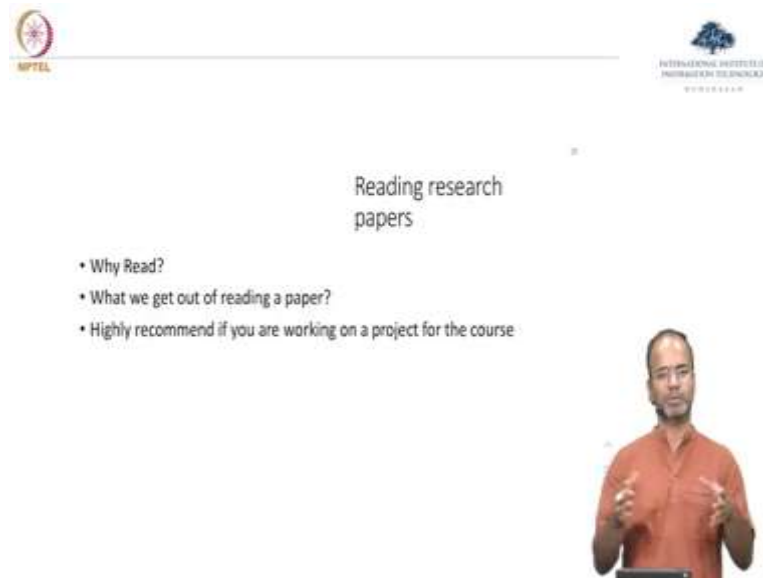


Online Privacy
Professor Ponnurangam Kumaraguru “PK”
Indian Institute of Technology, Hyderabad
Lecture 15
Research Paper Reading

(Refer Slide Time: 00:16)



Next part of this week is actually about reading research papers itself, one of the goals for this course is to also help you to understand how to read a research paper particularly within the context of privacy and secured by online privacy. And many meaning we have already seen the papers but I kind of discussed the content of the paper in the lectures, it will be nice now to flip it and see you actually do some readings of the papers and see how you can get understanding of topics from these papers.

Why read research papers? To understand what is going on in literature to know, what are the latest things, what are the normal work that people have done, why should you actually work on a problem, motivations for working on a problem all that can come from reading a research paper.

So, how do you place the work? Meaning the YouTube recommendation system that I mentioned, how do you actually know that it is actually novel nobody has done it, only the research papers can help us do that, help us understand that. Highly recommend if you are working on a project for the course, I think that if you are reading, I mean if it was a real class what I teach on campus I would let you read the papers, come discuss which I think I want to do this for the NPTEL model also which is I think later in the slides.

I have one paper that I posted there which if any of you are interested in reading it you read it, we will set up a time in the semester just around this week 7, where we can actually come and discuss the paper. Discussion on the paper also I have some ideas how better it can be done instead of just this model of one person reading the paper and coming and discussing the paper.

(Refer Slide Time: 02:01)

The slide is titled "Role-playing seminar" and lists four roles with their respective tasks:

- Scientific Peer Reviewer:** The paper has not been published yet and is currently submitted to a top conference where you've been assigned as a peer reviewer. Complete a full review of the paper answering all prompts of the official review form of the top venue in the research area (e.g., NeurIPS for Deep Learning and ACM SIGGRAPH for Geometry & Animation). This includes recommending whether to accept or reject the paper.
- Archivist:** This paper was found buried under ground in the desert. You're an archeologist who must determine where this paper sits in the context of previous and subsequent work. Find and report on one older paper cited within the current paper that substantially influenced the current paper and one newer paper that cites this current paper.
- Academic Researcher:** You're a researcher who is working on a new project in this area. Propose an imaginary follow-up project not just based on the current but only possible due to the existence and success of the current paper.
- Industry Practitioner:** You work at a company or organization developing an application or product of your choice (that has not already been suggested in a prior session). Bring a convincing pitch for why you should be paid to implement the method in the paper, and discuss at least one positive and negative impact of this application.

Source: <https://colerabel.com/blog/role-playing-seminar.html>

I stumbled on this like about a year back or no probably seven, eight months back which is to how reading research papers, different ways of reading research papers. We were interested in because I think if I mean general model is that you ask one student to come and you ask one student to read the paper, the one student will prepare some deck of slides, come, present the paper and others who can ask questions, that is the general way by which paper reading happens. One, two, three students do it together.

But here is another interesting way and I have been practicing this for last a semester or so in reading papers I think it is very effective not just with the students that I work with I also tried this in the class that I taught on campus it seems to have worked very well. How does it work?

Instead of one person reading the paper, there are going to be seven people reading the paper now and presenting it. Meaning I think the goal is that in a paper discussion everybody reads the paper because without that if you come to the paper reading discussion it is generally not going to be very useful. So, instead of one person presenting the paper, we are going to get like seven people to present the paper.

But these seven people will have different types of roles that they are going to play in discussing the paper. So, we should we will try this in the class in an online session or in an offline session also now that things may be better in the January 2022 semester again if people are interested in coming to campus and trying this out, we can try it. So, here is what the different roles are.

The first one is scientific peer reviewer, which is the paper has not been published and yet and is currently submitted to a top conference. Essentially the role of this is, the role of this particular viewer is to read the paper as though it is actually, they are reviewing the paper for a conference and giving suggestions on what good and what not good. So, generally a paper acceptance, paper reviews has this why should I accept the paper, why should I reject the paper, comments to the authors, all of that. Can the students reading the paper prepare that?

Archaeologist, this is an interesting one. This paper was found buried underground in a desert, you are an archaeologist who must determine whether this paper sits in the context of the previous or the subsequent work. So, the goal here is, a few minutes before I said, which is why do you need research papers, to place the research that you are doing, you need to understand what the literature is.

There is also this understanding that the quality of your research is highly dependent on the literature that you are aware of or even limited by the literature that you are aware of. So, find an report on a older paper cited within the current paper that substantially influenced the current paper blah, blah, blah. So, this basically archaeologist goal is to go look at, is the paper placed properly and then give their thought on the positioning of the paper.

Academic researcher, you are a researcher who is working on a new project in this area, propose an imaginary follow-up project, not just based on the current but only possible due to the existence and success of the current paper. So, this is basically an idea of okay now I want to do a follow-up study what would that be.

Industry practitioner, you work at a company or an organization developing an application or product of your choice, bring a convincing pitch for why we should, why we should be paid to implement the method in this paper. This is actually an interesting one because many papers do not get converted into a implementation.

So, the role of the industry practitioner is to say that look this paper does this, for example, let us take the Facebook not just paper, three nudges, the paper evaluated the three nudges

should we explore these types of nudges implement it in Facebook and actually try it out and see it with the users. Why would that, why should somebody try it, what is the, and discuss at least one positive or negative impact of this application. Let us take you build it and put it there why would users like it, why should users use it all that is a question that industry practitioners should look at.

(Refer Slide Time: 06:55)

The slide content is as follows:

- Hacker:** You're a hacker who needs a demo of this paper ASAP. Implement a small part or simplified version of the paper on a small dataset or toy problem. Prepare to share the core code of the algorithm to the class and demo your implementation. Do not simply download and run an existing implementation - though you are welcome to use (and give credit to) an existing implementation for "backbone" code.
- Private Investigator:** You are a detective who needs to run a background check on one of the paper's authors. Where have they worked? What did they study? What previous projects might have led to working on this one? What motivated them to work on this project? Feel free to contact the authors, but remember to be courteous, polite, and on-topic.
- Social Impact Assessor:** Identify how this paper self-assesses its (likely positive) impact on the world. Have any additional positive social impacts left out? What are possible negative social impacts that were overlooked or omitted?
- Archaeologist:** The paper was buried under a pile of papers in the drawer. You're an archaeologist who will determine where this paper sits in the context of previous and subsequent work. What role does it play in the literature? What are the key contributions that distinguish it from other papers and how does it relate to other papers that cite this research paper?
- Academic Researcher:** You're a researcher who is working on a new project in this area. Prepare an imaginary follow-up project not just based on the current but only possible due to the existence and success of the current paper.
- Industry Practitioner:** You work at a company or organization developing an application or product of your choice that has not already been suggested in a prior session. Using a concrete use case, who you should be paid to implement the method in the paper, and discuss at least one positive and one negative impact of this application.

At the bottom of the slide, there is a URL: <https://www2019.thw.ac.uk/track/track.asp?track=2019>

So, that was the four roles, the rest of the three roles are these. The first one is the hack the, sort to say, fifth one is the hacker. In this role you are a hacker who needs to, who needs a demo of this paper as soon as possible. Implement a small part or a simplified version of the paper on a small data set or a toy problem, prepare to share the code of the algorithm to the

class and demo your implementation, do not simply download and run an existing implementation.

