Review Assignment : DAI

1. Write the review of the paper "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey"(https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8294186). The review should be written in your own words. You can discuss with your colleagues to understand the concepts but the writing should be your own.

2. This paper was published in 2018. Since then, several attacks, detection and mitigation algorithms have been proposed. Include a summary of 6 additional papers (2 attacks, 2 detection and 2 mitigation), which are not covered in this paper. These papers should be published in top venues only.

3. Draw a table according to the categorization proposed in the 2018 paper (question 1) and fit the new papers you have studied in this categorization, with justification.

Task : 1

Paper Review: Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8294186

➔Write the review of the paper "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey" (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8294186). The review should be written in your own words. You can discuss with your colleagues to understand the concepts but the writing should be your own.

Solution:

Let's first understand with a short summary:

With the rise of deep neural networks we have come to a great remarkable achievements using different SOTA architectures of DNN. Even there are examples that have shown Super Human Capabilities. Just like defeating time limited humans in recognizing cat vs dog. However, the Computer Vision techniques are prone to attacks. Attacks are anything that shows undesired result or classification. A simple example in Image processing or CV studies, a small disturbances in the image which is called as adversary, can change small perceptible change however, changes the prediction label of the image. So, when you have a ML model along with ground truth, prediction truth without attacked model and prediction of the model when it is attacked, you can realize what an attacking algorithm can change in your Machine Learning. The paper specifically focuses on designing of adversarial attacking algorithms, detection of the attack, defense of such attacks. Detection and Mitigation. Paper also covers real world examples. Additionally, the paper opens a broad categories for research interests. Not only attack, the model is changed such a way that, it starts predicting an incorrect class with high confidence in laboratory settings. In the Ian Goodfellows paper we have seen that when dog's image is perturbed, it starts classifying it as hummingbird with 98% confidence than the dog with less than 40%. Ajmal Mian, in the paper mentions that this divides the community into two categories, the one who tries to make harder attacking algortihms and other ones which try to defend such powerful attacking algorithms. They also defined the terminologies which are standard in the study of Dependable AI course. Terms included in the paper are: adversarial example, adversarial perturbation, adversarial training, adversary, black box attacks, detector, fooling ratio, one-shot methods, quasi imperceptible perturbation, targeted attacks, threat model, transferability, white box attacks. In the paper, they have defined 11 types of attacks along with misc. attacks. Lets define each of them below:

1. Box Constrained L-BFGS: The methodology tells that with even a small perturbation, can lead to misclassification. This attack calculates small amount of perturbation which could lead to misclassification. It calculates loss of the classifier such that we get

maximum loss with the smallest perturbation. It was also observed that small amount of noise could lead to fool different neural networks.

