Piotr Król, 28.06.2021

Adversarial examples

Adversarial examples are inputs to machine learning models that do not change objective reality (class of input) but cause them to make false predictions. They have the power to fool the machines. Adversarial examples could be imperceptible to humans, unrecognizable to them, or have the form of adversarial patches (Shen et al., 2020).

In this project I implemented an example of adversarial attack called projected gradient descent (PGD). It is a so-called white-box attack, since the attacker has access to model gradients (Knagg, 2019). I checked the working of this attack on a simple convolutional neural network trained on the FashionMNIST[1] dataset. I checked how big the perturbations must be to fool this network. I've also checked if the model can be resistant to this attack, by providing adversary inputs (created with the same technique) during the training stage. Finally, I produced some examples of adversary-attacked images.

According to Fig.1. the value of epsilon (size of perturbations) for which accuracy of the model was dramatically affected and reached the level worse than a random-guesser was around **1,5**. Higher values of epsilon produced only slightly lower accuracy, whereas perturbations with such values could have been much more visible.

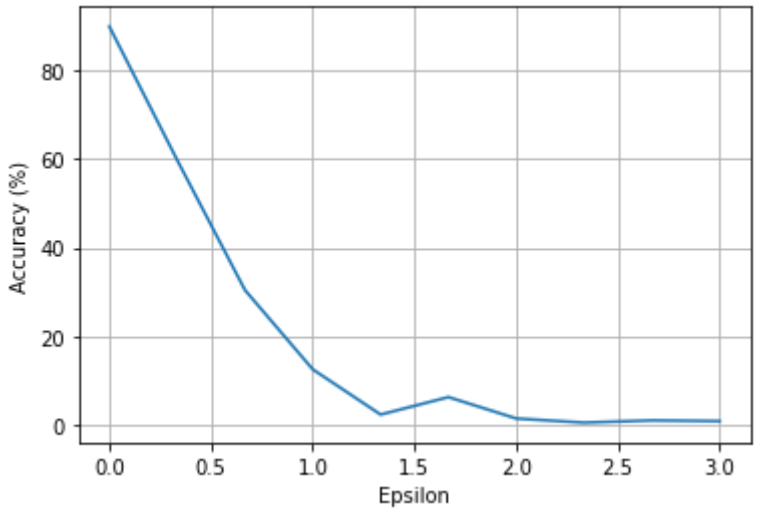

Fig.1. Accuracy of the model vs. size of perturbations (epsilon).

Thus, it was the value of epsilon (1,5) selected for adversarial attack. Accuracy of the model on clean inputs was around **90%**, whereas accuracy on the attacked images fell to almost **1%**. Clearly the attack was successful. When the network was given adversary-attacked images during the training stage its accuracy on clean inputs was a bit lower **~77%**, however its accuracy on PGD inputs was definitely much better **~ 54%**. To get an idea of what kind of perturbations were made by the attacker I produced some examples of clean and attacked images (Fig. 2., Fig. 3.).

[1] https://github.com/zalandoresearch/fashion-mnist

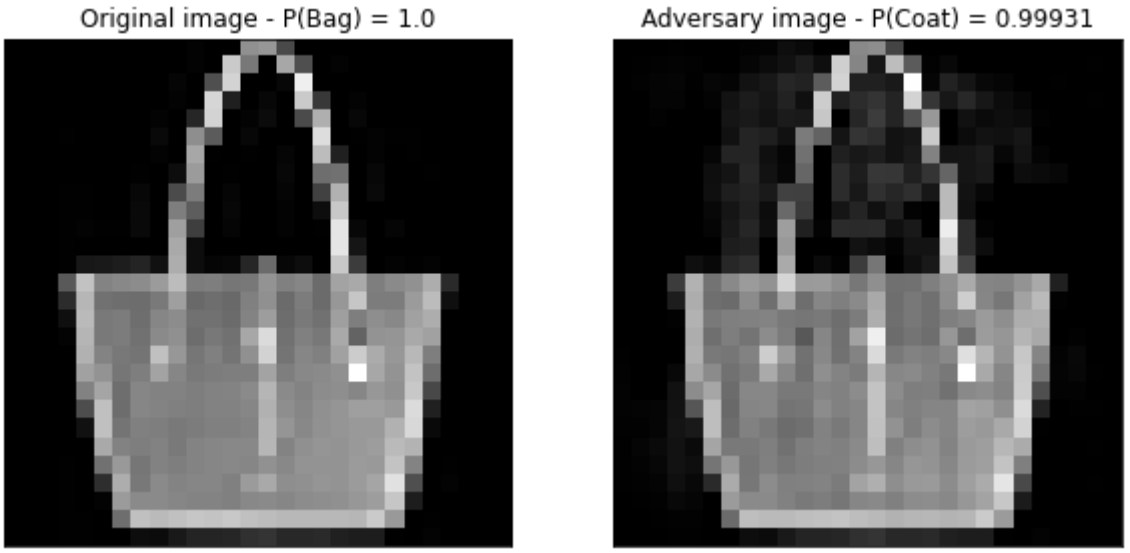

Fig.2. Example of original and attacked image. Picture on the right is not much transformed, although it's perceived by the network as coat with almost 100% certainty.

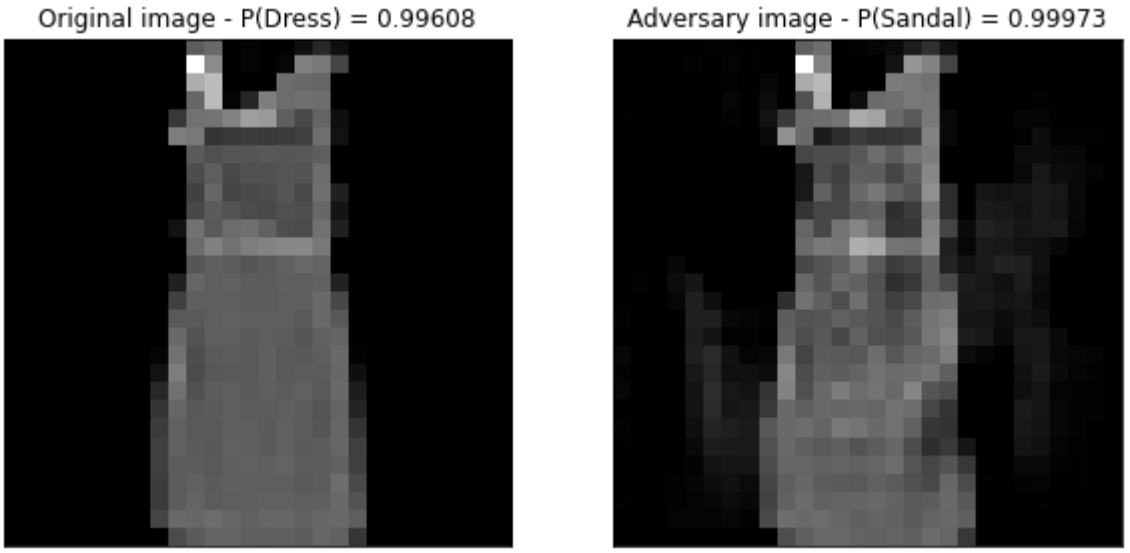

Fig.3. Example of targeted attack with the goal of making the network classify everything as sandals.

References

Knagg, O. (2019, January 6). *Know your enemy*. Medium. https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3

Shen, H., Chen, S., Wang, R., & Wang, X. (2020). Generalized Adversarial Examples: Attacks and Defenses. *ArXiv:2011.14045 [Cs]*. http://arxiv.org/abs/2011.14045