So, the idea here is that hackers should try out what is going on by themselves and bring the code, show the code, walk through the code all of that to be done. Though you are welcome to use an existing implementation for backbone code, it is helping users, helping the student who is reading the paper as a hacker to get a sense is to actually build it. It is slightly intense than how naturally paper reading is done.

Private investigator, you are a detective who needs to run a background check on one of the paper's authors, where have they worked, what did they study, what previous projects might have led to working on this one. So, essentially this private detective is to doing a background work on the authors.

He or she was the Ph.D. student at this institute and they were working with the last author, last author being the postdoc advisor or the Ph.D. thesis advisor, before this paper this to the first author actually wrote another paper which is published in this conference and that paper was also on privacy blah, blah, blah, blah, blah. So, that is how a private investigator would do.

Social impact assessor, I think the name itself gives it away. Identify how this paper self-assesses its impact on the world. Have any additional positive impacts left out? What are the positive, negative possible negative social impacts that were overlooked or omitted?

So, interesting things that roles do and I have been using this model for last about a semester reading papers and then playing these roles, coming up with hacks, coming up with actually social impact assessment all that, it is been very, very useful thanks to the blog which actually talks about methods to do these seven roles. And I would like to try out in the class also. So, we will see if we can actually do a session where some of you read the paper in these roles and come.

(Refer Slide Time: 09:36)



The slide features a title 'Loose Tweets: An Analysis of Privacy Leaks on Twitter' with logos for 'MPTEL' and 'INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY'. Below the title is the authors' information: 'Huina Mao, Xin Shuai, Aou Kapadia, School of Informatics and Computing, Indiana University Bloomington, Bloomington, IN, 47401, USA, [huinmao, xshuai, kapadia}@indiana.edu'. The slide is divided into sections: 'ABSTRACT' (discussing privacy leaks on Twitter), 'General Terms' (Human Factors, Measurement, Security), 'Keywords' (privacy, leaks, social networks, Twitter), and '1. INTRODUCTION' (starting with 'As a microblogging service, Twitter has become one of the most popular social networking tools today...'). A small video inset on the right shows a man in an orange shirt speaking.

So, I put a paper that we should do as part of this week's content itself. Please read this paper it is not a very intense paper, very simple, very, very highly cited paper in that sense but very, very intuitive, nicely done paper. So, if you can read this paper, we can actually do these roles and see how well you are able to capture the idea of reading the papers that are around privacy also.

So, to help you understand how these roles are done, the next 30 minutes or so. A paper was taken and that paper recording how students did interact with these roles is recorded and I have actually put it as part of this video itself. So, please watch it before you actually start reading it yourself, reading this paper for yourself and see how these roles can be done, it will be super nice if you try it once a couple of papers for the class also.

(Refer Slide Time: 10:45)

Fawkes: Protecting Privacy against Unauthorized Deep Learning Models

Shawn Sharif, Emily Wengerf, Jayun Zhang, Huiying Li, Haibo Zheng, Ben Y. Zhao

Motivation

- Facial Recognition Models are easy to build
 - Less time required to train models
 - Labeled data everywhere
 - Cheaper hardware
- Anyone with limited coding knowledge and computational power can train powerful facial recognition models
- Facial Recognition Models Easily Misused - one such example is Clearview

GAIN INTELLIGENCE. DISRUPT CRIME.

A laptop using F2 models without consent

Student 1: Good evening, everyone I will be presenting Fawkes protecting privacy against unauthorized deep learning models today. Due to the developments in deep learning facial recognition systems are successfully scanning millions of citizens in different countries without explicit consent.

This situation is compounded by the fact that facial recognition systems are easy to build. The models are only getting faster to train and the hardware to train them is getting cheaper. It is also extremely easy to misuse these models for your own game.

(Refer Slide Time: 11:14)

Example of Misuse of FR systems: Clearview

The Secretive Company That Might End Privacy as We Know It

Known: Clearview.ai customers include government agencies, law enforcement departments, and private citizens.

> Using FR systems, user privacy is violated in an unfair manner

One such example is clearview.ai which some of you must have heard about. It has up to 3 billion images in their database, they scrape this data from social networks without user consent, violating user privacy in an unfair manner. This data can be used for malicious purposes in the wrong hands, such as for extortion and stalking. Thus, securing our data against such usage is of paramount importance. To that end in this paper the authors provide a defence against such attacks called Fawkes to protect people from being identified by unauthorized facial recognition models.

(Refer Slide Time: 11:53)

The Idea

The authors present Fawkes, which tries to fool Facial Recognition systems.

Original Images → Fawkes (Clashed Images) → Model Training → Testing Data (Clashed) → Wrong Label

- To make the system mislabel images after deployment - Clash all images before uploading them on social media.
- Ensure that visually the face remains the same after 'clashing'
- The model trains on such 'clashed' images, leading it to predict incorrect labels during testing.

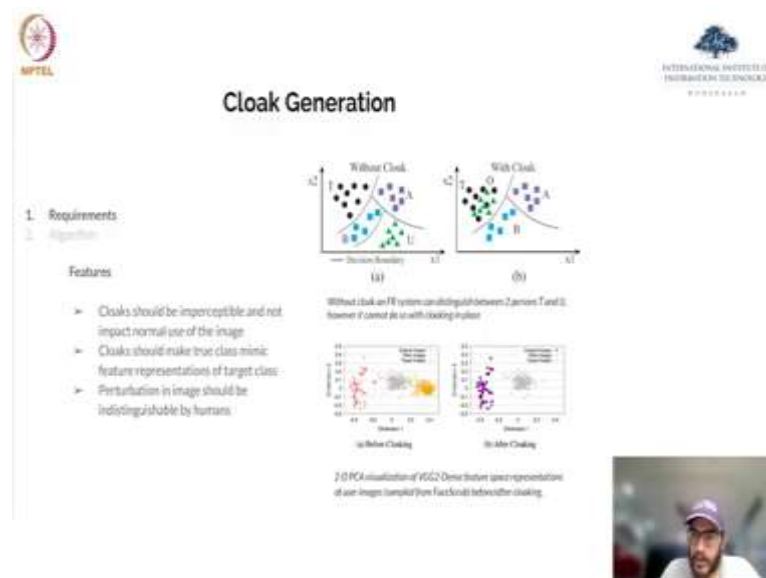
First let us understand how exactly this data is obtained and misused. The users upload images on the web to some social media such as Facebook, Twitter, etcetera, which are then

further scraped by a web crawler. These images are then used to train a facial recognition model with your images.

To circumvent this, the authors propose that users use Fawkes to first cloak images before uploading it to the web. This would help fool the model when these images are scraped as the model will be trained on false training data. And when a clean image is compared with this data it will not be able to recognize this person.

Since the features the, since the feature space is essentially changed by this cloaking, this helps the model to realize, to get fooled when the real image is presented to it and we shall see more of this in the next slides. An important point the authors note, is that these images do not change visually after cloaking. That is, they maintain the usability.

(Refer Slide Time: 13:01)

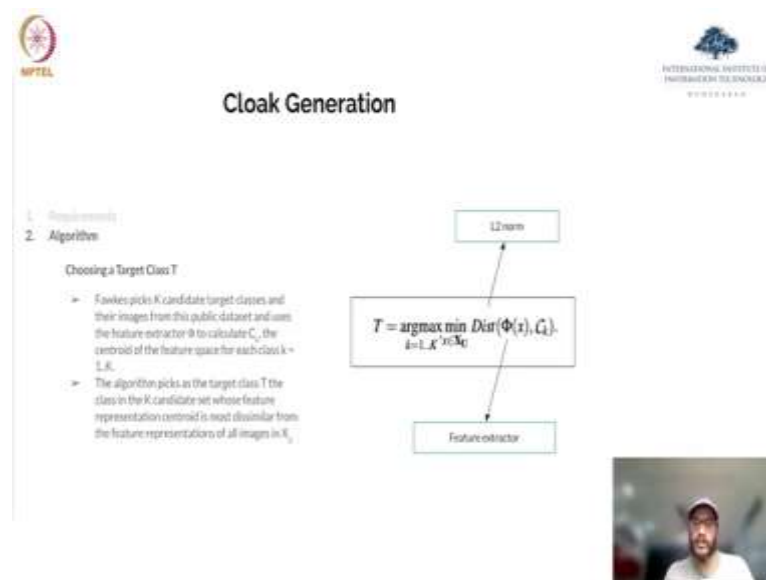


So, let us understand how these cloaks are generated and what are the requirements for the same. First cloak should be imperceptible and not impact normal usage of the image. Secondly, cloaks should make two class mimic feature representation of target class which we will talk more about later.