2. FGSM: The main idea of FGSM is that the robustness or strongness of the NN against adversarial example can be improbved with learning adversaries in the adversarial examples. It calculates the noise to be added by solving a mathematical equation which results gradients. So if we have a picture of panda we try to add a factor of noise such that it misclassify the panda to gibbon. It finds the gradient of the loss function wrt the clean image given, and takes element wise sign() function for the gradients, this tells the direction we want to go into. So when we apply higher epsilons it passes multiple decision boundaries of multiple classes from the dataset. This decides the final classification. Now this is a kind of threshold such that up to this threshold the model cannot add more noise. They also proposed linearity hypothesis, that is the model is presumed to be highly non linear but the attacks exploits the nature of linearity in the models in higher dimensional space. In other words, there is a kind of tradeoff which goes like, You work on deep models such that to make it work linear for computational benefits makes them prone to get adversaries. With this, the authors proposed that robustness can be increased by learning these adversaries back to the deep neural network.
3. Basic and Least likely Iterative Methods: It adds noise by taking a larger step in first go. Then it minimizes the noise such that the minimum noise added creates a large loss to the classification. The BIM proceed with clean image and runs iteratively. The extension of BIM is ILCM (Iterative Least Likely Method) Moreover ILCM has proved better results in misclassification of popular modern Deep Architectures.
4. JSMA: It is a white-box targeted attack. Jacobian Saliency Map Attack is an algorithm to generate adversarial attacks. The core to JSMA is saliency map. This works on the underlying principle of which region of the image is actually important that drives the classification of the example. Traverses pixel to pixel and the generated saliency map tells that what are the features that take part in deciding the classification. In other words, it look to the features which are important and tells that yes these are the pixels (for example in MNIST those can be vertical and horizontal features), which are important and those need to be attacked. So once we have this saliency map ready, we try to push our clean image towards a neighborhood image such that we want the new image to be an adversarial example. As it is written in the paper itself, “The algorithm chooses pixels that are most effective to fool the network and alters it.”
5. One Pixel Attack: When you add noise to just one pixel and it changes the complete distribution of the classification criteria. That is by adding only one pixel you achieve misclassification of the clean image. In fact with the experiment they found with only one pixel they got higher confidence of the misclassified example. This is Non targeted Black Box attack.
6. CW Attack: The underlying idea of CW Attack algorithms is that, suppose you have an input image X, then you envision for another X’ which is the neighborhood of the image in the input space. Or in other words you search for an img X’ which is the target image which you want the attacking algorithm to output. This is the target label. This works like as follows. First you minimize the distance between x and x’ to find the x’. Once you find x’ you get the target label. which is you decided in the start to get the output of the attacking algorithm. Since flavor of CW attack algorithm used here is the Targeted Attack and Whitebox. by minimizing we mean that in the input space we want our x and x’ to be near so that both becomes perceptibly similar. Function g tells how close we are to the target label. Keeps on updating in the attacking algorithm.

7. Deep Fool: It is advanced version of FGSM. The author proposed that Deep Fool adds less number of perturbations such that image is misclassified.
8. Universal Adversarial Perturbations: Moosavi proposed, there exist universal perturbations which when added to any image, it gets misclassified. These UAPs are imperceptible to naked eye. Or quasi imperceptible.
9. Upset and Angri: UPSET gets a target image and learns noise while ANGRI takes both target label and original image. Thus UPSET minimizes the norm of misclassification loss and ANGRI finds the direction where to proceed to find the gradient direction to find multiple decision boundaries.
10. Houdini: This is something different from all of the attacks. It tries to generate adversarial examples instead of adding noise to query image for ML models.
11. Adversarial Transformation Attacks (ATNs): It is proposed with the principle that learn networks such that it generates adversarial examples. With this an attacker has multiple images and one can select the best one among multiple network generated examples. Formally, this is learning to generate adversarial examples.

Rest of the paper describes miscellaneous attacks. Also author describes the attacks on Machine Learning models itself. They also describes about defensive mechanisms to mitigate these attacks.

Review:

