Training of stable neural ordinary differential equations

Arturo De Marinis, Nicola Guglielmi, Anton Savostianov, Stefano Sicilia, Francesco Tudisco
arturo.demarinis@gssi.it
GSSI - Gran Sasso Science Institute, Viale Francesco Crispi 7, L'Aquila, 67100, Italy

Neural ODEs [1] are ordinary differential equations whose vector field is a neural network. The numerical integration of neural ODEs defines a deep neural network.

As all neural networks, neural ODEs are vulnerable to adversarial attacks, i.e. imperceptible perturbations, added to the inputs of a neural network, designed in such a way that the output corresponding to the perturbed input is far away from the output corresponding to the original input. Nevertheless, neural ODEs are ordinary differential equations, thus the stability and contractivity theory of ODEs can be applied to make neural ODEs robust and stable against adversarial attacks.

Our contribution is in this direction [2]. We consider the neural ODE

$$\dot{x}(t) = \sigma(Ax(t) + b), \qquad t \in [0, T],$$

where $x : [0, T] \to \mathbb{R}^n$ is the feature vector evolution function, $A \in \mathbb{R}^{n,n}$ and $b \in \mathbb{R}^n$ are the parameters, and $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, assumed to be smooth and such that $\sigma'(\mathbb{R}) \subset [m, 1]$, with $0 < m \leq 1$.

Then, we notice that the neural ODE is contractive if

$$\sup_{D \in \Omega_m} \mu_2(DA) = -\alpha, \tag{1}$$

where $\Omega_m = \{D \in \mathbb{R}^{n,n} \ : \ D \text{ is diagonal and } m \leq D_{ii} \leq 1,\ \forall i = 1, \ldots, n\}$, $\alpha \geq 0$ and μ_2 denotes the logarithmic 2-norm of a matrix. Therefore, we compute the smallest $0 < m^\star(A) \leq 1$ such that (1) holds with $\alpha = 0$ and, if $m < m^\star(A)$, we set $\delta > 0$ a small constant and replace the matrix A by the shifted matrix

$$\widehat{A} = A - \ell\delta I,$$

where ℓ is the smallest positive integer such that (1) holds for $\widehat{A}$ in place of A, with $m^\star(\widehat{A}) < m$.

Of course, shifting the matrix A, and thus its entire spectrum, to get $m^\star < m$ is a greedy approach. As optimal one (the draft will be available soon), we compute the nearest matrix B to A in Frobenius norm such that

$$\sup_{D \in \Omega_m} \mu_2(DB) = -\alpha.$$

Eventually, after each step of gradient descent in the state-of-the-art training, we apply either the greedy approach or the optimal one to make the neural ODE contractive, i.e. it does not amplify the error in the input data to the output data, and thus robust and stable against adversarial attacks.

To illustrate our methodology, we compare the performance of two neural ODEs for MNIST and FashionMNIST classification against the Fast Gradient Sign Method (FGSM) attack: the former trained according to the state-of-the-art training strategy, and the latter trained according to our proposed strategy. Our experiments indicate that the latter shows a significant improvement in robustness against the FGSM attack.

References

[1] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt and David K. Duvenaud, *Neural Ordinary Differential Equations*, Advances in Neural Information Processing Systems 31, 2018.

[2] Nicola Guglielmi, Arturo De Marinis, Anton Savostianov, Francesco Tudisco, *Contractivity of neural ODEs: an eigenvalue optimization problem*, submitted.