Third perturbations in images should be indistinguishable by humans. The images on the show that without cloaking the feature extractor can distinguish between two classes U and T. But with the feature extractor we can confuse the model from a class U to T. As you can see the decision boundaries between class U and class T is now merged and it is not able to distinguish between the two classes.

Similarly, in the image below you can see something similar happening before cloaking the target class which is x and the original images which is shown by this delta this yellow colour on the right, are separate, but after cloaking both of them are present in the same feature space.

(Refer Slide Time: 14:03)



So, how should we go about choosing a target class T ? We can choose any random class but there is a better way to choose these classes and the authors propose an algorithm to do the same. Firstly, to choose this class T we want to select the class which is most dissimilar from the initial class. This helps the model; this helps the cloak to fool the model in a better manner.

The authors do this by taking K candid target classes and their images from a public data set and use a feature extractor file to calculate the centroid of the feature space for each of the class. Then the centroid, then to choose the centroid representation which is the most dissimilar for the initial class, let us say that class is U .

They calculate this distance using the L2 norm which is essentially the root over the squared difference of all dimensions of the feature space and the feature extracted can be anything you can use ResNet, you can use whichever feature extractor that you want, and the authors compare different feature extractors in their results.

(Refer Slide Time: 15:11)

The slide is titled "Cloak Generation" and features the MITEL logo on the left and the logo of the International Institute of Information Technology, Hyderabad on the right. It is divided into two main sections: "1. Requirements" and "2. Algorithm".

1. Requirements

2. Algorithm

Computing Per-image Cloaks

- Let X_t represent the set of target images available to user U .
- For each image of user U , $x \in X_t$, Fawkes randomly picks an image $x_c \in X_c$.
- It then computes a cloak $\delta(x, x_c)$ for x , subject to $\|\delta(x, x_c)\| \leq p$.

The algorithm is visualized by a box containing the following formula:

$$\min_{\delta} \text{Dist}(\Phi(x), \Phi(x_c \oplus \delta(x, x_c))) + \lambda \cdot \max(\|\delta(x, x_c)\| - p, 0)$$

An arrow points from the formula to a box labeled "Structural Dissimilarity Index".

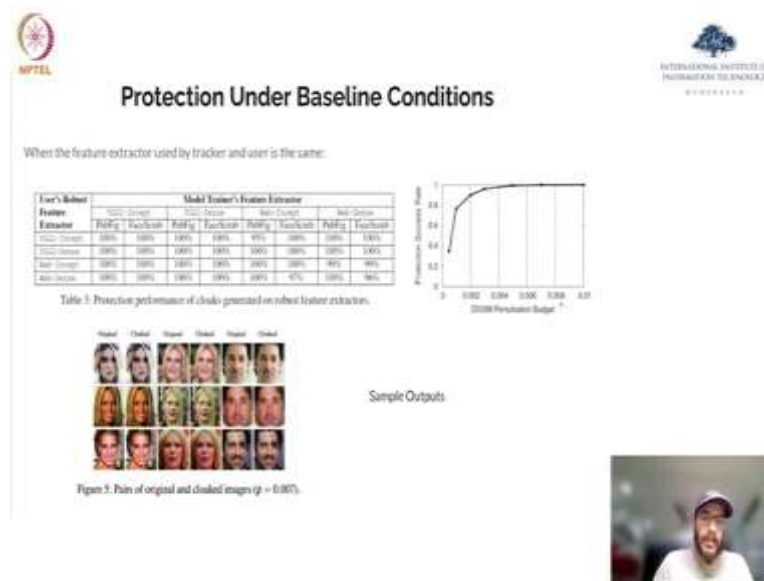
In the bottom right corner, there is a small video inset showing a man with a beard and a white shirt.

So, now that you have chosen this class T , we want to understand how this cloak is generated, what perturbations do you do to make sure that your images are a cloak. To that end the authors use a structural dissimilarity index to calculate this cloak. This index maintains the usability of images that is, you will not recognize the difference, visually you will not be able to recognize the difference between two images but the pixels are edited slightly to make sure that these images are cloaked.

They make sure that this, this difference between the two images which is called $\delta(x, x_T)$, right here in the formula this is less than p , which is a value that they keep which is the perturbation value that they have to keep below a certain amount such that visually the images do not change.

So, essentially, they want to transform one image class to another class, where maintaining a certain degree of similarity with the initial class. And so, they use this formula where they keep, where they minimize this, they minimize this difference δ and transform one feature space to the other, maintaining some properties of the initial feature space.

(Refer Slide Time: 16:24)



The authors show their efficacy of their model in different conditions. They call the baseline conditions the one where the extractor used by the tracker or the hacker if you want to call them and the user is the same. In this case the authors show that the cloaking works in almost 100 percent with almost 100 percent success rate, in all of the models that they use as you can see if they use vgd2 with inception, vgd2 with Dense.

The accuracy with which they are able to cloak is almost 100 in each of the cases. And if you look at this figure down below you will see that the cloaked images and the original images are pretty much the same, we cannot visually tell the difference between the two, for us both of these images remain the same but there is slight variations in them so that the facial recognition model which is trained on say one image let us say on the cloaked image is not able to recognize the original image when compared against it.

On the right-hand side, you can see this graph where based on the perturbation you can see the protection success rate. So, if you increase the perturbations from say 0.002 to 0.001, the success rate for protection increases but an important point to note here is that the images, so the cloaked image and the original image will start appearing dissimilar if you approach this value 0.001. So, that is why you need to keep this budget smaller, you need to either stay below 0.008 to maintain, to visually maintain the similarity between the original and the cloaked images.

(Refer Slide Time: 17:59)

Protection Under Realistic Conditions

- Realistically, we will not know the feature extractor used by tracker
- It is possible that the tracker uses transfer learning or trains a network from scratch
- In both cases protection rate stays above 95%
- Works well even with Deployed APIs available online

Face Recognition API	Protection Success Rate		
	Without protection	Protected by normal cloak	Protected by robust cloak
Microsoft Azure Face API	0%	100%	100%
Amazon Rekognition Face Verification	0%	100%	100%
Face++ Face Search API	0%	0%	100%

The slide also contains a line graph showing Protection Success Rate vs. Number of Labels in Student Dataset, and a scatter plot showing Dimension 2 vs. Dimension 1 for Original, Cloak, and Robust images.

So, under realistic conditions we cannot really hope to assume that we will have, that we will have the same feature extracted in both cases, that when a tracker is using a feature extractor and you are using a feature extractor that they would be the same. Realistically they would be different, you do not know how the facial recognition system is trained.

Because since it is possible for the tracker to even use transfer learning or scratch a model from scratch it makes sense that the feature, that the feature extractor they use will be different from yours. So, the conditions that we saw earlier were not realist, were not realistic at all, these conditions are baseline conditions, they will not happen in real life.

So, the authors present their cloaking technology, the cloaking method in this case as well. And they show that in all of these cases, in cases where a tracker trains the model from scratch or uses transfer learning from pre-trained model. For example, they are using ResNet and using transfer learning on that model the protection rate remains above 90 percent in those cases as well.

It even works well in deployed systems. For example, if they try to Microsoft Azure Face API, Amazon Recognition Face Verification and Face Plus Plus. As you can see down here, they used all three of these APIs and protect and with protection they were able to achieve 100 percent success rate.

(Refer Slide Time: 19:25)

Protection Under Realistic Conditions

- Since the tracker has original, un-cloaked images, the model struggles
 - The protection success rate drops below 39% when more than 15% of the user's images are uncloaked.
- Sybil Accounts help boost protection
- Detection of Cloaked effects is possible using
 - Image Transformation
 - Anomaly Detection

These methods however are not as efficient as increasing perturbations helps mitigate these effects.

Figure 16: When the user's feature extractor can distinguish the difference, the tracker can separate the protection success rate by increasing the 100% success of the 100% success. (Caption: Not visible)

Figure 17: Protection success rate decreases when the tracker has more original user images. (Caption: Not visible)

Figure 18: Protection success rate is high when the user has a Sybil account, even if the tracker has original user images. (Caption: Not visible)

Figure 19: Protection success rate is high when the user has a Sybil account, even if the tracker has original user images. (Caption: Not visible)

So, if you go even more in depth in case, so in a realistic condition we can expect that some of our images are uncloaked which are presented to the model. Since we already have images on Facebook even if we cloak our images right now, we will have some uncloaked images that would have been scraped by the web crawler. So, authors consider that condition as well and they run some experiments and they show that the success rate drops below 39 percent, if more than 15 percent of the user's images are uncloaked.