The paper starts with describing the attacks under laboratory settings but are as powerful that can be seen in real world as well. Attacks mentioned in the paper comprises of techniques whether they are Whitebox or Blackbox, targeted or non-targeted and what's their learning process whether it is iterative of single shot. All these attacks essentially have one underlying concept which is minimizing the noise to be added such that it maximizes the loss. All work beyond this paper was mainly focused on the same concept. Different Authors came with flavor of targeted and non-targeted, in targeted attack, they tried to add noise such that the network classifies it as the target class. Some attacks uses L0 norm by adding noise to complete image and then shrinking the noise where as other tried to add L2 norms by adding minimum noise to the clean image to make it misclassified. However they mentioned about one attack called Houdini which works completely different from other attacking techniques. It is based on generating adversarial examples and attacking works on alternative loss function, though the working remains same as it also target backpropagating gradients. Not only classification attacks, they also mentioned how attacks are evolving by modifying the architectures also and leveraging the properties of different deep neural networks. These tasks are beyond classification, in the sense that they try to address attack on different classes in the image by object detection.As it is rightly said by K. Reddy (IISc), "gradient is a double edged knife", he further mentioned in the conference that "attacking an ML system is easier than defending it." This requires a rigorous training of adversarial examples and robust training procedure. Moreover, Adversarial training may approximately doubles the training time and might be specific to a particular attack. With the understanding of new architectures like Houdini and other generative adversarial techniques, this can be reduced, where you might not be dealing with each and every adversarial example generated and back throwing it to the model for training. So the thought process goes like, Why do we need to train adversarial images, why not learn the underlying principle of the attack and generate less number of examples on contrary to the adversarial example of each image in training process. People also work on data augmentation to make model robust beforehand. Apart from this, the paper opened (in 2017) a new study for the researchers to formulate new losses and defensive mechanisms to counterpart these attacks. Community found this paper-survey interesting as a handy tutorial which comprises all past attacks and other mechanisms at one place. More interestingly, there exist real world attacks which the paper mentions are road sign attacks and face attacks. There are face masks or some objects which can fool the system, moreover, a special class of masks which can results targeted fooling. As paper describes wearing lipstick and not, Classifying a female as male etc. Road sign attacks are more dangerous when come to autonomous driving. A simple attack to stop signal board might lead to fool the autonomous car driving system. Community has come with several attacks to autonomous cars as well, example, "emergency brake attack", "AV freezing attack" which leads the car to remain stopped even after red sign turns green. Researches wants to address the vulnerabilities by finding newer attacks such that it triggers companies like Tesla inc. to spend more on R&D of Trustworthy AI. It becomes more adversarial that Tesla car was attacked to change lanes, it was shown that a small adversarial sticker when placed in multiple places, tricked the car to change lanes. The paper also mentioned about fundamental existence of adversarial examples that is how do they exist in nature. The methods presented in the survey (Ajmal Mian et al.) which are limits, space boundary tilt, uncertainty, robustness correlation, linearity hypothesis and existence of universal perturbations. According to me, the existence of these adversarial examples is no more than responding to bad gradients and learn incorrect features. As, when noise is added, the model have to learn multiple dimensions (spaces) which somewhere cause it to understand wrong features and thus result bad predictions. I agree with the author where they have mentioned about universal noises. This is clearly true because we know there exist universal

stickers also which when added to a classification engine, changes the class label. Even this can be seen with Google Lens and other object detection tools. As you might know, the human adversarial example in interstate 60 movie, black hearts and red spades! "Experience had conditioned you into thinking that all hearts are red and all spades are black, more because their shapes are similar! It is because based on past experience, mind is not open to idea that they could be different" So this again calls to the paper of Ian Goodfellow, about fooling time limited humans. The lesson was, "the people who don't play cards often tend to tell that there are black hearts and red spades". This is what we want from our machine learning model also. And once you know that there exist black hearts and red spades (that is adversaries) you will be able to perceive them! Thus in my point of view, once we get to know why do the adversaries exist with a strong notion may be specific to a dataset or universal, we also envision about the vulnerabilities which can be corrected. This categorizes a new research scheme called Detection and Mitigation! There is one paper "Robustness May Be at Odds with Accuracy (2019)" which tells that robustness and mitigation is a tradeoff. So when you try to do for example adversarial training in penalty you reduce the overall accuracy. There are question posed by the authors that, "would it always be preferable to have a robust model instead of a standard?" and "what are the costs you incur in adversarial training?". The paper also claims that, a classic ML model would try to overfit with large dataset, however, against this statement they have showed that features learned by classic ML (Standard Classifier) technique and features learned by large dataset adversarial training (Robust Classifier) can be fundamentally different. Thus in my opinion, defensive mechanisms just not do the plain adversarial training, but also a model should understand the features which are actually creating disturbances. There are techniques mentioned in the paper are based on 3 notions, modified inputs, modified networks and network addon. One of the method in modifying the network is Gradient Regularization, we all know gradient helps in learning models along with attacking models. Thus gradient has to be handled correctly. So one thing which is mentioned in the paper is minimizing the loss of the model over worst case adversarial examples instead of data.

