Adversarial Attack and Learning

Anosh Billimoria, billimor@rhrk.uni-kl.de, @anbilli (Mattermost)

Setup: The constraints on δ can be summarized where, $|\delta|$ is the norm of the noise, ϵ and α are integers.

(i) …. $|\delta| = |\delta|_{\infty} + \alpha|\delta|_2$

such that, (ii) …. $|\delta| \leq \epsilon$

Procedure: To keep matters simple LeNet-5 CNN is selected.

Constraint Simplification: To bound $|\delta|$ within desired ϵ and we estimate the $|\delta|_{\infty}$ using gradient method and then clipping the result. To that end here we have selected some methods to modify the constraints over the norm using some simplifications. Namely,

1. Since it is not specified the noise is randomly drawn from a uniform distribution between -ϵ and ϵ of size equal to the image size (28,28)
2. The $|\delta|_2$ is calculated and δ is normalized such that $|\delta|_2 = 1$. Making (i).
   (iii) …. $|\delta'| = |\delta'|_{\infty} + \alpha.1$
   where δ' is the new normalised version. We can modify (ii) similarly.
   (iv) …. $|\delta'|_{\infty} \leq \epsilon - \alpha$
   dividing by α in (ii) and take $\alpha^* = \frac{1}{\alpha}$, we get,
   (v) …. $\alpha^*|\delta'|_{\infty} \leq \alpha^*\epsilon - 1$
3. By reducing the problem to simply depend on the $|\delta'|_{\infty}$ of new normalised δ'. We can apply gradient estimation method and clip results within $(\alpha^*\epsilon - 1)$
4. Here, we can note that to replicate α^* is multiplied with norm-infinity. We can also use it as step-size during the estimation.

Effect of varying ϵ and α: Now we perturb the images.

1. Constant ϵ and varying α^*: We observe that α value can affect the transparency of the digit. A lower value almost blacks out any grey pixels, while a higher value causes extreme alterations making the digit incomprehensible even for humans.
2. Varying ϵ: A larger ϵ distorts images beyond recognition and leaves lesser room for selecting appropriate α^* to balance out the effect. On the other hand, a smaller ϵ can allow images to have more contrast for a wider range of α^* .
3. We can also say that choosing a just right ϵ can offer better tuning opportunities on α^*, especially since our modification to α^* makes it act like step-size for the infinity-norm estimate.

Training: We have selected α^*= 0.4, ϵ = 0.7, as the adversarial training parameters.

1. **Choice of α^* and ϵ:** For this combination the images are still recognizable by the human eye, yet the accuracy dropped to two-thirds of the original. Since the values are well-balanced with respect to each other this combination can potentially train a more universally robust model.
2. **First training:** We decided to retrain the model and add an adversarial image to the dataset every tenth batch using the above values. Validation is performed on the adversarial images from the test set. The results are interesting since testing on other combinations of ϵ and α^* tend to respond with better accuracy and lower loss than before.
3. **Second training:** Train a newly initialized model with every tenth batch is replaced with an adversarial one. The results of which are not impressive probably due to using a simple model. It may be beneficial to modify the model slightly to incorporate more robustness.