On the Robustness of Metric Learning: An Adversarial Perspective

Dmitry Sotnikov   Nikita Fedyashin

Vega Institute

31 may 2022

Problem statement

Due to the increasing popularity of metric-learning, we should worry about the reliability of the neural networks. This work is related to the adversarial attacks and possible defence mechanisms.

Goal

The project was based on the article of Huai et al.[1] Our goal was to implement and verify the proposed adversarial attack and defence mechanisms.

Possible application

Generalization and extension of the proposed method to a wider class of architectures.

[1] Huai, M., Zheng, T., Miao, C., Yao, L., Zhang, A. (2022). On the Robustness of Metric Learning: An Adversarial Perspective. ACM Transactions on Knowledge Discovery from Data (TKDD), 16(5), 1-25.

Metric learning

Set of instances $\mathcal{X} = \{x_i\}_{i=1}^{N} \subset \mathbb{R}^d$. Data is normalized to $[0, 1]^d$.

Goal

Learn mapping $f\colon \mathbb{R}^d \to \mathbb{R}^m$ into feature space, based on which the similarity can be calculated.

The similarity degree:

$$D(x_i, x_j) = \|f(x_i) - f(x_j)\|_2^2.$$

Two instances x_i, x_j are supposed to be similar iff

$$D(x_i, x_j) \le \gamma$$

for a priori given γ.

Adversarial attack (AckMetric)

The pairwise robustness of the given distance metric function D with respect to the instance pair (x_i, x_j) is given by

$$\rho(D; (x_i, x_j)) = \min_{\delta_i\ \delta_j \in \mathbb{R}^d} \{||\delta_i||_\infty + ||\delta_j||_\infty \colon D\left(x_i + \delta_i, x_j + \delta_j\right) > \gamma\}.$$

Here minimizers δ_i and δ_j are called adversarial perturbations. Generation of adversarial perturbations can be reduced to the following optimization problem:

$$D\left(x_i + \delta_i, x_j + \delta_j\right) - \gamma \longrightarrow \max_{||\delta_i||_\infty + ||\delta_j||_\infty < \varepsilon}$$

$$s.t.\ x_i + \delta_i \in [0,1]^d,\ x_j + \delta_j \in [0,1]^d.$$

Projected gradient descent

$(r+1)$-th iteration update:

$$x_k^{r+1} = \Pi_{clip}\left(x_k^r + \xi \,\mathrm{sign}\left(\frac{\partial D(x_i^r,\, x_j^r)}{\partial x_k^r}\right)\right),$$

$$k = \mathrm{argmax}_{\{i,\, j\}}\left\{\left\|\frac{\partial D(x_i^r,\, x_j^r)}{\partial x_i^r}\right\|,\ \left\|\frac{\partial D(x_i^r,\, x_j^r)}{\partial x_j^r}\right\|\right\}.$$

For linear mapping ($f(x) = Wx$) one can obtain explicit formula

$$x_i^{r+1} = \Pi_{clip}\left(x_i^r + \xi \,\mathrm{sign}\left(2W^T W(x_i^r - x_j^r)\right)\right),$$

$$x_j^{r+1} = \Pi_{clip}\left(x_j^r + \xi \,\mathrm{sign}\left(-2W^T W(x_i^r - x_j^r)\right)\right),$$

which can be applied to arbitrary argument as the derivative norms are equal in this case.

Upper bound for perturbed distance

For one layer fully connected neural network ($f(x) = Wx$) the following upper bound for $\|\delta_i\|_\infty \leq \varepsilon$ can be proved:

$$D(x_i + \delta_i, x_j) \leq D(x_i, x_j) + \varepsilon \sup_{s \in [-1,1]^d} \|2W^T W s\| \leq$$

$$\leq D(x_i, x_j) + \varepsilon \sup_{s \in [-1,1]^d, t \in [-1,1]^d} 2t^T W^T W s \leq$$

$$\leq D(x_i, x_j) + \varepsilon \sup_{t \in [-1,1]^d} 2t^T W^T W t \leq D(x_i, x_j) + 2d\varepsilon\lambda_{max},$$

where λ_{max} is maximum eigenvalue of $W^T W$.

