Artificial intelligence is the theory and development of computer systems capable of performing activities ordinarily carried out by humans, such as visual perception, speech recognition, decision-making, and language translation. In this paper, the authors explore one of the numerous vulnerability’s artificial intelligence faces, specifically focusing on how intelligence can be susceptible to adversarial attacks.

Adversarial assaults involve harmful attacks on data that appear normal to the human eye but result in misclassification in a machine-learning pipeline. These attacks are often disguised as specially tailored ‘noise,’ which can lead to misclassification. The paper illustrates this by placing a patch in front of the AI where a banana was already located, causing the AI to misclassify in the machine-learning pipeline. This action led the AI to identify a real toaster from the patch, resulting in an incorrect classification. This illustration demonstrates how these changes are designed to exploit the weaknesses of these models, eventually causing the model to misinterpret predictions or produce biased results. These attacks provide attackers insight into the model’s internal workings and potential vulnerabilities.

In conclusion, the paper reveals that AI has numerous vulnerabilities that, if studied and analyzed by attackers, can provide them with knowledge on how to exploit AI for their advantage.

**New Summary**

Artificial intelligence is the theory and development of computer systems capable of performing activities ordinarily carried out by humans, such as visual perception, speech recognition, decision-making, and language translation. In this paper, the authors explore one of the numerous vulnerability artificial intelligence faces, specifically focusing on how deep learning models are susceptible to carefully chosen inputs that cause the network to change output without a visible change to a human, called adversarial patch attacks.

The paper explores this vulnerability by illustrating a construct of an attack that does not change one object into another. Instead, this attack generates an image-dependent patch that is noticeable in the network. This patch can be placed anywhere in the testing area of the classifier (deep learning model) and causes the network to change the output to how it should be. In the experiment, the attack was completed by replacing a part of the image with a patch. The masked patch allowed it to take any shape and then trained over a variety of images, applying a random translation, scaling, and rotation on the patch in each image, optimizing using gradient descent. This patch exploits this feature by producing output inputs much more obvious than objects in the frame. This causes the patch to be classified as a toaster thus not affecting other portions of the image.

This experiment was tested out by comparing the efficacy of two Whitebox attacks, a Blackbox attack, and a control patch. The attack was evaluated by averaging the win rate across all five models. The white box single model attack does the same but only trains

and evaluates on a single model. The Blackbox attack jointly trains a single patch across four of the ImageNet models, and then evaluates the black box attack on a fifth model, which we did not access during training. The control is a picture of a toaster. This test was performed by using a physical patch generated by the white-box ensemble method. When a photo of a tabletop with a banana and a notebook (top photograph) is passed through the model, it reports it was a ’banana’ with 97% confidence (top plot). If we physically place a sticker targeted to the class "toaster" on the table (bottom photograph), the photograph is classified as a toaster with 99% confidence

In conclusion, deep learning models (AI) are susceptible to carefully chosen inputs that cause the network to change output without a visible change to a human (adversarial patches). It shows that created and targeted attacks such as adversarial attacks can fool classifiers regardless of the scale or location of the patch and does not require knowledge of the other items in the scene that it is attacking.

New Summary Part 3

Artificial intelligence is the theory and development of computer systems capable of performing activities ordinarily carried out by humans, such as visual perception, speech recognition, decision-making, and language translation. In this paper called “Adversarial Patch” by Tom B. Brown et al. explores how carefully chosen computerized inputs such as 3D image-like patches such as toaster that are placed into scene (a tabletop with a banana and notebook) is modified into a scene with the aim to deceive or manipulate a machine learning model. This type of attack is called adversial patch attack.

This paper explores this vulnerability by illustrating a construct of an adversarial attack where a carefully chosen input such as the 3D-image-like patch of toaster is silently placed into the scene. When this image-like patch is placed into the scene of a tabletop with a banana and notebook, the model’s decision-making process was manipulated to cause them to misclassify the scene and produce inaccurate predictions. This image-like patch is placed at certain points in the scene of a tabletop with a banana and notebook. This image-like patch is placed within in the scene of the classifier, it will cause the classifier to misclassify the scene which lead to output a targeted class The aim of this patch is to cause the classifier to give inaccurate information of what it should be it targeted class. It is designed to manipulate the classifier into misclassification, is added to a scene where a banana is already present, the classifier must consider both the original features of the scene (including the banana) and the perturbations introduced by the adversarial patch. The conflicting information between the actual banana in the scene and the adversarial patch designed to resemble a toaster can lead to misclassification. The classifier might struggle to reconcile these conflicting cues, potentially resulting in a misclassification where the output is labeled as a toaster even though there is no actual toaster present.

This experiment was tested out by comparing the efficacy of two Whitebox attacks, a Blackbox attack, and a control 3D-patch. Black box attacks are characterized by a lack of detailed knowledge about the targeted system, while white box attacks involve a higher level of understanding of the system's internals. The attack was evaluated by averaging the win rate across all five models. The white box single model attack does the same but only trains and evaluates on a single model. The Blackbox attack jointly trains a single patch across four of the ImageNet models, and then evaluates the black box attack on a fifth model, which we did not access during training. The control is a picture of a toaster. This test was performed by using a physical patch generated by the white-box ensemble method. When a photo of a tabletop with a banana and a notebook is passed through, the model classified it as banana in the scene. The model classified it with a high confidence of a banana in the scene. But when the image-like patch was placed into the scene, the model was manipulated. This causes the model to misclassify to give an inaccurate prediction. The model gives a prediction of the output of a toaster. This prediction was inaccurate due to the classifier struggle to reconcile these conflicting cues, potentially resulting in a misclassification where the scene is labeled as a toaster even though there is no actual toaster present. The only prediction that should be given is a banana. When the crafted image-like patch is introduced into the scene, the white-box attack successfully manipulates the decision-making process of the model, causing it to misclassify the scene as a toaster instead of correctly identifying the banana. While in the black-box attack still succeeds in manipulating the decision-making process, leading to the misclassification of the scene. This showed why the adversarial patch is more effective than a picture of a real toaster, even when the patch is significantly smaller than the original picture of a real toaster.

In conclusion, deep learning models (AI) are susceptible to computerized chosen inputs such as image-like patches. When these inputs are placed within the scene of a model, it will cause the model to misclassify which eventually leads to inaccurate predictions.

ChatGpt Link: <https://chat.openai.com/share/f5e3a4c9-2736-4b5c-9b5f-d4a953cdee2e>