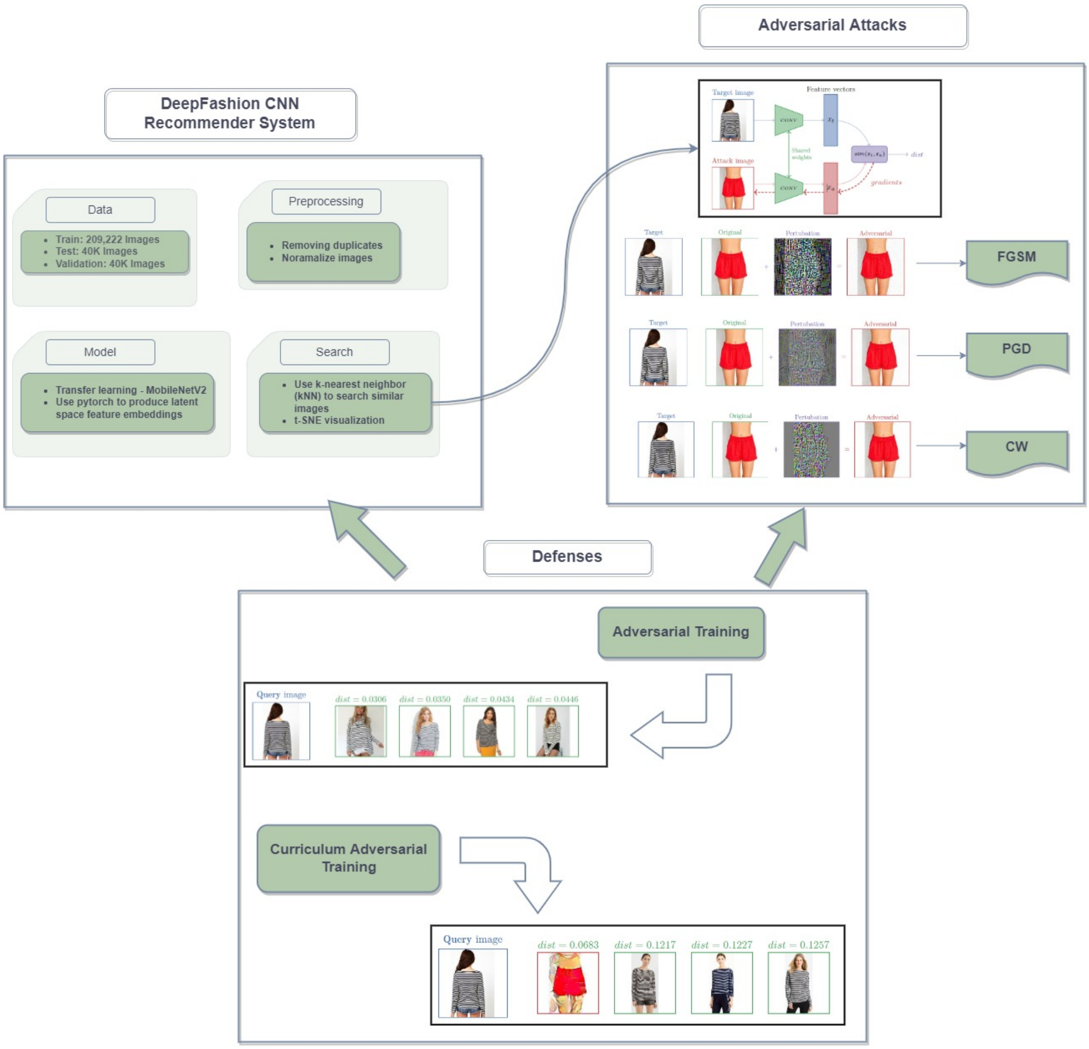


Figure 4: Design architecture for Adversarial Attacks

5 Implementation

This deep learning framework was implemented in steps. In first step, CNN and kNN machine learning techniques were used to process images and discover nearby objects that are similar to the garments image in RS model. CNN extracted image's features for classifying them and kNN used for finding similar image for recommendation. RS was trained and tested by using clothing images in DeepFashion dataset. The undefended model was attacked by FGSM, PGD and CW white-box targeted attacks. Figure 5 demonstrates implementation of adversarial attacks. By utilizing success rate metrics, the effect of attacks was assessed in order to determine the effectiveness and system efficiency declines. The next step consisted of training RS with adversarial examples with the help of AT and CAT methods. In the last, performance of model was compared with ordinary and trained RS. The positive impact of defense methods was evaluated using success rate metrices.