Important conclusions with my review:

1. There is a large fight between ML community being completely denying for adversarial attacks:
   I completely deny with this, I agree with that fact that adversarial attacks do exist and are severe in some cases as we have seen many examples throughout. Major one being autonomous car driving. So the "Threat is Real"
2. Even you come up with a *best* deep neural architecture, it would be able to generate misclassifications when attacked. So one thing you can do is learn and adapt the changes so that your model does not perform same mistakes again. Because it is a general phenomenon.
3. It is observed that most adversarial examples shows misclassification among different Deep Architectures, in that case it is better that to have one standard set of adversaries that should be learned along with the machine learning model training. This shall improve accuracy and reduces classification loss.
4. Also, I agree with the fact that the vulnerability lies inside the linearity of deep neural architectures, or in other words, adversaries exploit the linear nature of the DNN (where output is just a linear combination of the inputs), it is because most of the neural architectures are linear in kernel spaces or in other words, not using non linearity in the activation side of the neural networks. But this is one reason. There are modifications and community coming with highly nonlinear activations. But why Goodfellow posed that models are linear? As all the models in the deep neural world uses non linear activations.
5. The paper gives an additional advantage such that it is a large study of categorization of Adversarial Attack generation, Its Detection, and Defensive mechanisms in single place. The paper bridges the gap between different attacking algorithms and help readers to understand the nature & properties of these algorithms in a single table.
6. I think we shall start building awareness to make people understand to think about these adversaries they face in daily lives, but accept it as a machine learning model fault. This includes examples, from "Adversarial Examples from Physical world"
7. The study of vulnerability is still confined to the computer vision community, but a broader thought process is required where at industry level, all companies leveraging ML models whether to perform Object recognition, speech synthesis or Analytics, they should care for vulnerabilities. More specifically, when it comes to Natural language processing, adversarial attacks can be severe. In computer vision the perturbed image still looks somewhat perceptible to human eyes, however noisy audio or adversarial audio are not easy to interpret.(Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey 2019)
8. I really liked the paper because it index major attacks since it was written in 2017, I know at least these were the major attacks, detection and mitigation techniques. This opens a channel to research about each of the survey topic they have covered.
9. The paper uses less number of experiment data though it is not expected in survey but, still some accuracy charts, speedups, classification reports could have helped better understanding of Attack generation, Detection algorithms mentioned in the paper.
10. Table 1. is the key takeout from 1st half of the paper.
11. Finally I would like to say the survey paper is well written and putted on for a early researcher to start understanding the adversarial attacks.

Questions:

1. It is still unclear to me about linearity hypothesis as most of the models uses nonlinear activations. In 2017 also people have used non linearity in their architectures. This could be a small reason for vulnerability but not major reason!
2. Gradients are very important. They help to learn at the same time they also are helpful for adversarial examples. When detection, why not come with different gradient learning approach such that we drop those gradients which try to take our image at a distant point in the hypothesis space. In other words, why not come up with an approach such that we detect in the classification process that when our classification logits start giving unusual behaviors or sudden drastic changes in it. However I know there are adversarial attacks which uses fast flipping attributes (fig 5 from paper) which will fail the former approach which I mentioned!

Some Corrections:

1. There are some typos:
    a. Most of the places they have written “maleware” instead of “malware” (pg1)
    b. In section 3.A.7 (Deepfool) (pg 6) “polyhydron” ➔ “polyhedron”
    c. Section Cyberspace attacks (Pg12) “maleware” , “lrearning”
    d. Section Miscellaneous Approaches (Pg 16) “prposed”
2. In table 1 description of the “Strength” should be provided.

Task : 2

➔This paper was published in 2018. Since then, several attacks, detection and mitigation algorithms have been proposed. Include a summary of 6 additional papers (2 attacks, 2 detection and 2 mitigation), which are not covered in this paper. These papers should be published in top venues only.

