Recurrent Generative Stochastic Networks for Sequence Prediction

Abstract

We present a new generative model for unsupervised learning of sequence representations, the recurrent generative stochastic network (RNN-GSN).

1. Introduction

Unsupervised sequence learning is an important problem in machine learning given that most information (ranging from speech to video to even consumer behavior) is often unlabeled and has a sequential structure. Most of these sequences consist of high-dimensional, complex objects such as words in text, images in video, or chords in music. Recently, recurrent neural networks (RNN) have become state-of-the-art for sequence representation because they have an internal memory that can learn long-term temporal dependencies.

Maybe I should write this at the end?

2. Generative Stochastic Networks

Generative stochastic networks (GSN) are a generalization of the denoising auto-encoder and help solve the problem of mixing between many major modes of the input data distribution.

Denoising auto-encoders use a Markov chain to learn a reconstruction distribution $P(X|\widetilde{X})$ given a corruption process $C(\widetilde{X}|X)$ for some data X. Denoising auto-encoders have been shown as generative models (Bengio et al., 2013b), where the Markov chain can be iteratively sampled from:

$$X_t \sim P_\Theta(X|\widetilde{X}_{t-1})$$
$$\widetilde{X}_t \sim C(\widetilde{X}|X_t)$$

As long as the learned distribution $P_{\Theta_n}(X|\widetilde{X})$ is a consistent estimator of the true conditional distribution $P(X|\widetilde{X})$ and the Markov chain is ergodic, then as $n \to \infty$, the asymptotic distribution $\pi_n(X)$ of the generated samples from the denoising auto-encoder converges to the data-generating distribution $P(X)$(Bengio et al., 2013b)).

2.1. Easing restrictive conditions on the denoising auto-encoder

A few restrictive conditions are necessary to guarantee ergodicity of the Markov chain - requiring $C(\widetilde{X}|X) > 0$ everywhere that $P(X) > 0$. Particularly, a large region V containing any possible X is defined such that the probability of moving between any two points in a single jump $C(\widetilde{X}|X)$ must be greater than 0. This restriction requires that $P_{\Theta_n}(X|\widetilde{X})$ has the ability to model every mode of $P(X)$, which is a problem this model was meant to avoid.

To ease this restriction, Bengio et al. (Bengio et al., 2013a) proves that using a $C(\widetilde{X}|X)$ that only makes small jumps allows $P_\Theta(X|\widetilde{X})$ to model a small part of the space V around each $\widetilde{X}$. This weaker condition means that modeling the reconstruction distribution $P(X|\widetilde{X})$ would be easier since it would probably have fewer modes.

However, the jump size σ between points must still be large enough to guarantee that one can jump often enough between the major modes of $P(X)$ to overcome the deserts of low probability: σ must be larger than half the largest distance of low probability between two nearby modes, such that V has at least a single connected component between modes. This presents a tradeoff between the difficulty of learning $P_\Theta(X|\widetilde{X})$ and the ease of mixing between modes separated by this low probability desert.

2.2. Generalizing to GSN

While denoising auto-encoders can rely on X_t alone for the state of the Markov chain, GSNs introduce a latent variable H_t that acts as an additional state variable in the Markov chain along with the visible X_t (Bengio et al., 2013a):

$$H_{t+1} \sim P_{\Theta_1}(H|H_t, X_t)$$
$$X_{t+1} \sim P_{\Theta_2}(X|H_{t+1})$$

The resulting computational graph is shown in Figure 1.

The latent state variable H can be equivalently defined as $H_{t+1} = f_{\Theta_1}(X_t, Z_t, H_t)$, a learned function f with an independent noise source Z_t such that X_t cannot be reconstructed exactly from H_{t+1}. If X_t could be recovered from H_{t+1}, the reconstruction distribution would simply converge to the Dirac at X. Denoising auto-encoders are

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

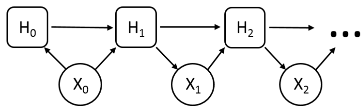

Figure 1. GSN computational graph.

therefore a special case of GSNs, where f is fixed instead of learned.

GSNs also use the notion of walkback to aid training. The resulting Markov chain of a GSN is inspired by Gibbs sampling, but with stochastic units at each layer that can be backpropagated (Rezende et al., 2014).

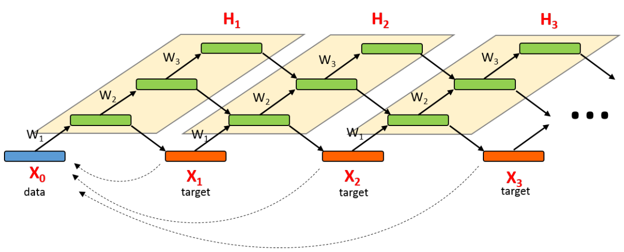

Figure 2. Unrolled GSN Markov chain.

3. Recurrent Neural Networks

This is the RNN section.

4. The RNN-GSN

This is the main RNN-GSN section.

5. Experiments

This is the experiments section.

5.1. Sequences of MNIST digits

arbitrary sequences of images.

- Sequence1 is a simple linear sequence of digits 0-9 repeating.
- Sequence2 introduces one bit of parity by alternating sequences 0-9 and 9-0 repeating.
- Sequence3 gives a slightly longer-term time dependency.
- Sequence4 creates a more non-linear sequence with two bits of parity.

5.2. Sequences of polyphonic music

midi stuff.

- Piano-midi.de ....
- Nottingham .....
- MuseData ....
- JSB chorales .....

6. Conclusion

This is the conclusion.

References

Bengio, Yoshua, Thibodeau-Laufer, Eric, and Yosinski, Jason. Deep generative stochastic networks trainable by backprop. *CoRR*, abs/1306.1091, 2013a.

Bengio, Yoshua, Yao, Li, Alain, Guillaume, and Vincent, Pascal. Generalized denoising auto-encoders as generative models. *CoRR*, abs/1305.6663, 2013b.

Rezende, Danilo J., Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generateive models. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.