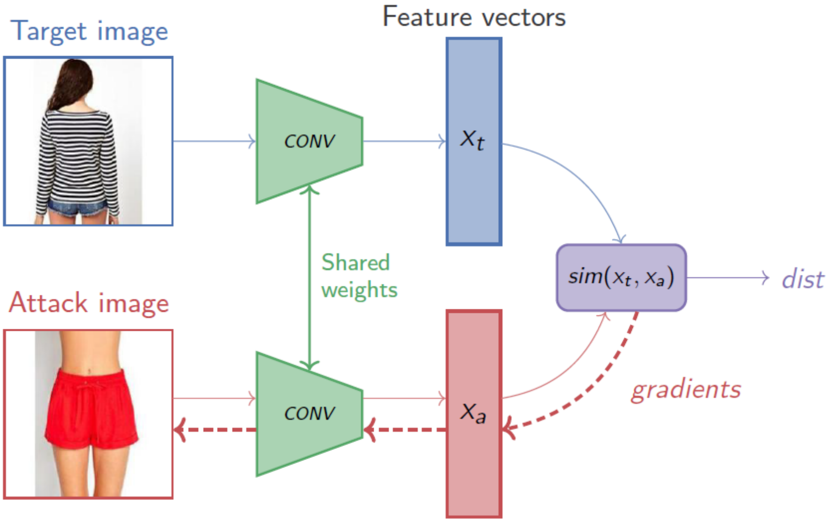
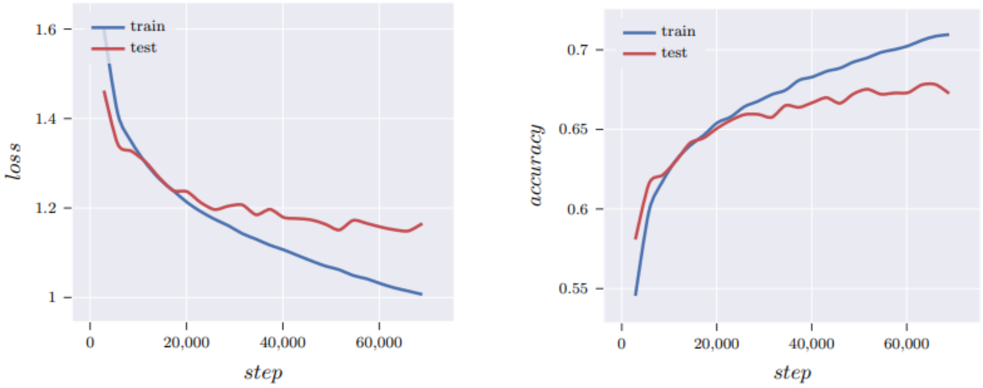


Figure 5: Implementation of adversarial attacks

6 Results

6.1 Recommender System

For this RS, CNN was used for classifying images, to predict category and garment's texture, and kNN for searching similar images based on the features extracted by classifier. In referenced article Tuinhof et al. (2018), two architectures of CNNs were evaluated which were: 1) Inception with batch normalization and 2) AlexNet. After that, improved architectures were presented such as ShuffleNet and MobileNetV2 and we have decided to use MobileNetV2 for our CNN by considering its superiority. We employed MobileNetV2 Adam optimizer with 32 batch size and 0.001 learning percent. Figure 6 shows trained classifier's history.