2 Attacks

1. https://arxiv.org/pdf/2101.02562.pdf (**DeepPoison: Feature Transfer Based Stealthy Poisoning Attack for DNNs**)
   Deep poison: Poison attacks are the ones which makes the vanilla model ill and destroy it. Or it makes a model to misclassify the examples. Data poisoning is a type of adversarial attack where a perturbation or adversarial patch is added in training, which allows models to misclassify at test time. It is studied that one single poison image can change the behavior of the classifier. Deep poison methodology comprises of an iteratively carefully adding small amount of poison (examples) into the training data such that the classification results start misclassifying the test data. This can be achieved sometimes even with a single image. In general for small datasets, there exist a poison image for each corresponding training example, However most of the poison images are transformations of one image. Deep poison works with three principle components, Feature Extractor (FE), Generator (G) and Discriminator (D). G generates the poison images to add noise. Iteratively, D finds the similarity between the generated image and the clean image. This way FE ensures that there are similar features in the perturbed image and poison image.
2. https://arxiv.org/pdf/2101.00989.pdf (**Fooling Object Detectors: Adversarial Attacks by Half-Neighbor Masks HNM**)
   This method is used to fool many classifiers and object detection techniques. The method creates a region inside the clean image. This is also called as mask. The attacking algorithms picks some region inside the masked region and put it to the learning algorithms. Picking small regions from a masked area of the image and then add noise to that. This way they achieved that perturbed images look very clear in perception. This also preserves the important features of the clean image and rest of the pixel distribution. The novelty of this algorithms lies inside the mask which does not hinder any pixel arrangement in rest of the image. This was rarely identified in previous work. The method identifies the adversarial patches in the image automatically under l^0 constraint. This method can be assumed working iteratively by covering a large area of the image and shrinking the mask such that minimum noise to be added and maximum classification loss is gained. PGDs are targeted gradient flow. Here we use PGD to minimize the noise function wrt subject to class C. We direct to negative gradients and project to feasible set.

2 Detection

1. https://arxiv.org/pdf/2101.02899.pdf
   (**Adversarial Attack Attribution: Discovering Attributable Signals in Adversarial ML Attacks**)
   This detection mechanism works on finding the underlying architecture of the attacking algorithm used. It finds feature maps, signals that exposes attacking algorithm and hyperparameters used in the attack. Author devised a tactic and technique procedure by which they have identified the class of attacking algorithm. Once found the class and nature of the attacking algorithm, you can understand the big picture. The procedure learns about the CNN architecture, Attacking Algorithm, and Parameters such that it generates attributes and classified accordingly. Author uses signal and noise interchangeably in the paper. They not only identify or detect the noise with previous methods, but also try to understand the signals from different attacks. It is like, creating a categorization that if I get this kind of signal distribution into the adversarial example, I will classify the misclassification algorithm to a particular category according to the signal distribution. The author also claims that each type of attacking algorithm generates a distinguishable kind of signal vector. By this way of learning new kind of noise or signal distribution into the adversarial example, they categorize the noises and place new incoming signals into one of them. They have shown such signal vectors in figure 2. So they come up with some genuine questions asked into the paper,
   a. How to learn such signals or how to store such signal vectors so that in future query images we are ready to classify the attacks or more importantly the categorization of the attacks.
   b. How do you distinguish with models with such noise vectors.
   c. Author envision about coming up with hyperparameters used during the attacking technique.

   This paper puts a step ahead and claims all these statements by assuming that we have already identified the adversarial example in the previous steps. Now putting all together, the fundamental approach that authors claim is, they want to put all adversarial examples to a dataset and end up creating a dataset of all adversarial examples. Now that we have created a dataset, we run an Unsupervised ML model which learns the underlying signal patterns and categorizes the signal vectors. This way not only we will detect the noise, but also ready for a whole class of similar attacks using the property of inheritance and universal nature.

