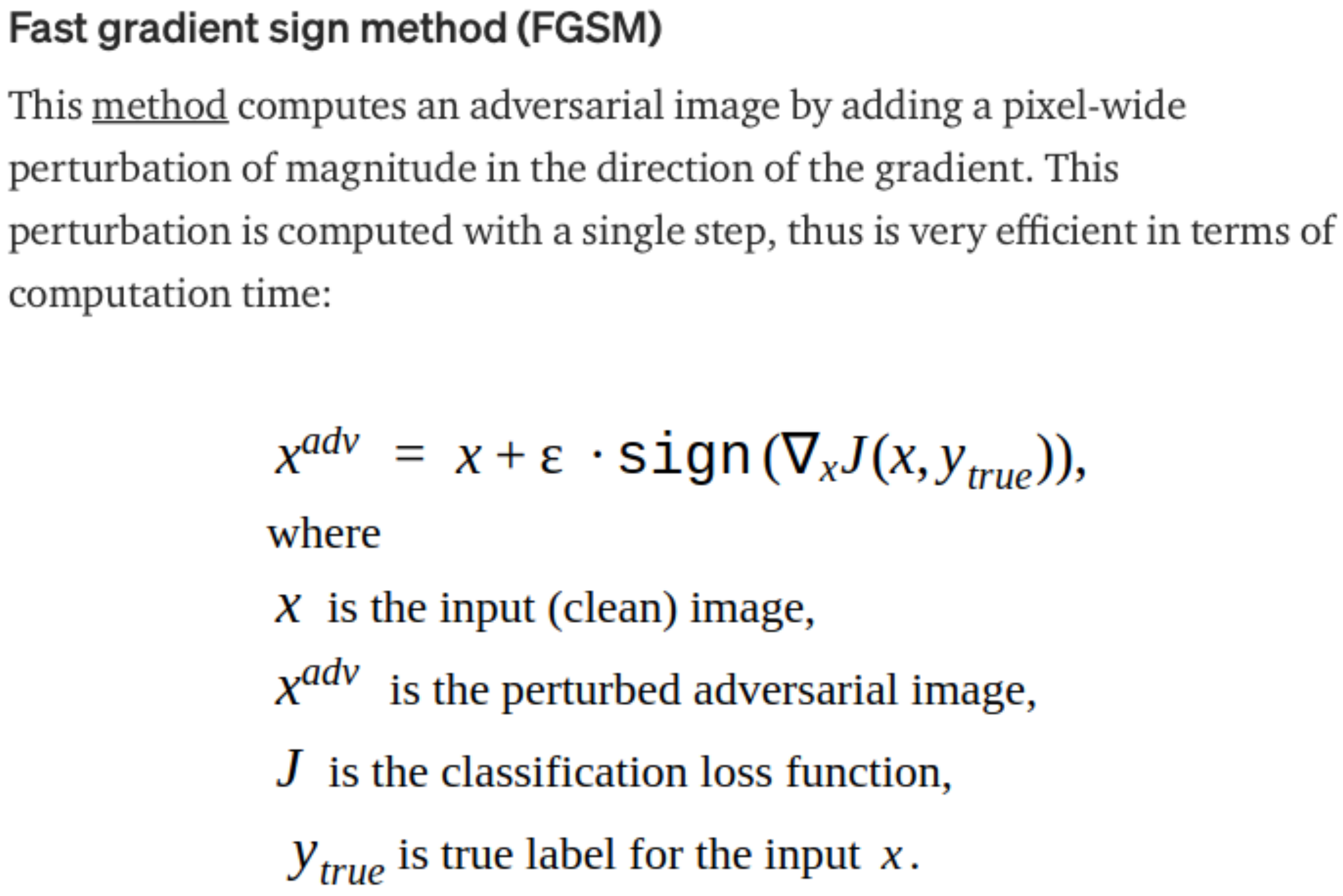
Advanced Computer Vision

Практический курс

Савельева Юлия Олеговна, 3 семестр, 30.11.2021

Adversarial Examples

Взять уже обученную на Stanford Online Products нейронную сеть на задачу классификации.
Реализовать FGSM или T-FGSM, НО не I-FGSM.