(a) Plot of combined model loss on training and test dataset

(b) Plot of category accuracy on training and test dataset

Figure 6: Trained classifier's history with 24-epochs

- As can be seen in Table 1 (a), we improved the category classification accuracy by 5% compared to the original accuracy.

- The second texture classification task was approached as a multi-label challenge by the designers of the DeepFashion dataset. We chose to treat the texture label as a multi-label problem and use the entire dataset. For the loss function of the texture classifier, we used the binary cross-entropy function. The highest anticipated textures out of 156 probable classes were classified by our network with an accuracy of about 44%, as shown in **??** (b).

- For each image used as input, CNN created feature vectors with a size of F(X) d = 1280, which were then applied to all the images displayed in the dataset. y stacking the resultant vectors, an embedded matrix with nxd dimensions was produced for the kNN technique. Performance was improved by combining very powerful data structures called hierarchical navigable small world (HNSW) graphs with kNN algorithm. Malkov and Yashunin (2018) were the ones who first introduced this data structure.

Category	Our Model	Tuinhof et al. (2018)
Accuracy	68.25	63.00
Top-5 Accuracy	93.14	84.00
CE-Loss	1.09	1.27

(a) Category Classifier results for our model as compared to Tuinhof et al. (2018)

Texture	Our Model
Top-1 Precision	43.68
BCE-Loss	0.03

(b) Results for texture Classifier

Table 1: Results of test sets used to evaluate the multi-task *DeepFashion* classifier

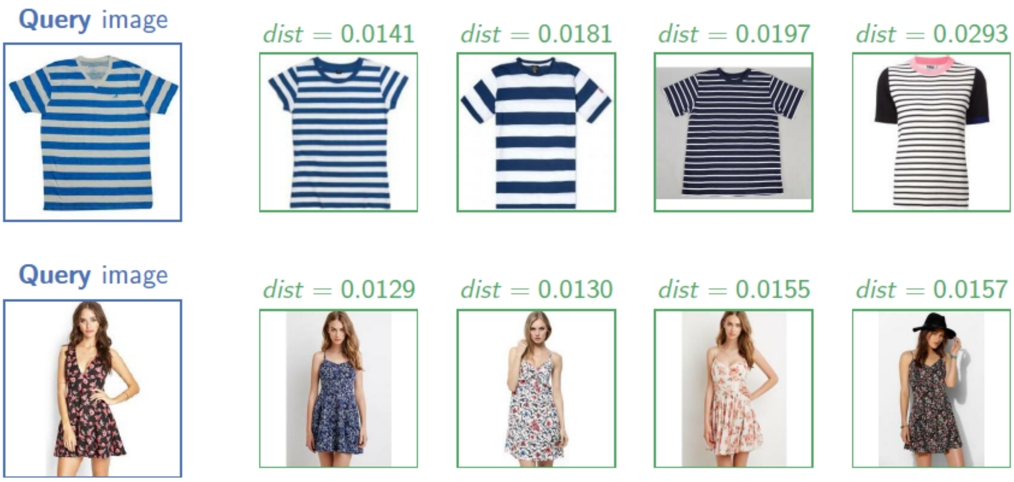


Figure 7: kNN ranking results for two items chosen at random

6.2 Adversarial Attacks

There are three different adversarial attacks- FGSM, PGD and CW which were executed on our recommender system and the results are discussed in this section.

1. **Fast Gradient Sign Method :** For FGSM, the cosine distance between picture embeddings produced by our CNN feature extractor was reduced and the maximum success rate achieved by this method was 0.27 in percentage with $rank_{min}$ as 3. It showed that it wasn't enough to achieve our adversarial goal, and the efficiency dropped as step sizes increased.