2. http://bit.ly/daipaper
   (Detecting Adversarial Examples through Image Transformation)
   what is adversarial example. An adversarial example is any disturbed image which is a product of a clean image perturbed by some attacking algorithm. With the rise of community building attacking algorithms, there is another group of researchers who defend against these attacking techniques. To study the detection mechanism, let us first understand the CW Attack algorithm. Because when the paper was published it was thought that CW Attacks bypasses all the detection techniques back in 2017. Thus the need of detection mechanism to combat such algorithms was required. The paper is written completely to defend from CW algorithm. What is

CW Attack algorithm? The underlying idea of CW Attack algorithms is that, suppose you have an input image X, then you envision for another X' which is the neighborhood of the image in the input space. Or in other words you search for an image X' which is the target image which you want the attacking algorithm to output. This is the target label. This works like as follows. First you minimize the distance between x and x' to find the x'. Once you find x' you get the target label. which is you decided in the start to get the output of the attacking algorithm. Since flavor of CW attack algorithm used here is the Targeted Attack and Whitebox. by minimizing we mean that in the input space we want our x and x' to be near so that both becomes perceptibly similar. Function g tells how close we are to the target label. Keeps on updating in the attacking algorithm. The novel idea in this algorithm comes as a flavor of the same algorithm, it uses L0 norm. where you perturb complete image in one go then you shrink the modified pixels such that you end up with the minimum pixels perturbed. As you can see in the image. there are very less number of pixels perturbed in the region of the image such that it classified dog as hummingbird even with 98% confidence. What is the optimization problem? So the equation $\min||\delta||2 + c \cdot f(x+\delta)$ where delta is the distance, x is the example or clean image, function f returns whether the image is perturbed or not. A Boolean return type with yes or no kind of return values which tells that if the image is perturbed or not. So recollecting and putting in single picture it says that, I want to minimize the distance such that the perturbed image lies close to the clean image in input space. Also, we want to do this iteratively (remember we have 2 type of algorithms, single shot and iterative in Ajmal mian paper.) so this is the iterative kind of algorithm which iterates and finds the minimum perturbation which makes our function f to return the success value. Additionally, c is the hyperparameter which controls the tradeoff distance and the function. this hyperparameter is auto tuned by the algorithm. In addition to that, going into the formulation of the function f we come to know that Nicholas Carlini removes the dependency on the function g which used to tell about how close we are to the target label and merged in single function f. Again recollecting this formulation, we come to know that x+delta = x' which is perturbed image, when passed to the function f, calculates the distances between clean and perturbed image. the max is term is sitting there which says that, I want to maximize the perturbation such that our clean image x changes it label from lx to i. This is the illustration for image transformations based adversarial example detection. This is what author has proposed that, Classification results of Clean images are immune to transformations. Classification results of Adversarial images are prone to transformations. The methodology used by author in describing how to detection adversaries in an image example using their approach. In addition, they also tells about how do they build their idea analog from human behavior.

Methodology is as follows, First they add more data to the corpus they already have. They added the rotations, transformations and shifts of same image to the dataset. so if they had m examples earlier after applying k transformations, they end up with K*M examples. Now passing the transformed data to the learning algorithm, to get the predictions and classification process, they could essential detect the adversarial examples. in more detailed terms, suppose you have x1 as clean image. now you pass this example to get corresponding adversarial example. x1'. Now that we have x1', we transform this x1' and get k transformations. When those k transformations passed to the model, it predicted different labels for most of the examples. This complete thing utilizes the property that, "You can classify the image as adversarial or normal image as Classification results of Adversarial images are prone to

transformations" or "minor transformation of an image may result in a significant change of the classification results. The basic idea is to apply several transformation operations on an image. "

The goal is to build a detector that is able to distinguish adversarial examples from the normal ones. Inherently, a detector is also a classifier, which is trained on a set of normal and adversarial examples. Detectors learn to capture the difference caused by the added perturbation. The detector will tell you whether the model is attacked or not. Because if the image is clean image it must get correct classification where as if the image is perturbed image it shows wrong classifications, however we also ask question to the classifier if the transformed image gives us the correct label or not. While performing experiments, the authors also got to know about the change in arrangement of pixels when the image is transformed. In the subsequent papers following same ideology proved that the small gray part added in the transformed images so as to resize it back to 32X32, they found that this rotations lead to change in the arrangement of pixels and decision boundaries of popular Deep Neural Architectures.