Fast gradient sign method (FGSM)

This method computes an adversarial image by adding a pixel-wide perturbation of magnitude in the direction of the gradient. This perturbation is computed with a single step, thus is very efficient in terms of computation time:

$$x^{adv} = x + \varepsilon \cdot \mathtt{sign}(\nabla_x J(x, y_{true})),$$

where

x is the input (clean) image,

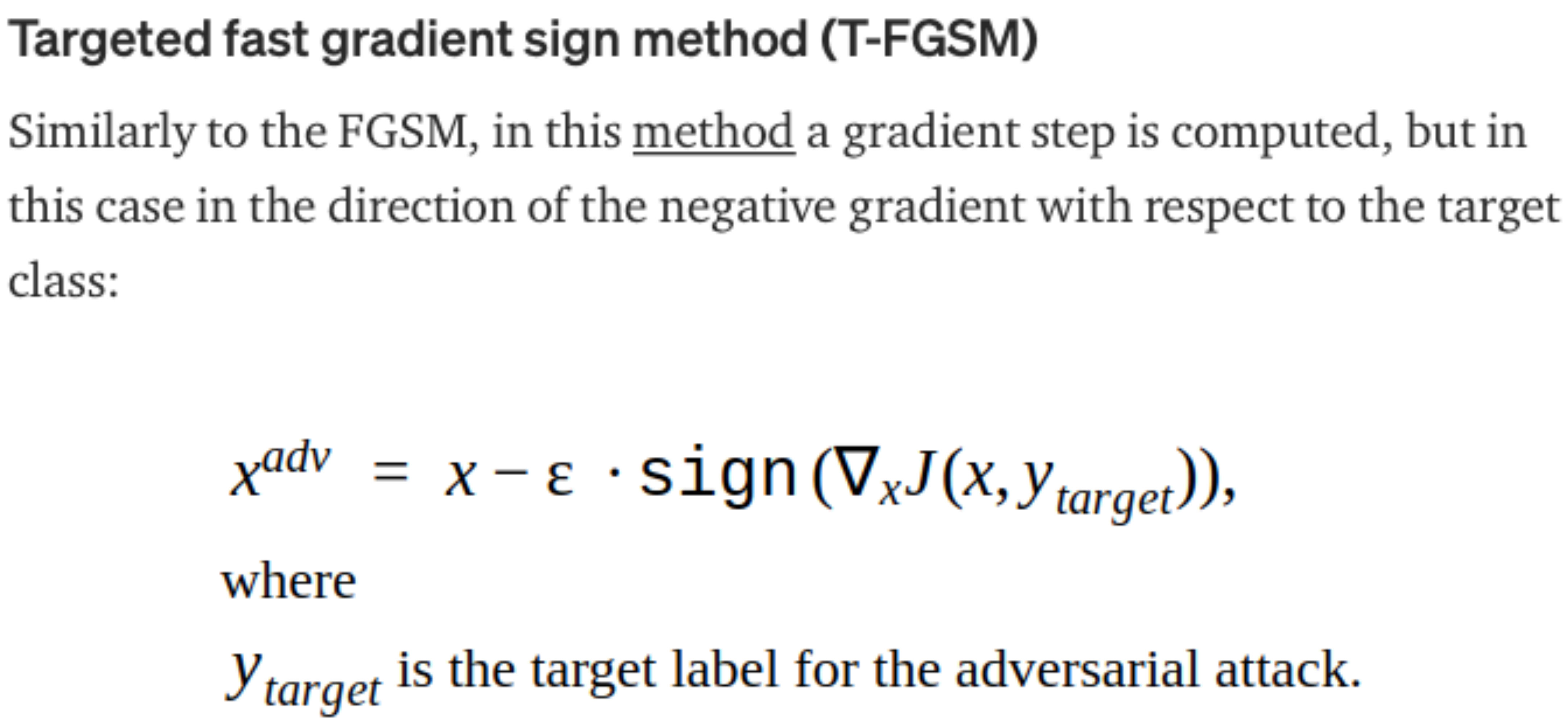
x^{adv} is the perturbed adversarial image,