2. **Projected Gradient Descent :** The step size used for PGD was $\alpha = \epsilon/k$ where k represented num of iterations. To optimize the adversarial aim, we evaluated the success rate of PGD approach using k as 8, 16, 32, 64 and 128. We were able to achieve 97.09 percent success rate for $rank_{min}$ as 3 with 64 PGD iterations, demonstrated that an iterative optimization method like PGD is noticeably more effective at finding worst case perturbations that fully achieve our adversary's goal.

3. **Carlini & Wagner :** In the end, we tested strongest method CW for our adversary goal. We used Adam optimizer with learning rate of 0.005 and 1,000 optimization steps were performed. The highest success rate was 99.7 percent for $rank_{min}$ as 3 with 1,000 iterations and proved that the more advanced CW approach was more effective in locating worst-case perturbations that fit our adversarial goal.

The aggregated result of all three attacks is shown in Table 2 for $rank_{min}$ as 3 and $\epsilon = 0.3$.

		Success Rates for different ϵ		
Attack Method	$Rank_{min}$	$\epsilon = 0.01$	$\epsilon = 0.03$	$\epsilon = 0.05$
FGSM	3	0.27	0.14	0.07
	10	0.64	0.32	0.13
PGD	3	44.33	94.06	97.09
	10	50.21	95.9	98.22
CW	3	83.1	99.4	99.7
	10	86.6	99.5	99.9

Table 2: Summary of Attack Methods

Comparison : The comparison between all three attacks and their success rates are shown in graph 8.

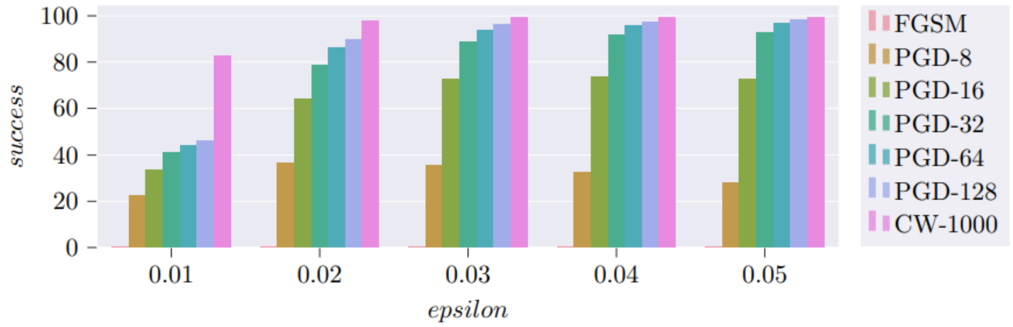


Figure 8: Success rate percentage for $rank_{min}$ as 3, computed for 10,000 randomized tuples (1,000 in CW method), for all assessed attacks and ϵ parameters

- PGD and CW achieved remarkably high success rates (almost 99.7%). FGSM achieved insufficient results and was not able to achieve the adversary aim.
- When attacking an undefendable model, the performance of CW was better than PDG-12 and was relatively small and arguably not worth the extra computational cost (approx. ×7.5).
- We believed that PGD-128 offers the optimal balance between attack power and computational expense.

6.3 Defense Methods

In the section, we applied two defense mechanisms (both are based on adversarial training) to train our CNN feature extractor and demonstrate how they affect our suggested attacks against our kNN recommender system.

1. **Adversarial Training:** During adversarial training, the adversary's goal was to raise the possibility of misclassification for categories and texture features. As with our standard classifier, we ran the adversarial training for 24 epochs. The accuracy on clean images suffers as a result of this increase in robustness. In our scenario, the adversarial trained model lost 12.19 percent of its accuracy in identifying the proper garment category on clean images, but we improved 48.69 percent for the identical job on adversarial images. We were capable to lower success rate metric of attack for the perturbation budget of ϵ =0.05 by utilizing adversarial training from 99.7% to only 0.3% and it started rising when considered perturbation budget of ϵ as 0.1 or higher. Again, CW technique significantly surpasses every other attack showcasing its greater efficiency. This experiment led to the conclusion that successful attacks were unfeasible due to the achieved drop-in success rate metric for ϵ values less than or equal to 0.05, and for higher values, users would be more likely to observe the tampering and depletion in image quality would be intolerable.

