From Variational to Deterministic Autoencoders

1. Overview

Sampling a stochastic encoder can be interpreted as injecting noise into the input of a deterministic decoder.

Non-stochastic regularization can lead to equally smooth and meaningful latent representations.

Introduce a density estimation step that allows you to generate from deterministic autoencoders.

Problems with the variational objective: 1. Tradeoff sample quality and reconstruction quality. 2. Over-regularization induced by the KL term (as seen in WAE paper). 3. Trivial solutions that lead to posterior collapse. 4. Approximate expectations through sampling at the cost of increased variance in gradients. 5. Even when things go well the posterior distribution rarely matches the assumed latent prior hurting the quality of generated samples.

And people spend effort trying to fix these drawbacks.

Instead... throw out the variational framework.

2. Constant Variance Encoders: Intermediate Step

Fix the variance of $q_\phi(z|x)$ to be constant for all x,

$$E_\phi^{CV}(x) = \mu_\phi(x) + \epsilon$$

where $\epsilon\ N(0, \sigma I)$. This simplifies the KL loss term. With learned variance we had,

$$2L_{KL} = ||\mu_\phi(x)||_2^2 + d + \sum_i^d \sigma_\phi(x)_i - \log \sigma_\phi(x)_i$$

With constant variance we get,

$$L_{KL}^{CV} = ||\mu_\phi(x)||_2^2$$

We will investigate this model later empirically later. One problem is that treating σ_ϕ as a constant breaks the assumption of $p(z)$ as an isotropic Gaussian.

3. Deterministic Regularized Autoencoders

Concept: In common VAE implementations the input to the decoder is the mean augmented by random Gaussian noise. The noise injection is a key factor in having a regularized decoder. But this is a technique that has been known for decades (Sietsma Dow 1991) which helps smooth the function learned at the price of increased variance.

Core Idea: Substitute noise injection by an explicit regularization scheme for the decoder. New objective:

$$L = L_{rec} + \beta L_Z^{RAE} + \lambda L_{REG}$$

where L_{REG} is the explicit regularizer for the decoder and $L_Z^{RAE} = \frac{1}{2}||z||_2^2$ from simplifying the KL is equivalent to constraining the size of the learned latent space.

4. Regularization Schemes for RAEs

There are a number of options: 1. Tikhonov regularization (equal to addition of low magnitude input noise). $L_{REG} = ||\theta||_2^2$ which is like weight decay on the decoder parameters. 2. Instance noise like in GANs. 3. Weight clipping like in GANs. 4. Gradient penalty on the discriminator $L = ||\nabla D_\theta(E_\phi(x)))||_2^2$ which bounds the gradient norm of the decoder wrt to the input. 5. Spectral normalization.

Appendix presents a principled derivation of RAE framework

5. Ex-Post Density Estimation

We no longer parameterize a simple distribution so we have lost our generation.
Fit a density estimator q_δ to $z = E_\phi(x) | x \in X$.