Centre for Analysis, Scientific Computing and Applications

Technische Universiteit Eindhoven University of Technology

Sparsifying Continuous-Time Neural Networks for Density Estimation Problem

Anna Shalova, a.shalova@tue.nl,
supervisor: Prof. W.H.A. Schilders

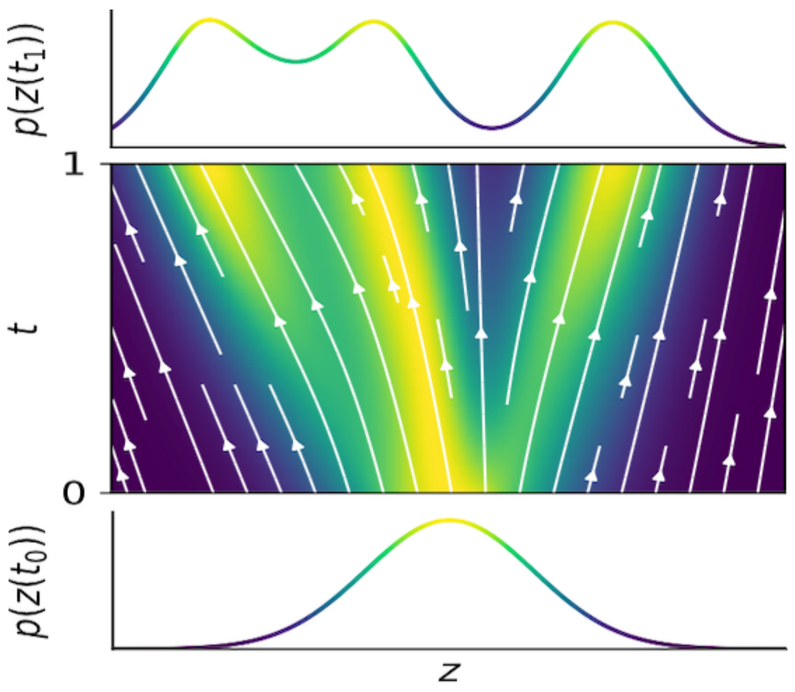

Figure 1: IVP solution as an invertible mapping for the density estimation problem, illustration from [2].The model is trained to maximize likelihood of $z_i(t_0) = \mathcal{M}^{-1}(X_i)$

UNRAVEL

Is an NWO XL project that is aimed to unravel neural networks through structure-preserving computing. Despite various impressive results of deep neural networks in generative and predictive tasks, they have several strong issues, including high computational costs and heuristic nature of the approach. At the same time, structure-preserving methods are strong in providing accurate solutions to complex mathematical problems. So the goal of the UNRAVEL project is to build neural networks using the computationally efficient structure-preserving methods and apply them in various fields of study, from astrophysics to fluid dynamics.
We are a group of scientists from TU/e, CWI and University of Leiden working on 6 subprojects. For more details please visit our website: `https://nwo-unravel.nl`

Dealing with Over-Parametrization

Almost all currently used deep neural networks are overparametrized, which means that the model has more parameters than can be estimated from the data. Even though it does not reduce the generalization ability, the overparametrized models have high memory and computational costs, so many attempts have been made to decrease the number of parameters in neural networks. In particular, three main approaches are:

- distillation (distill knowledge from larger models into smaller ones)
- matrix and tensor decompositions (SVD, TT, Tucker)
- sparsification, which is used in this project.

Continuous-time Models for Density Estimation

Density estimation is a generative problem: having a set of samples $X_1, ...X_k \sim \mathcal{P}_X$, the goal is to get a sampler from the distribution $\mathcal{P}_X$. One of the key approaches is to learn an invertible mapping between the data distribution $\mathcal{P}_X$ and a latent distribution $\mathcal{P}_Z$ that we can easily sample from (usually Gaussian). In [2] it was proposed to model such a mapping $\mathcal{M} : X \to Z$ as a solution of an IVP (1) defined by a parametric function $f_\theta(z,t)$, $\mathcal{M}(x_0) = z(t_0)$.

$$\dot{z} = f_\theta(z,t), \; z(t_1) = x_0 \qquad (1)$$

Jacobian-Spectrum Pruning at Initialization

Is a sparsification method applied before training and aimed to speed up the model's inference. Global truncation error of the IVP solution is related to the spectrum of the Jacobian $J(z) = \frac{df_\theta(z)}{dz}$: the larger are the singular values, the higher is the error. Or, alternatively, the less is the norm of the Jacobian, the less integration steps are needed for the fixed accuracy.
Consider the Taylor expansion of the cumulative Jacobian norm R:

$$R = \int_{t_0}^{t_1} \|J(z(t))\|_F dt \qquad (2)$$

$$R(\theta + d\theta) = R(\theta) + \left\langle \frac{dR}{d\theta}, d\theta \right\rangle + O(\|d\theta\|^2). \qquad (3)$$

During training the gradient steps are made in the direction of the loss function gradient $d\theta = -\frac{dL}{d\theta}$. Based on this fact we propose the following sparsification algorithm:

- evaluate the gradient of the loss function $dL = \frac{dL}{d\theta}$,
- evaluate the gradient $dR = \frac{dR}{d\theta}$,
- choose the sparsity pattern with the largest parameters' scores $s_i = dR_i * dL_i$.

In such a way, for the selected set of parameters the gradient of the loss is expected to be aligned with gradient of the Jacobian norm. The method is an alternative to the Jacobian norm regularization [1].

First Results

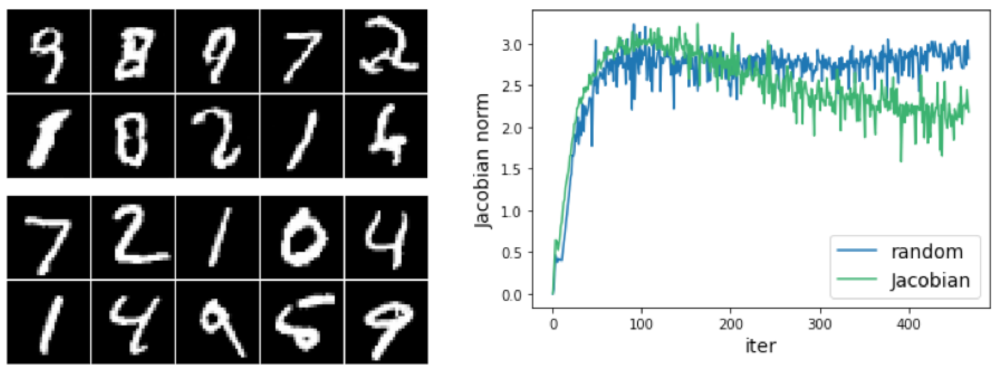

Figure 2: Jacobian Spectrum Pruning performance for the MNIST dataset. On the left: samples from the pruned model (above) compared to the samples from the validation dataset (below). On the right: cumulative Jacobian norm during training for random and Jacobian spectrum pruning.

References

[1] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ode: the world of jacobian and kinetic regularization. In *International Conference on Machine Learning*, pages 3154–3164. PMLR, 2020.

[2] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2018.

45th Woudschoten Conference,
6-8 October, 2021