2 Mitigation

1. https://arxiv.org/pdf/1801.08926.pdf
   (**Deflecting Adversarial Attacks with Pixel Deflection**)
   Pixel Detection and Wavelet denoising are core to this methodology. The authors were motivated by important observation : CNNs can learn inherent adversaries in the image but are prone to added noise. With this observation, author experimented with adding more noise along with the existential noise and the attacked noise so as to diminish the attacked noise. This resulted in lowering the power of adversarial noise, due to the property that CNNs can learn inherent adversaries or noises. This is also intuitive because, this is what we do in adversarial training. Collecting the adversarial examples and throwing to the learner so that when same adversary come in it tells that it is not Y'(Incorrect label), but Y (correct label).This is done by taking a pixel randomly and iteratively from the image, then replacing it with another chosen pixel which is present close to the previous pixel in the space. They also proposed Targeted Pixel Deflection. When looked to the feature maps or the activation maps, experiments showed that attacking algorithms tend to attack complete image with some distribution of the noise. While clean images have high pixel intensity with foreground objects and less pixel intensity for background objects. Using this property, author found, pixels with less intensity which were actually the background objects also got some perturbations or noise or signals. In the experimentation part, firstly a feature map or thermal map is devised. This will tell you the objects with sharp edges in the image. In this maps when compares with clean image and adversarial image, it was found the less important parts in the clean image feature maps tend to show some noises in the perturbed image. Now that we have found, the denoising process works further. This try to soften and diminish the effect of such noises. Due to its nature of deciding foreground and background principles, it is more relevant to use it in real world images with high resolutions. This resulted in good accuracy on many real world cases.

2. https://arxiv.org/pdf/2012.01701.pdf
   (**FENCEBOX: A Platform for Defeating Adversarial Examples with Data Augmentation Techniques**)
   This paper focuses on defending with adversarial attacks by preprocessing techniques. Author claims that, past works uses attacks and defensive mechanisms to mitigate the adversaries. The novelty of their work lies in that they focus only on preprocessing based solutions. Specifically, the paper deals with 15 different kinds of data augmentation functions. The methodology works as firstly, they collect results of each attacks on clean image along with success rate with different configurations. Author also tells that using this mechanism as a cloud service and call it as preprocessing as a service. Which means you come up with the data, preprocess using described algorithm and then pass it to the classification tasks. The preprocessing includes reducing the effect of carefully crafted adversarial attacks. Preprocessing the image includes three approaches: First, image distortion. Secondly, image compression. Thirdly, inject artificial noise. The important takeout of the paper is to add multi-dimensional transformations, such that attacks are minimized. Using the tool they have devised called FenceBox, you augment data. Then you concatenate it to the defending

mechanisms. *g(·) = g1 ◦ g2. Function g1 is a highly randomized function that can mitigate the BPDA attack. This can be selected from the image distortion category. Function g2 is a non-differentiable function to thwart the EOT attack* The main idea is to reduce the effect of gradients learning for adversaries. As I already mentioned "gradient is double edged knife". The three basic functions used in the paper are:

a. Image Distortion: This remaps the pixels from their original position to another. This makes gradients to learn different scheme when attacks are applied. There is a great difference between a fencebox preprocessed image and clean image.
b. Image compression
c. Image noise injection: Adds intentional noise to the images which are learned with the classifier CNN model but not learned by the attacking algorithm. Randomizes pixels by adding gaussian noise.

Task : 3

Draw a table according to the categorization proposed in the 2018 paper (question 1) and fit the new papers you have studied in this categorization, with justification.

DEEP POISON	Whitebox	Non Targeted	Image Specific	L2 for model L0 for pois	Iterative
HNM-PGD	Whitebox	Targeted	Image Specific	L0	Iterative
ATTR. SIGNALS	Blackbox	Targeted	Image Specific	-	Iterative
IMAGE TRANS.	Whitebox	Targeted	Universal	-	Single Shot
PIXEL DEFLECTION	Whitebox	Non Targeted	Image specific	-	Iterative
FENCBOX	Whitebox	Non targeted	Image specific	-	Iterative