Now, this is a little problematic because most of our images are already online, so either we have to delete them and upload cloaked images which is not, which is not feasible. So, the authors also present a different method to a sort of a hack to help you maintain these images online which are also, which are uncloaked. These are called Sybil accounts.

Sybil accounts are essentially a duplicate account that you have on the same social networking site for example on Facebook I will have one original account and one duplicate account, where I will be uploading similar images, this helps because when the web crawler is crawling images from a social network, they will crawl images from your Sybil account as well as your original account, thus, providing you more protection if your Sybil account has cloaked images.

So, this essentially helps, this essentially mitigates that effect that if you have some uncloaked images online which are already available to the web crawler, these accounts help you mitigate that. As you can see in this image down below without Sybil accounts the feature extractor is able to distinguish and create a decision boundary between different images of class, they are able to essentially distinguish.

For example, if you look at the green deltas these are leaked images of you and these are the cloaked images of you and they are able to distinguish between the two. However, with Sybil this distinguished, they are not able to distinguish between these two classes.

(Refer Slide Time: 21:30)

Conclusion

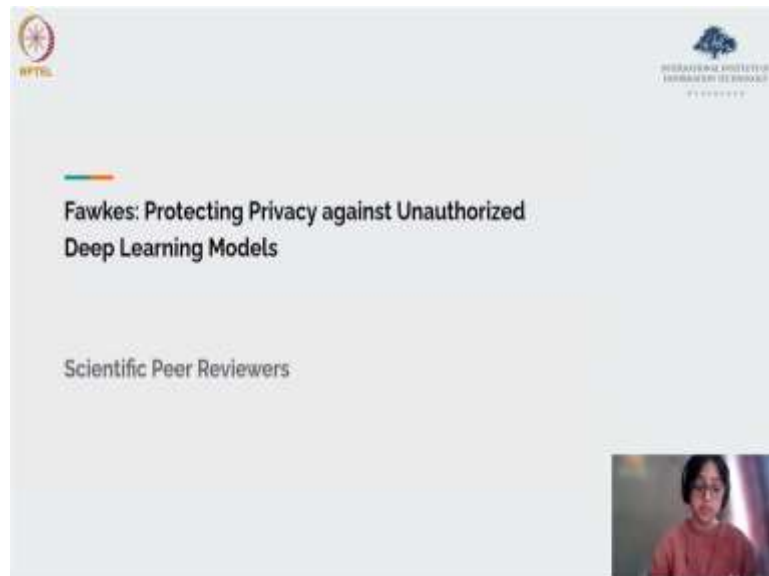
- Fawkes can provide protection against Facial Recognition systems in the real world
- The images generated are visually similar to originals, thus, making it hard to recognise cloaked images from uncloaked ones (on right)
- The model works well in all cases as long as uncloaked images are not available to the model
- If uncloaked images are available, the efficacy of Fawkes decreases, but it is helped by Sybil accounts.

Figure 3: Pair of original and cloaked images ($\rho = 0.807$)

The slide features the MITEL logo on the top left and the International Institute of Information Technology logo on the top right. A grid of 12 face images is shown, with the first column labeled 'Original' and the second column labeled 'Cloaked'. Below the grid is a small video frame showing a man in a white shirt and a cap.

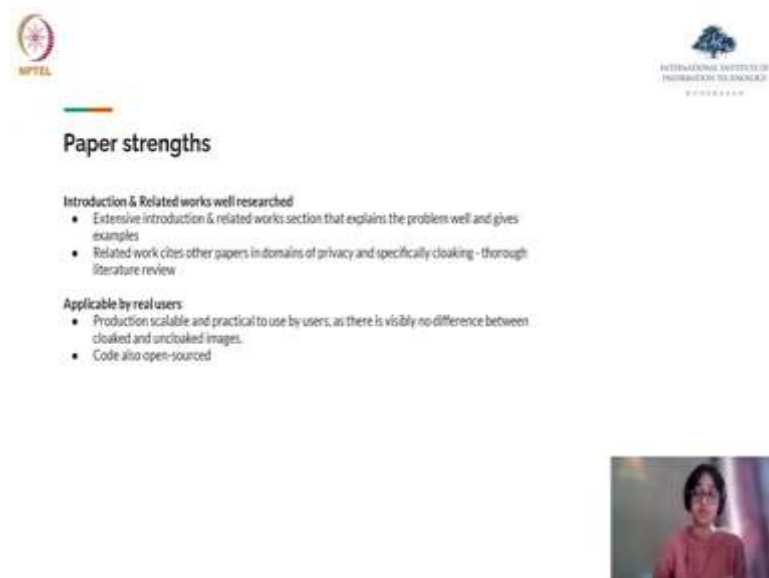
And so yeah so just to conclude Fawkes can provide protection against facial recognition systems in the real world. The images generated are visually similar to originals, thus, making it hard to recognize images that you have cloaked versus images that are uncloaked. The model works well unless some uncloaked images are available to it miss the features purpose but the authors also present an alternate way to secure images which are called Sybil accounts which essentially help protect your privacy even if uncloaked images are available to the web crawler. Thanks, that is it from me.

(Refer Slide Time: 22:37)



Professor: Yeah, thanks, thanks to Shawn, I think that was fantastic for getting the summary of the paper and now we have academic reviewer. As I said earlier in the video, there are different roles that we will see in students playing in reviewing this paper, giving what they think about the paper. So, now we will have reviewers talk about what they think of the paper.

(Refer Slide Time: 22:41)



Student 2: Scientific reviewers, we complete a full review of the paper and recommend whether the paper should be accepted or rejected in the conflict. So, some of the paper strengths which we will discuss here are, first point, the introduction and the related works are very well researched.

There is an extensive introduction and related work section that spans over almost five pages that explains the problem well and gives real world examples. The related work cites other papers in the domain of privacy and specifically cloaking which and there is a very thorough literature review in the field.

Second point, the paper is the model is applicable by real world users. The model is production scalable; it is practical to use by users as there is visibly no difference between cloaked and uncloaked images. The code of the model is open source, which implies that anyone can use it or modify it according to their own, according to their own case.

(Refer Slide Time: 23:34)



Paper strengths

Pertinent problem statement

- Highlight the importance of protecting one's images and the danger of face recognition models
- Give real world examples such as Clearview.ai

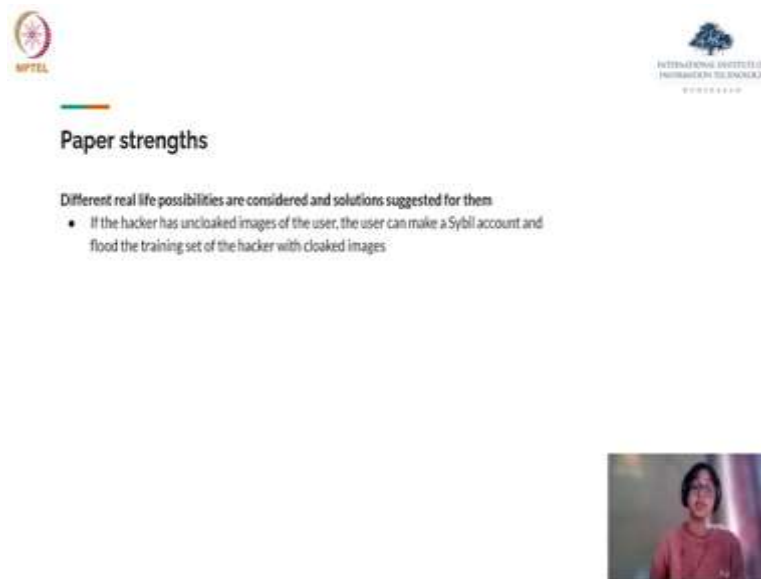
Highly reliable and robust model

- Provides 95% plus chances of success irrespective of how the model is trained.
- 100 % success of robust cloaking against state-of-art facial recognition services from Microsoft (Azure Face API), Amazon (Rekognition), and Face++.
- Effectiveness of robust cloaking remains same after data augmentation and transformations



The model, the paper has a very pertinent problem statement. It highlights the importance of protecting one's images and the danger of face recognition in the present world scenario by giving real world examples such as of clearview.ai, the model is highly reliable and robust it provides a 95 percent plus chance of success irrespective of how the face identification model is trained, it provides 100 percent success of robust cloaking against state of art facial recognition services from Microsoft that is the Azure Face API, Amazon which is the recognition model and Face Plus Plus. The very important point here is that the effectiveness of robust cloaking will remain same after data augmentation and transformations.


