Momentum Residual Neural Networks

Michael E. Sander
DMA, ENS and CNRS
michael.sander@ens.fr

Abstract: The training of deep residual neural networks (ResNets) with backpropagation has a memory cost that increases linearly with respect to the depth of the network. A way to circumvent this issue is to use reversible architectures. We propose to change the forward rule of a ResNet by adding a momentum term. The resulting networks, momentum residual neural networks (Momentum ResNets), are invertible. Unlike previous invertible architectures, they can be used as a drop-in replacement for any existing ResNet block. We show that Momentum ResNets can be interpreted in the infinitesimal step size regime as second-order ordinary differential equations (ODEs) and exactly characterize how adding momentum progressively increases the representation capabilities of Momentum ResNets. Our analysis reveals that Momentum ResNets can learn any linear mapping up to a multiplicative factor, while ResNets cannot. In a learning to optimize setting, where convergence to a fixed point is required, we show theoretically and empirically that our method succeeds while existing invertible architectures fail. We show on CIFAR and ImageNet that Momentum ResNets have the same accuracy as ResNets, while having a much smaller memory footprint, and show that pre-trained Momentum ResNets are promising for fine-tuning models.

Joint work with: Pierre Ablin, Mathieu Blondel and Gabriel Peyré [1].

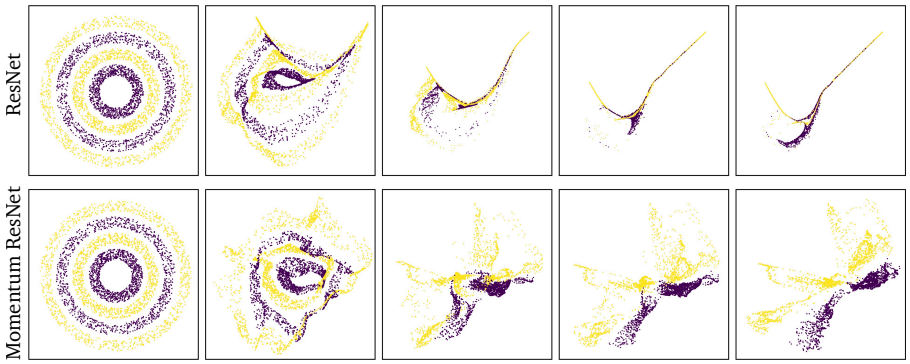

Figure 1: **Separation of four nested rings** using a ResNet (upper row) and a Momentum ResNet (lower row). From left to right, each figure represents the point clouds transformed at layer $3k$. The ResNet fails whereas the Momentum ResNet succeeds.

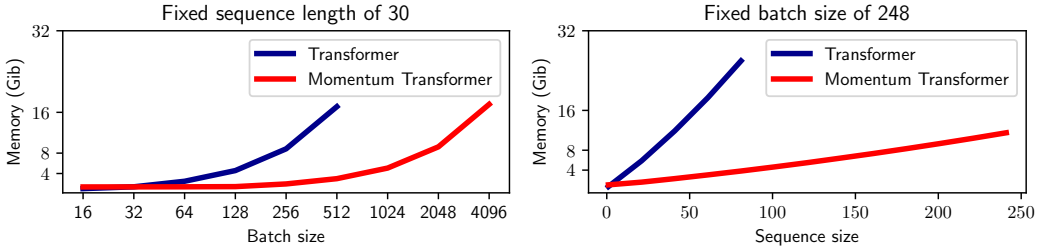

Figure 2: **Memory used** (using a profiler) for a Transformer and a Momentum Transformer on one training epoch, as a function of the batch size (left) and sequence size (right).

References

[1] Sander, M. E., Ablin, P., Blondel, M., and Peyré, G. Momentum residual neural networks. *In Proceedings of the 38th International Conference on Machine Learning, volume 139*, 9276–9287, 2021.