J is the classification loss function,

y_{true} is true label for the input x.

https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7

Adversarial Examples



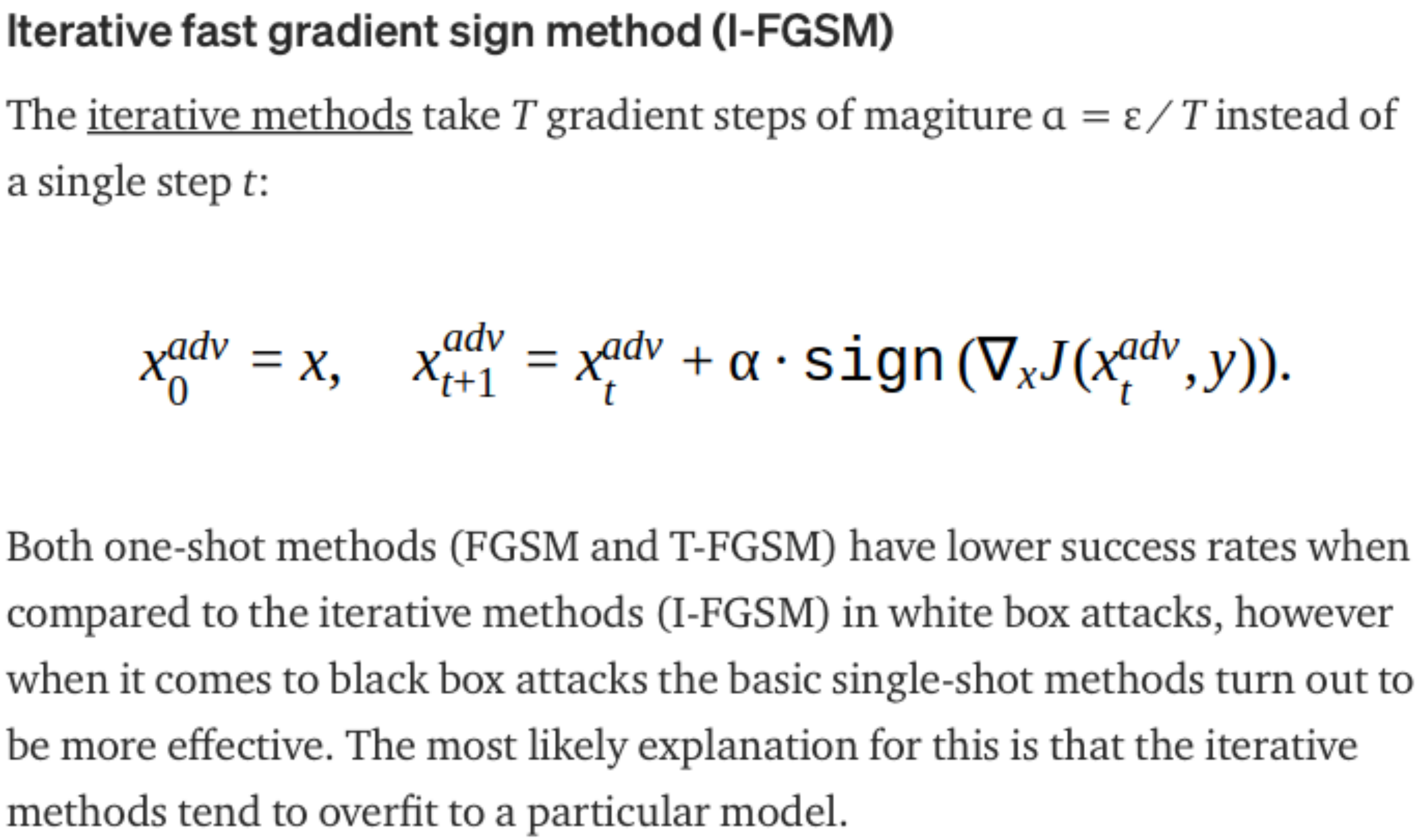
Targeted fast gradient sign method (T-FGSM)