(Refer Slide Time: 24:27)



Paper strengths

Different real life possibilities are considered and solutions suggested for them

- If the hacker has unclocked images of the user, the user can make a Sybil account and flood the training set of the hacker with cloaked images



Another the last point is that different real-life possibilities have been considered in the model and solutions have been suggested for them. For example, in the case where the hacker has unclocked images of the user, the user can make a Sybil account to flood the training set of the hacker with cloaked images. Thus, making the face identification model perform poorly and keeping the identity of the user private.

(Refer Slide Time: 24:54)



Paper weaknesses

Can be misused by malicious agencies to hide their identity.

- Conveniently leave discussing this trade-off between user privacy and authorized use to future work.

Not as useful if the user's pictures are already online. Cumbersome for users to upload more.

- User needs to be proactive by making Sybil accounts, and remembering to pass their images through FAWK model everytime they post them on a public forum.



Student 3: Now, we will discuss the paper weaknesses. As Ayoshi mentioned the box paper that is an important problem and proposes a unique solution, as we have discussed the strengths of the paper as reviewers, we must also look at its weaknesses. Some of the

weaknesses we found were one, this model can be used by malicious agencies to hide their identity.

Of course, this is a problem that occurs with various privacy solutions but this paper fails to discuss this in detail. What do some malicious agencies like criminals come into play and we need to use the visual recognition model to identify. The paper conveniently leaves discussing this off, this trade-off between the user privacy and the authorized use of the same to future work.

Second, we find that this model is not as useful if a user's pictures are already online in abundance. They say that the user then has to create Sybil accounts and upload more such photos that are cloaked. Now, this is very cumbersome for a user to do.

(Refer Slide Time: 25:56)



The next weakness that we found is that the UI of the model is not very well explained, we believe that the paper could have included more flow diagrams and better illustrations to explain how the model really works. Finally, we find that the software is not easy to use, we do appreciate the authors making the code open source and putting the model up to the public but the instructions are not very detailed. And so only tech savvy people can use this model for now. If a regular user wants to cloak their images, they may find it very difficult to do so.

(Refer Slide Time: 26:29)



NPTEL

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDRABAD

Overall Reviews

Reviewer #1: Strong Accept

Reviewer #2: Strong Accept

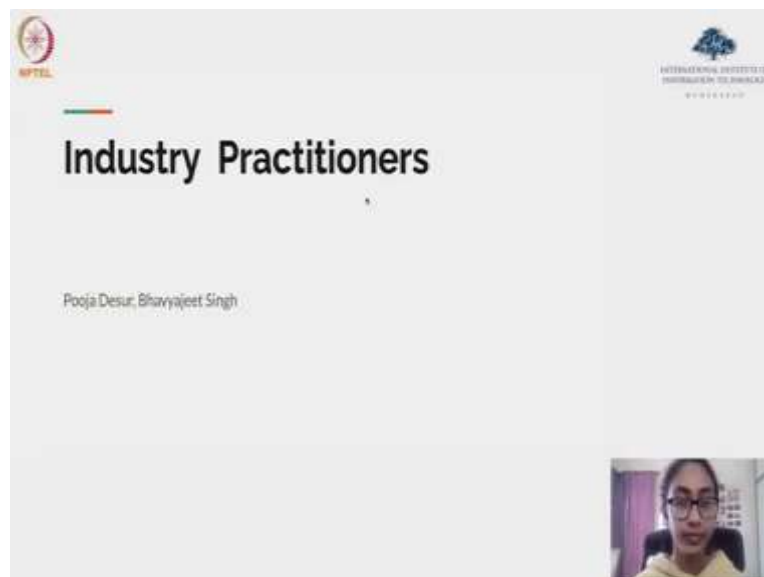
1 2 3 4



Overall, we find that the paper despite having some weaknesses it has much stronger reasons for acceptance. Thus, both reviewer one and reviewer two believe that this paper deserves a strong accept and a scalar strong reject we project weak accept and strong accept, where strong accept is the highest score the paper can receive and that is evenly it should hold up with the publication. Thank you that is all we have from the peer reviewers.

Professor: Thanks Ayoshi and Shraddha for reviewing the paper, I was just wondering as you were speaking whether, if the authors of the paper actually listen to this, they would actually look at us and say that our students actually reviewing our paper and giving reviews thanks again. Now, let us move on to the next role industry practitioners we have Bhavijit and Pooja.

(Refer Slide Time: 27:20)




NPTEL

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDRABAD

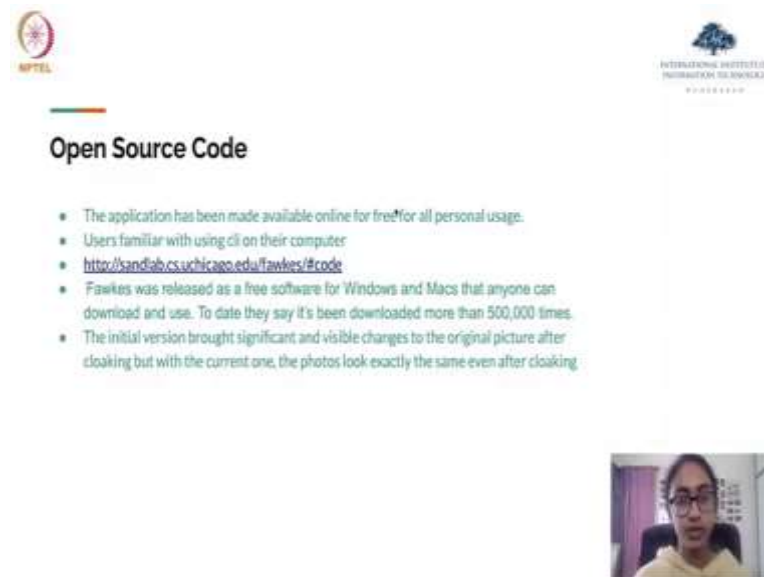
Industry Practitioners

Pooja Desur, Bhavyajeet Singh



Student 4: So, hi everyone, I am Pooja and for this pm my role is that of an industry practitioner. So, an industry practitioner goes through the paper and tries to see if it has some real-world implementations and if currently any companies or industries are using this methodology or if possible if this could be used in the future in a company or industry-based implementation.

(Refer Slide Time: 27:44)



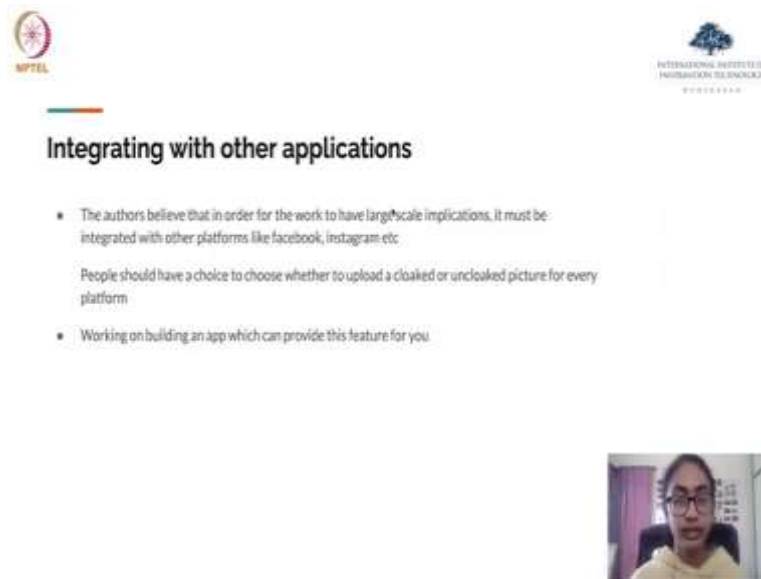
The slide features the MITEL logo on the left and the International Institute of Information Technology logo on the right. The main heading is "Open Source Code". Below it, a list of five bullet points provides details about the software's availability and usage. A small video inset in the bottom right corner shows a person speaking.

- The application has been made available online for free for all personal usage.
- Users familiar with using cli on their computer
- <http://sandlab.cs.uchicago.edu/fawkes/#code>
- Fawkes was released as a free software for Windows and Macs that anyone can download and use. To date they say it's been downloaded more than 500,000 times.
- The initial version brought significant and visible changes to the original picture after cloaking but with the current one, the photos look exactly the same even after cloaking.

So, as has been mentioned before the code for this Fawkes model is open source and is available on their website. However, it is only meant for users who are familiar with using a command line interface on their computer. Currently it is been downloaded more than 500,000 times so we can say that it has been, has seen a widespread usage. And they have released this free software for both windows and mac operating systems.


