Adversarial Attacks and Adversarial Training

Ruthwik Ganesh
230702930
ec23759@qmul.ac.uk

This is a supporting document for ICV coursework 2 contains images and tables that should be referred as one progresses through the report.

Fig.1 Adversarial examples from overfitting

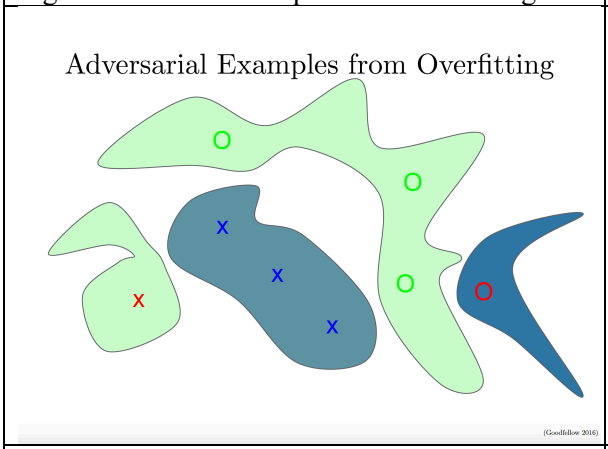

Fig. 2 Adversarial examples from underfitting.

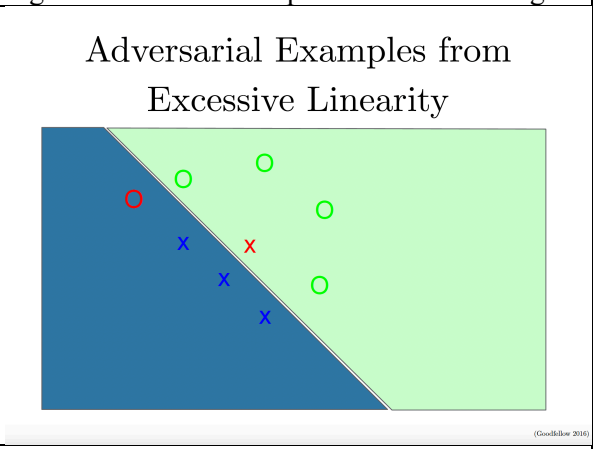

Fig. 3 Reason for linearity in DNNs

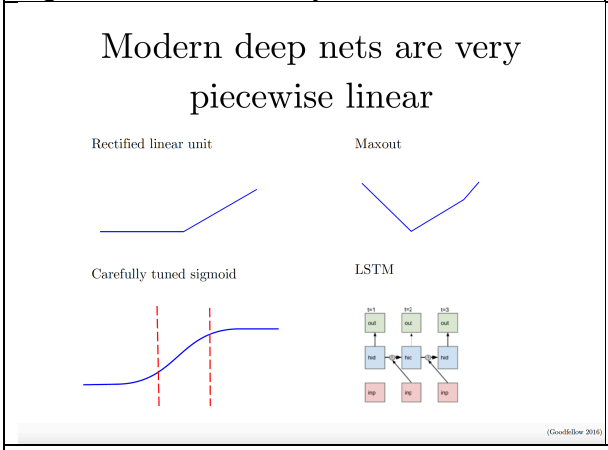

Fig. 4 Generating an adverserial image.

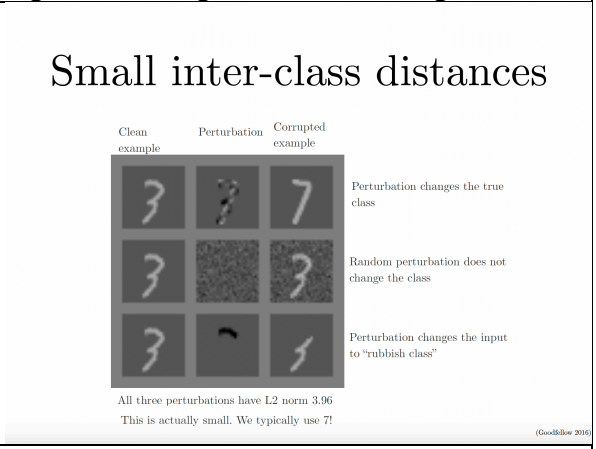

Fig. 5 Don't know/Null class label for unfamiliar images.

Table 1: Summary of the attributes of diverse attacking methods

Method	Black/White box	Targeted/Non-targeted	Specific/Universal	Perturbation norm	Learning	Strength
L-BFGS [22]	White box	Targeted	Image specific	ℓ_∞	One shot	$***$
FGSM [23]	White box	Targeted	Image specific	ℓ_∞	One shot	$***$
BIM & ILCM [35]	White box	Non targeted	Image specific	ℓ_∞	Iterative	****
JSMA [60]	White box	Targeted	Image specific	ℓ_0	Iterative	$***$
One-pixel [68]	Black box	Non Targeted	Image specific	ℓ_0	Iterative	$**$
C&W attacks [36]	White box	Targeted	Image specific	$\ell_0, \ell_2, \ell_\infty$	Iterative	$*****$
DeepFool [72]	White box	Non targeted	Image specific	ℓ_2, ℓ_∞	Iterative	****
Uni. perturbations [16]	White box	Non targeted	Universal	ℓ_2, ℓ_∞	Iterative	$*****$
UPSET [146]	Black box	Targeted	Universal	ℓ_∞	Iterative	****
ANGRI [146]	Black box	Targeted	Image specific	ℓ_∞	Iterative	****
Houdini [131]	Black box	Targeted	Image specific	ℓ_2, ℓ_∞	Iterative	****
ATNs [42]	White box	Targeted	Image specific	ℓ_∞	Iterative	****