Proposed defence method

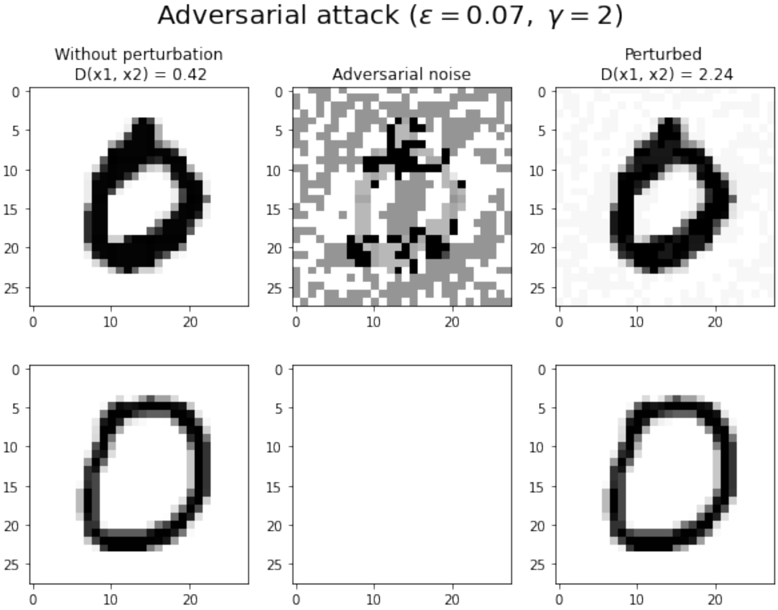
For binary classification problem let $y_i \in \{-1, 1\}$ denote the class of x_i, and $y_{ij} = y_i y_j$.
Margin loss (no defence):

$$\mathcal{L}_1 = \frac{2}{N(N-1)} \sum_{i<j} (1 + y_{ij}(D(x_i, x_j) - \gamma))^+ \to \min_W$$

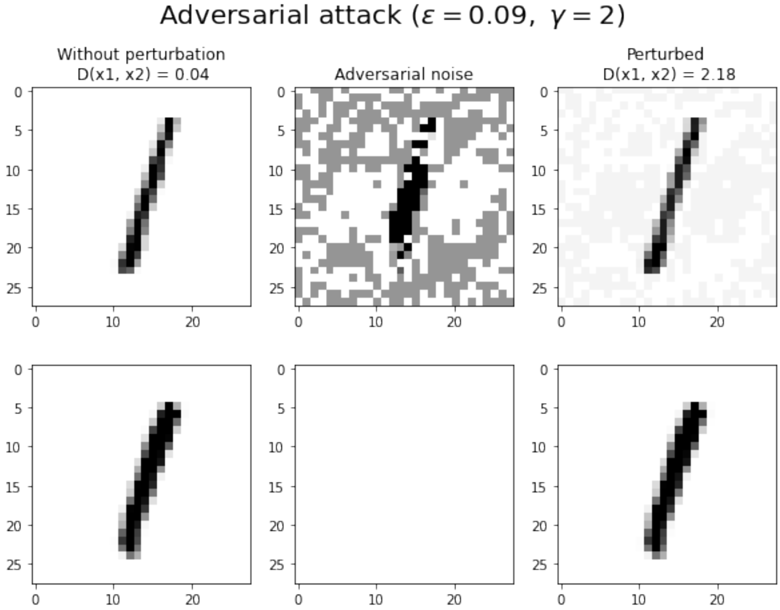
Let $\lambda_d, \lambda_{d-1}, \ldots$ denote top maximum eigenvalues of $W^T W$.
Then the robust loss function has the form

$$\mathcal{L}_2 = \mathcal{L}_1 + \sum_{j=1}^{k} \alpha_j \lambda_{d-k+j} \to \min_W .$$

Attack evaluation



Attack evaluation



Defence Evaluation