So, there were two versions that were released, the initial version brought significant and visible changes to the original picture. So, after there were complaints and feedback given to the authors of the paper, they updated the initial version and released the second iteration of the model, where the photos look the same even after cloaking.

(Refer Slide Time: 28:33)



The slide features the NPTEL logo on the left and the International Institute of Information Technology logo on the right. The title "Integrating with other applications" is centered. Below the title, there are three bullet points:

- The authors believe that in order for the work to have large scale implications, it must be integrated with other platforms like facebook, instagram etc
- People should have a choice to choose whether to upload a cloaked or uncloaked picture for every platform
- Working on building an app which can provide this feature for you



So, the authors believe that in order for this their work to have large scale impact and in order for this to blow up in a global scale it has to be integrated with platforms, social media platforms such as Facebook and Instagram where photo sharing is a predominant part of using the platforms.

So, people it would be in the future it would be a nice option if users automatically had a choice if they would want to upload cloaked photos and if the social media platform themselves provided this option rather than the user having to cloak their photos beforehand and then uploading onto these platforms.

Currently the authors of the paper are working on building an app which can provide this feature for you which is much more user friendly than having to download the software and figuring out how to use it on your own.

(Refer Slide Time: 29:24)



Clearview.AI Feedback

- Company that scraped over 3 billion photos of faces online
- Founder Hoan Ton-That: "a system like Fawkes would not only fail against a gargantuan facial recognition database such as Clearview's own, but in fact make their recognition algorithms stronger"
- could use images cloaked by Fawkes to improve its ability to make sense of altered images
- "There are billions of unmodified photos on the internet, all on different domain names," Mr. Ton-That said. "In practice, it's almost certainly too late to perfect a technology like Fawkes and deploy it at scale."
- Author feedback: "What the Clearview CEO suggested is akin to adversarial training, which does not work against a poisoning attack. Training his model on cloaked images will corrupt the model, because his model will not know which photos are cloaked for any single user, much less the hundreds of millions they are targeting."



So, clearview.ai which was the company that has scraped over 3 billion photos of faces online and which was talked about during the paper summary discussion. So, when asked the founder of this company when they were asked for a statement against a model like Fawkes which could fool their facial recognition model.

He stated that he believes Fawkes would fail against his massive facial recognition database and also stated that they believe that the recognition algorithm would actually be made stronger because of a model called Fawkes. However, when the author was asked for a comment against what Mr. Ton-That said he had stated that this the Fawkes model still holds.

Since, what the Clearview CEO suggested has to do with adversary, adversarial training which does not work against a poisoning attack which is what the cloaking model is based on. So, he still stands by the Fawkes model and states that it would still continue to fool those facial recognition models.

(Refer Slide Time: 30:22)



The slide features the MITEL logo on the top left and the International Institute of Information Technology logo on the top right. The main title is "Real World Performance". Below the title, there is a bulleted list of companies and their associated facial recognition APIs:

- Microsoft Azure Face API
 - Uber
 - Jet.com
- Amazon Rekognition
 - NFL
 - CBS
 - Nat Geo
 - ICE
- Face++
 - Alipay

A small video inset in the bottom right corner shows a woman with glasses speaking.

So, if we look at real world performance, so actual companies that use the APIs that were talked about before that this Fawkes model would end up fooling and would provide protection against, some companies such as Uber and Jet.com use the Microsoft Azure Face API facial recognition system.

Organizations such as the NFL, CBS, National Geographic and ICE which is the immigration and customs organization in the US also use this Amazon Recognition API. And also, Alipay which is a major form of payment service in China uses the Face Plus Plus API. So, all these companies the Fawkes model would provide a protection against, that is all for the industry practitioners. Thank you.

(Refer Slide Time: 31:07)



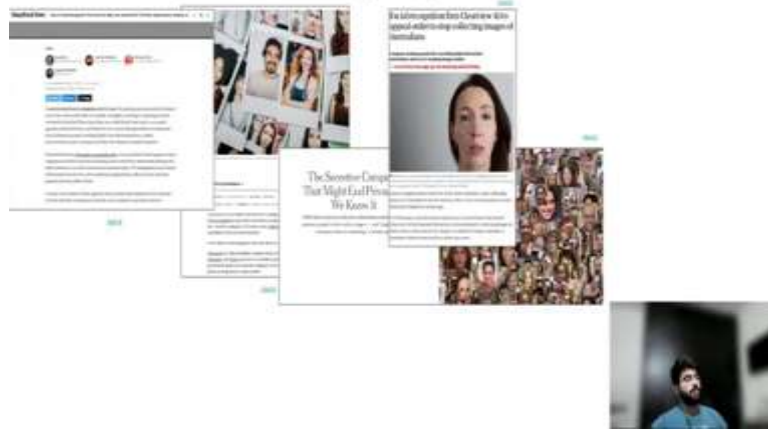
The slide features the MITEL logo on the top left and the International Institute of Information Technology logo on the top right. The main title is "Fawkes: Protecting Privacy against Unauthorized Deep Learning Models". Below the title, the authors' names are listed: "Shaan Shrivastava, Emily Wenger, Jinyan Zhang, Huijing Li, Hailan Zhang, Bai Y. Zhao". At the bottom of the slide, the text "Social Impact Assessor" is displayed. A small video inset in the bottom right corner shows a man with a beard speaking.



The social problem being addressed/ solved



Unauthorized and unaccountable use of facial recognition systems.



Student 5: Yeah, hi everyone I am the social impact assessor for the paper Fawkes. I will be discussing the positive and the negative social impacts of the paper. Yeah, so the main problem, the social problem that is being addressed by the paper is the unauthorized and the unaccountable use of facial recognition system.

The continuing example that we have discussed in this paper is of clearview.ai which was able to scrape multiple images from online social media platforms and was able to collect and create a big facial recognition system which not only is unauthorized but unaccountable to the various government sites and can access your information and recognize your photos. So, to attack this problem they had created a new tool called Fawkes which was able to mitigate this problem.

(Refer Slide Time: 31:58)



Positive Impacts



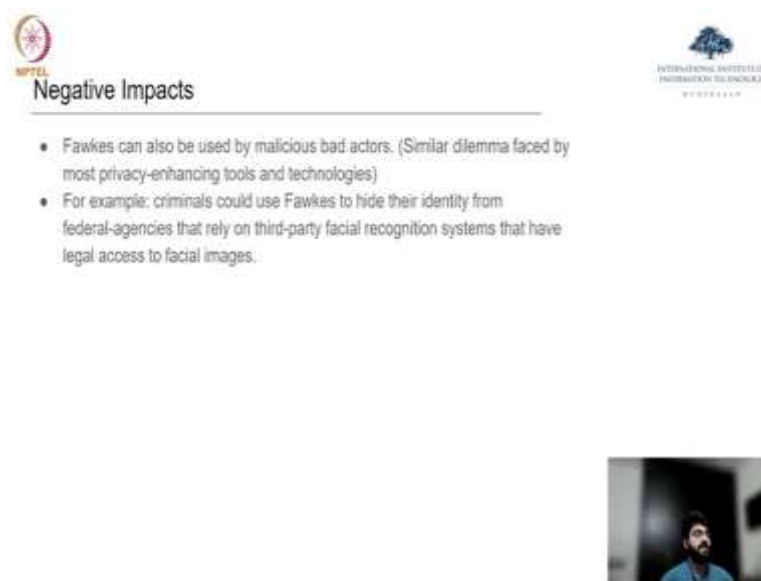
- Help people safeguard their photos on online social media platforms from being used to train a facial recognition system without their consent.
- Create a safe environment for people on online social media platforms.



But the positive impacts of the paper could be seen as that the Fawkes algorithm helps safeguard the photos on online social media platform from being used to train facial recognition systems without their consent. So, this creates a safe environment for people to post on these online social media platforms without being worried about their photos being used for unauthorized purposes.

So, this is an example of where people or researchers had created an algorithm to distinguish between discriminated society, discriminated segments of the society and were asked to take off their paper from the internet.


(Refer Slide Time: 32:38)



The slide features a header with the MITEL logo on the left and the International Institute of Information Technology logo on the right. The title 'Negative Impacts' is centered below the logos. Two bullet points are listed below the title, and a small video thumbnail is positioned at the bottom right of the slide content area.

Negative Impacts

- Fawkes can also be used by malicious bad actors. (Similar dilemma faced by most privacy-enhancing tools and technologies)
- For example: criminals could use Fawkes to hide their identity from federal-agencies that rely on third-party facial recognition systems that have legal access to facial images.