Similarly to the FGSM, in this method a gradient step is computed, but in this case in the direction of the negative gradient with respect to the target class:

$$x^{adv} = x - \varepsilon \cdot \mathtt{sign}(\nabla_x J(x, y_{target})),$$

where

y_{target} is the target label for the adversarial attack.



Iterative fast gradient sign method (I-FGSM)

The iterative methods take T gradient steps of magiture $\alpha = \varepsilon / T$ instead of a single step t:

$$x_0^{adv} = x, \quad x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \mathtt{sign}(\nabla_x J(x_t^{adv}, y)).$$

Both one-shot methods (FGSM and T-FGSM) have lower success rates when compared to the iterative methods (I-FGSM) in white box attacks, however when it comes to black box attacks the basic single-shot methods turn out to be more effective. The most likely explanation for this is that the iterative methods tend to overfit to a particular model.

https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7

Adversarial Examples

Для картинок из списка вычислить adversarial examples. Найти такой ε, чтобы нейронная сеть справлялась с классификаций хуже всего, но при этом на картинке не появлялось видимого шума. Изменить картинки и прислать на проверку.

bicycle_final/111265348817_0.JPG

bicycle_final/111265348817_1.JPG

bicycle_final/111265348817_2.JPG

bicycle_final/111265348817_3.JPG

bicycle_final/111469262153_0.JPG

bicycle_final/111588452395_0.JPG

bicycle_final/111588452395_1.JPG

bicycle_final/111612717975_0.JPG

bicycle_final/111612717975_1.JPG

bicycle_final/111612717975_2.JPG

bicycle_final/111612717975_3.JPG

bicycle_final/111612717975_5.JPG

bicycle_final/111620439270_1.JPG

bicycle_final/111620439270_3.JPG

bicycle_final/111645993618_0.JPG

bicycle_final/111645993618_1.JPG

bicycle_final/111645993618_2.JPG

bicycle_final/111645993618_3.JPG

bicycle_final/111661598505_0.JPG

bicycle_final/111661598505_2.JPG

bicycle_final/111661598505_3.JPG

bicycle_final/111661603079_0.JPG

bicycle_final/111661603079_2.JPG

bicycle_final/111661603079_3.JPG

bicycle_final/111687027218_0.JPG

bicycle_final/111687027218_1.JPG

bicycle_final/111687027218_10.JPG

bicycle_final/111687027218_11.JPG

bicycle_final/111687027218_2.JPG

bicycle_final/111687027218_3.JPG

bicycle_final/111687027218_4.JPG

bicycle_final/111687027218_6.JPG

bicycle_final/111687027218_7.JPG

bicycle_final/111687027218_8.JPG

bicycle_final/111687027218_9.JPG

bicycle_final/111688507119_1.JPG

bicycle_final/111688507119_2.JPG

bicycle_final/111702172753_0.JPG

bicycle_final/111702172753_1.JPG

bicycle_final/111702172753_2.JPG

bicycle_final/111708813226_0.JPG

bicycle_final/111708813226_1.JPG

bicycle_final/111720342775_0.JPG

bicycle_final/111720342775_1.JPG

bicycle_final/111720342775_10.JPG

bicycle_final/111720342775_11.JPG

bicycle_final/111720342775_3.JPG

bicycle_final/111720342775_4.JPG

bicycle_final/111720342775_5.JPG

bicycle_final/111720342775_8.JPG

Дедлайн 14.12.2021 00:00