2. **Curriculum Adversarial Training:** The Curriculum Adversarial Training (CAT) was used to improve efficiency of clean and adversarial input images for more challenging tasks. By employing PGD attack to train a model with up to 8 iterations (k), we deployed and assessed this method, along with the batch mixing optimizations. We also limited L_∞ perturbations ϵ to 0.03 during this process. Table 3 displays the evaluated outcomes for this classifier. Results showed that CAT performed better than the traditional AT on clean data. However, it reduced the reliability against adversarial cases. By using CAT, the probability of worst case success metric reached to 32.8% same as we acquired using 1000 iterations in CW technique. The success rate metric increased with the increment in ϵ budget, peaking at a 99.50% for CW and ϵ = 0.3.

Training Method	Clean/Adversarial	Accuracy after Defense	Clean Accuracy	Difference
AT	Clean Accuracy	56.06	68.25	-12.19
	Adversarial Accuracy	48.71	0.02	48.69
CAT	Clean Accuracy	62.29	68.25	-5.96
	Adversarial Accuracy	27.45	0.02	27.43

Table 3: Summary of Defense Results

Comparison :

	Attack		
Defense	FGSM	PGD-128	CW-1000
Unsecured	0.07	98.32	99.70
Adversarial Training	0.03	0.07	0.30
Curriculum Adversarial Training	0.00	14.89	32.80

Table 4: Success rate percentage for $rank_{min}$ as 3, computed for 10,000 randomized tuples (1,000 in CW method), for all assessed attacks and ϵ as 0.05

- According to our experiments, we successfully lower the success metric of 0.05 perturbations budget by using conventional adversarial training. This finding demonstrated how employing adversarial samples can effectively boost a CNN feature extractor's robustness against adversarial attacks aimed at the underlying features space.
- CAT did succeed in improving robustness in contrast to an unprotected system, the improvement is much less pronounced than conventional AT (refer to table 4).
- We must point out that no quantified robustness metric proves general resistance against additional unidentified attacks; rather, it simply indicates robustness against the types of attacks and ϵ boundaries we investigated in our studies.

7 Conclusion

The main goal of this study was to examine the vulnerability of a deep learning visual content-based recommender system to adversarial attacks and also to analyse the impact of potential defense methods. In the first step, a visual content-based RS was produced by using DeepFashion dataset images. The next step consisted of creating and executing three targeted item-to-item attacks utilizing state-of-the-art white box techniques (FGSM, PGD and CW). Also, we evaluated how well they worked to compromise the integrity of the attacked undefended recommender system by using success rate metric. According to our results as shown in table 4, PGD-128 and CW attacks achieved 98.32% and 99.7% success rate which was extraordinarily high which demonstrated that deep learning based RSs are vulnerable to adversarial attacks. The following step involved testing two defense methods (Adversarial Training and Curriculum Adversarial Training) based on adversary training. According to our experiment, both trained models reduced the success rate of attacks, however, traditional adversarial training (AT) had significantly enhanced the robustness of our system in comparison of Curriculum Adversarial Training (CAT). It showed that adversarial training had positive impact on RSs against adversarial attacks and made the system more robust. The conclusion of our research is that deep learning based recommender systems are vulnerable to adversarial attacks and they've also shown hopes that adversarial robust recommender systems based on DNN might be feasible.

Despite the fact that our study showed a high reliability against our examined white-box attacks, it is uncertain if and to what extent these observations extend for black-box or upcoming unknown attacks. Additionally, there is an area to explore the impact of these attacks and defense mechanisms on deep learning based hybrid recommender system. These could be the areas and scopes for the future works.

References

Aggarwal, C. C. et al. (2016). *Recommender systems*, Vol. 1, Springer.