So, one of the negative impacts that was also discussed in the paper but was left as the future work was that similar to many privacy enhancing tools and technology Fawkes can also be used by malicious bad actors. Criminals can use Fawkes to hide their identity from federal agencies and would be unaccountable to the law.

(Refer Slide Time: 32:58)

The slide is titled "Other social problems that arise with facial recognition systems". It features three main components:



- Bar Chart:** Titled "Accuracy of Face Recognition Technology". The y-axis is "Accuracy (%)". The x-axis is "Face Recognition Technology". The chart compares accuracy for four groups: Darker female, Darker male, Lighter female, and Lighter male. The accuracy values are: Darker female (28.2%), Darker male (28.7%), Lighter female (34.4%), and Lighter male (33.4%).
- MIT News Article:** Titled "Study finds gender and skin-type bias in commercial artificial-intelligence systems". The sub-headline reads: "Examination of facial-analysis software shows error rate of 6.8 percent for light-skinned men, 34.7 percent for dark-skinned women." Below the article is a thumbnail for a Netflix documentary titled "CODED BIAS".
- Video Thumbnail:** A small, dark video thumbnail showing a person's face.



So, one of the impacts that I think the authors missed for the papers was that even though the facial recognition systems might be authorized but they still have a lot of bias in them. So, as we can see multiple studies have shown that bias in terms of the gender or the colour of the skin creeps into this model because of the data sets that are being used to train these models.


So, most of the times these models perform for people with darker skin tone or which would also perform worse for a female gender. There are multiple such studies and one such recent studies was also in the form of a Netflix documentary called “Coded Bias.”

So, other innovative ideas that have also been introduced by many researchers have been as in terms of clothing accessories to help people stay away from these facial recognition softwares, where people can add these adversarial patch and they can then be protected from these facial recognition software.



(Refer Slide Time: 33:45)


 **Other Innovative Ideas (Clothing/ Accessories based) [1/2]** 


 



(Refer Slide Time: 34:07)

 **Other Innovative Ideas (Clothing/ Accessories based) [2/2]** 





So, another example of this innovative ideas was of LED glasses which people can use to mitigate the problem of facial recognition software. So, in all the paper addresses a very important social problem and it is not just the unauthorized use but also the bias that creeps into these facial recognition software that the paper helps to stop. Thank you.

(Refer Slide Time: 34:35)



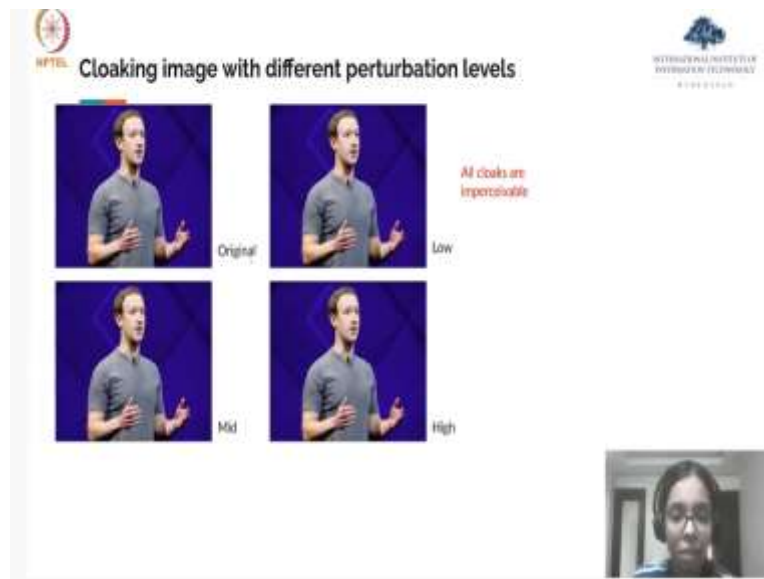
Professor: Yeah, thanks, thanks Mudit. Now, we will have a hacker role, who will talk about the paper. Akshala?

(Refer Slide Time: 34:44)



Student 6: Yeah, so as a hacker I had to look at the implementation of the paper. The authors have made the source code publicly available it is available at this GitHub link. So, when you are running the code there are different modes which are available, so there is like a minimum medium and high mode, so the higher the mode is the more perturbation will be added in the cloaked image and it would provide a stronger protections.

(Refer Slide Time: 35:04)



So, I tried out the code with different images and different perturbation levels so like this first one is the original image, this is with a low perturbation, this is with a medium perturbation, and this is high perturbation. So, as you can see all of these images look identical only, so the cloaks are imperceivable.

So, this is verifying what has been mentioned in the paper that their implementation would yield images which are imperceivable. So, like someone would not be able to distinguish between the original and cloaked image with the naked eye.

(Refer Slide Time: 35:39)



So, what basically is happening in this code is that the model is look it is picking up pictures from public data set, so it picks up like K groups of images in which each group has photos of

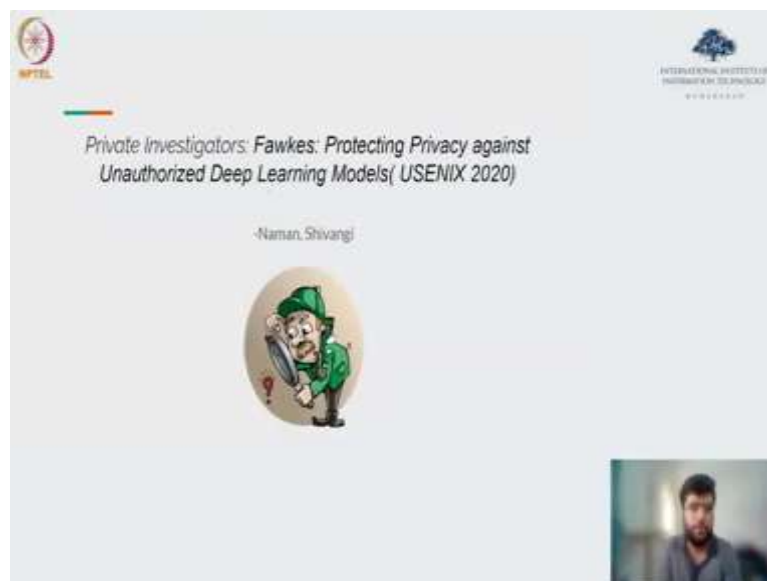
different people. Then using a feature extractor, the centroid of feature space of these images is being calculated so this is happening for the K groups of images of different people.

And for the cloaked image also the centroid of feature space is being calculated then afterwards the L2 distance between the cloaked images and all of these groups is calculated. So, we select the group which has the highest L2 distance. So, highest L2 distance would imply that this group is most dissimilar to the images that have to be cloaked.

Now, from the selected group, we randomly pick up one image and using this image we calculate the structural dissimilarity index and using that we get the cloak. Now, this structure dissimilarity index has an input parameter which controls the perturbation. So, different modes as we saw previously these low, medium, high modes this can be given as an input parameter while calculating the cloak and you can get different images as that we had seen in the previous slide. So, this was the implementation and algorithm of the paper.

Professor: Thanks, thanks, Akshala. Now, let us go to a private investigator, Naman?

(Refer Slide Time: 37:08)



Student 7: So, hi I am Naman and I was the private investigator for the project for this paper and what my job was to look into the history of the authors, what kind of research that they have been doing and what may have led to this project. So, we will talk about the first authors over there, there were two first authors for the paper who were also the project leads and then the advisors for the papers.

(Refer Slide Time: 37:36)



The slide features a header with the MPTEL logo and the name 'Shawn Shan' in blue. To the right is a portrait of Shawn Shan. Further right is the logo of the International Institute of Information Technology, Hyderabad. The main content is a bulleted list of his academic and professional background, including his B.Tech and Ph.D. from the University of Chicago, his role at Facebook, and his research at the SAND Lab. It also lists his previous research work, such as 'Oh, the Places You've Been!' and 'Tracking Transparency'.

- B.Tech CS University of Chicago (2016-2020), PhD (University of Chicago, 2020-), advisors: [Ben Y. Zhao](#), [Heather Zheng](#)
- Worked as a Senior Software Engineer at Facebook (2019-2020)
- Researcher at the SAND Lab and the SEPER Lab at U. Chicago
- PhD at SAND Security, Algorithms, Networking and Data's Lab. Research spans topics in security, machine learning, networked systems, HCI, data mining and modeling.
- Previous research that could have influenced this work:
 - Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Influencing (2018)
 - Tracking Transparency: A privacy preserving browsing extension that visualizes examples of long-term information third party trackers could have inferred from user's browsing. Gave users a better perception of the extent of tracking.

