Training Generative Reversible Networks

Reversible networks are by-design invertible neural networks, that had previously been trained by maximum likelihood and adversarial approaches on the generated data [1,2,3,4]. Here, we try:

1. Optimal transport on minibatches for MNIST
2. Adversary on latent space (as in adversarial autoencoders) for CelebA

Reversible Network Architecture

Our F/G functions (for the reversible Block)

Reversible block (as in prior work)

Top: forward

Bottom: inverse

Our subsampling step

Applied twice to a 4x4 input.

On the right, individual squares represent individual channels, so each channel has a single value at the end. At the end, both streams have access to pixels that cover the entire 4x4 input.

Artifacts in earlier training phases

Generated samples of untrained RevoNet with green and purple artifacts, caused by some latent dimensions strongly influencing latent dimensions they correspond to (details in paper).

Network architecture

Numbers behind RevBlock indicate number of filters.

Optimal transport on MNIST

Optimize RevNet parameters and prior distribution parameters (with one distribution per class) using optimal transport on minibatches:

- Outputs from RevNet
- Samples from prior Matching

- Outputs from updated RevNet
- Samples from updated prior

Despite based gradients, still reasonable results on MNIST:

Samples on MNIST

Realistic digits, somewhat blurry, lack diversity

Varying three dimensions of latent prior

These dimensions with largest standard deviations seem to roughly correspond to tilt, thickness and size, respectively. Different classes use same dimensions even though all dimensions usable by model.

Adversarial approach on CelebA

Optimize RevNet parameters using adversary in latent space:

- Outputs from RevNet
- Samples from prior

- Outputs from updated RevNet
- Samples from same prior

Reconstructions

Top row: original; middle row: reconstruction from latent space restricted the prior distribution; bottom row: reconstruction from full latent space

- Interpolations restricted

Interpolations in the encoding space restricted to narrow sections of the prior. Intermediate images clearly resemble human faces.

Generated Samples

Generated Samples

< Interpolations full

Interpolations in the full encoding space. More details, but also more artifacts.

Conclusion

- Reversible Networks can be trained inside adversarial autoencoder framework
- More comparison to prior work on invertible generative models needed
- Optimal transport approach interestingly leads to same dimensions encoding same concepts for different classes
- Optimal transport approach needs further ideas how to scale to larger and more diverse datasets (semi-dual approach?)
- Further improvement of existing hierarchical invertible architectures [2] might be promising

Contact and References

Robin Tibor Schirrmeister, robintibor@gmx.de
Translational Neurotechnology Lab, Albert-Ludwigs-University Freiburg

Department of Neuroinformatics, University of Freiburg, Freiburg, Germany