Bollacker, K. D., Lawrence, S. and Giles, C. L. (1998). Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications, *Proceedings of the second international conference on Autonomous agents*, pp. 116–123.

Cai, Q.-Z., Du, M., Liu, C. and Song, D. (2018). Curriculum adversarial training, *arXiv preprint arXiv:1805.04807* .

Chen, L., Yang, F. and Yang, H. (2017). Image-based product recommendation system with convolutional neural networks.

Cohen, R., Sar Shalom, O., Jannach, D. and Amir, A. (2021). A black-box attack model for visually-aware recommender systems, *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 94–102.

Dalvi, N., Domingos, P., Sanghai, S. and Verma, D. (2004). Adversarial classification, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108.

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T. and Zhu, J. (2019). Efficient decision-based black-box adversarial attacks on face recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7714–7722.

Hirano, H., Minagi, A. and Takemoto, K. (2021). Universal adversarial attacks on deep neural networks for medical image classification, *BMC medical imaging* **21**(1): 1–13.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition, *Neural computation* **1**(4): 541–551.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks, *arXiv preprint arXiv:1706.06083* .

Malkov, Y. A. and Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, *IEEE transactions on pattern analysis and machine intelligence* **42**(4): 824–836.

Melville, P. and Sindhwani, V. (2010). Recommender systems., *Encyclopedia of machine learning* **1**: 829–838.

Oh, S. and Kumar, S. (2022). Robustness of deep recommendation systems to untargeted interaction perturbations, *arXiv preprint arXiv:2201.12686* .

Pal, M., Jati, A., Peri, R., Hsu, C.-C., AbdAlmageed, W. and Narayanan, S. (2021). Adversarial defense for deep speaker recognition using hybrid adversarial training, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6164–6168.

Rakin, A. S., He, Z., Li, J., Yao, F., Chakrabarti, C. and Fan, D. (2021). T-bfa: Targeted bit-flip adversarial weight attack, *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Rathore, P., Basak, A., Nistala, S. H. and Runkana, V. (2020). Untargeted, targeted and universal adversarial attacks and defenses on time series, *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–8.

Sadrizadeh, S., Dolamic, L. and Frossard, P. (2022). Block-sparse adversarial attack to fool transformer-based text classifiers, *arXiv preprint arXiv:2203.05948* .

Sharma, S., Rana, V. and Kumar, V. (2021). Deep learning based semantic personalized recommendation system, *International Journal of Information Management Data Insights* **1**(2): 100028.

Sitawarin, C., Chakraborty, S. and Wagner, D. (2021). Sat: Improving adversarial training via curriculum-based loss smoothing, *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, pp. 25–36.

Sravani, C. S. and Kurumeti, N. K. (2021). Cnn based movie recommendation system.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2013). Intriguing properties of neural networks, *arXiv preprint arXiv:1312.6199* .

Tuinhof, H., Pirker, C. and Haltmeier, M. (2018). Image-based fashion product recommendation with deep learning, *International conference on machine learning, optimization, and data science*, Springer, pp. 472–481.

Uchendu, A., Campoy, D., Menart, C. and Hildenbrandt, A. (2021). Robustness of bayesian neural networks to white-box adversarial attacks, *arXiv preprint arXiv:2111.08591* .

Wei, Z., Chen, J., Wei, X., Jiang, L., Chua, T.-S., Zhou, F. and Jiang, Y.-G. (2020). Heuristic black-box adversarial attacks on video recognition models, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 12338–12345.

Ye, L. (2021). Thundernna: a white box adversarial attack, *arXiv preprint arXiv:2111.12305* .

Żelasko, P., Joshi, S., Shao, Y., Villalba, J., Trmal, J., Dehak, N. and Khudanpur, S. (2021). Adversarial attacks and defenses for speech recognition systems, *arXiv preprint arXiv:2103.17122* .

Zhang, C., Costa-Pérez, X. and Patras, P. (2022). Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms, *IEEE/ACM Transactions on Networking* .