So, the first project lead was Shawn Shan, who is a Ph.D. student at the University of Chicago. So, this research this paper was research was done at the University of Chicago as the SAND Lab which stands for Security Algorithms Networking and Data basically and they work in research topics like security, machine learning, networking systems, HCI, data mining and modelling.

So, Shawn Shan had did his B. Tech in Computer Science from the University of Chicago and then he started his Ph.D. at the SAND Lab under Ben Y. Zhao and Heather Zheng who are the co-advisers for this project also. So, before this project I think talking about the previous projects for Shawn Shan.

In 2018, he published a paper on user reactions to longitudinal transparency about third-party web tracking which basically tried to give users a better understanding of what all data can third party apps track from a user's browsing history. So, the aim of this project was to again give, tell the users that okay all of this what is the extent to which third party apps can track your data which was I think a good, it was a very crucial research that was done to make that from the users end to understand how privatized their data on the internet.

(Refer Slide Time: 39:04)

The slide features the NPTEL logo on the left and the IIT Madras logo on the right. The main content is a list of research achievements:

- Neural Course: Identifying and Mitigating Backdoor Attacks in Neural Networks (2019)
 - Built a generalizable detection and mitigation system for DNN backdoor attacks by identifying backdoors and reconstruct possible triggers. Identified multiple mitigation techniques via input filters, nearest pruning, and unlearning. (Published at IEEE Symposium on Security and Privacy)
- Lead to his PhD at the SAND Lab. Currently exploring limitations, vulnerabilities, and privacy implications of neural networks. Works include protecting neural networks from backdoor and adversarial attacks, using imperceptible perturbation to protect user privacy.

Below the text are two profile pictures: one of Shashan Shan with a cityscape background, and another of Shashan Shan with a white background. To the right of the second profile picture is the text: "Shashan Shan, University of Chicago, shashan@cs.uchicago.edu - @shashan, Machine Learning, Security, Privacy". At the bottom right of the slide is a small video feed showing a man speaking.

So, apart from this in 2019 he also worked on a very similar project in identifying and mitigating back door attacks in neural networks. So, basically in this, this was published in the IEEE symposium on security and privacy which where they were trying to basically prevent backdoor attacks on neural networks which and what backdoor attacks basically mean is that if the input image contains a certain possible trigger, it can bring down the whole neural network and it can basically misidentify the image into anyone they want to.

So, to mitigate these backdoor attacks it would filter neural planning and unlearning with some of the techniques that they proposed in this paper. So, I think these research that he did as an undergraduate and then in 2020 he started his Ph.D. at the SAND Lab where he has done his previous research.

(Refer Slide Time: 40:02)



The slide features the NPTEL logo and the name 'Emily Wenger' in blue text. To the right is a portrait of Emily Wenger, a woman with long blonde hair wearing an orange top. Further right is the logo of the International Institute of Information Technology, Hyderabad. The main content is a bulleted list of her academic and professional background:

- BS Mathematics and Physics from Wheaton University (2007-2010), PhD University of Chicago, 2018- present, advisors: [Ben Y. Zhao](#), [Hendrik Zeng](#)
- Worked as a research assistant at Wheaton University, as a mathematician at the US Dept. of Defense.
- Intern at Facebook AI Research (Sept 2021- present)
- PhD at SANDS Security, Algorithms, Networking and Data Lab. Her research explores the practical limitations, privacy violations, and security threats of deep neural networks.
- Previous research that could have influenced this work:
 - Privacy Resistant Watermarks for Deep Neural Networks (with Shawn Shan)
 - Tackling piracy attacks against false claims of ownership by embedding privacy-resistant watermarks. Proposed null-embedding, a new way to build privacy-resistant watermarks into DNNs that can only take place at a model's initial training



So, talking about the second author now which was Emily Wenger, she hired her a Bachelor's of Science in Mathematics and Physics from Wheaton University and then she started her Ph.D. in the University of Chicago from 2018 so she has worked there is an internet Facebook AI research, she is working as an intern and her research basically explores practical limitations, privacy violations and security threats of deep neural networks.

So, a lot of work that Emily Wenger has done in this domain has been with Shawn Shan which is the other lead for the project, so one relevant project that they worked on was piracy resistant watermarks for deep neural networks which was basically a strategy to tackle piracy attacks against false claims of ownership on deep neural networks.

So, basically what they proposed was a null embedding system which is basically a new method to build a watermark into a deep neural network which can only be put in the time of training. Therefore, nobody can so to say steal it or basically claim a false ownership on a network.

(Refer Slide Time: 41:10)

Common Works Between Emily Wenger and Shawn Shan

- Privacy resistant watermarks for deep neural networks (2019)
- Fawkes: Protecting privacy against unauthorized deep learning models (USENIX 2020)
- Blacklight: Defending black-box adversarial attacks on deep neural networks (2020)
- Gotta Catch'Em All: Using Honeypots to Catch Adversarial Attacks on Neural Networks (2020)

Emily Wenger
University of Chicago
Verified email at uchicago.edu · @emwenger
Machine Learning Security Privacy

So, some of the projects that as I said before like Emily Wenger and Shawn Shan have worked together in the past, they have worked in this domain for a while and after Fawkes also they have published two papers, one was called Blacklight which is again defending black box adversarial attacks on deep neural networks and Gotta Catch'Em All again using Honeypots to Catch Attacks on Neural Networks.

So, basically Emily Wenger and Shawn Shan because of the Ph.Ds. also lie in a very similar domain against exploring security and privacy of deep neural networks, I think have been, yeah so this is their research.

(Refer Slide Time: 41:49)

Advisors: Ben Y. Zhao & Heather Zheng

Heather Zheng / Co-Director SAND Lab at (Univ. Chicago)

- PhD from University of Maryland in Electrical and Computer Engineering
- Over 20,000 citations in mobile computing, wireless networks, security and privacy
- Worked at Bell Labs (4 years), Microsoft Research Asia (1 year)
- Professor at University of California, Santa Barbara
- MIT Technology Review (2005) for her work on cognitive radios
- Areas of Research:
 - Mobile/Internet and its implications in privacy/security
 - Security and Privacy of Deep Learning Systems

Now, let us talk about the advisors of the lab. So, first was Heather Zheng who is the co-director at the SAND Lab at University of Chicago, her Ph.D. was from University of Maryland in Electrical and Computer Engineering. And she has over 20,000 citations in mobile computing, wireless network, security and privacy.



So, she was also a part of the MIT technology review in 2005 for her work on cognitive radios and before being the director at University of Chicago she was a professor at the University of California Santa Barbara.

(Refer Slide Time: 42:20)

Advisors: Ben Y. Zhao & Heather Zheng

Ben Y. Zhao Co-Director SAND Lab at Uni. Chicago

- BS in Computer Science from Yale (97), MS and PhD from University of Berkeley (2000 and 04 resp.) in Computer Science advised by John Kulkarni and Anthony Joseph
- Previously assistant professor at University of California, Santa Barbara
- Over 32000 citations in security, adversarial machine learning and HCI
- ACM Distinguished Scientist, Recipient of NSF Career Award
- Areas of Research:
 - P2P networks, online social networks, SDN, open spectrum systems, graph mining and modeling, user behavior analysis, to adversarial machine learning.
 - Since 2016, Mostly working on security and privacy problems in machine learning and mobile systems.



And interestingly Ben Y. Zhao was also a professor there at university of California Santa Barbara. His BS was in Computer Science was from Yale 97 then his Master's in Ph.D. from Berkeley 2000 and 2004. So, he is also a very renowned scientist in the field of adversarial machine learning and human computer interaction security, he has over 32,000 citations and he is been awarded the ACM distinguished scientist award and the NSF career award.

So, his area of researches include P2P networks, online social networks, user behaviour analysis to name some. And since 2016, he has been working in this domain of security and privacy problems of machine learning and mobile systems. Yeah, so that is all from myself.

Professor: Yeah, thanks Naman for the background on the authors. So, that is how paper reading is done so I hope that gave students an idea of how to read a paper, what kind of different rules you can actually look at the paper and what to take from the paper.

(Refer Slide Time: 43:30)



What we covered?



Conducting (User, Lab, and Online) Studies
Reading research papers



So, what we have covered for this week is user studies, then how to read research papers, what are the roles that you could play. I think we can get into the details of generally what the paper is all that, let me see how the appetite for the class is, how much students are interested in this.

Depending on your interest I could actually add more content, we can add that content even during the semester that you are going to take this class. And we can go through some more papers, understand what are the details in writing the paper, how the paper is built all that. So, that is the content for week 7. Thank you for watching.