Fixing a broken ELBO

Alemi, et al. 2018

Motivation

- Progress in VAEs have led to good generative performance.
- Training via ELBO does not necessarily lead to good representation performance.
- This work focuses on developing losses that give good representations.

Mutual information based loss

• "Good" representation is analogous to higher mutual information between the observations and the latents

$$I_{e}(X;Z) = \iint dx \, dz \, p_{e}(x,z) \log \frac{p_{e}(x,z)}{p^{*}(x)p_{e}(z)}.$$
 (1)

• Computing the mutual information can be intractable, lower and upper bounds can be developed:

$$H - D \le I_e(X; Z) \le R \tag{2}$$

Mutual information based loss

Here,
Constant for a given
$$\longrightarrow H \equiv -\int dx \, p^*(x) \log p^*(x)$$

Distortion $\longrightarrow D \equiv -\int dx \, p^*(x) \int dz \, e(z|x) \log d(x|z)$ (4)
Rate $\longrightarrow R \equiv \int dx \, p^*(x) \int dz \, e(z|x) \log \frac{e(z|x)}{m(z)}$ (5)

"Marginal", approximation to p(z)

Distortion-Rate phase maps



- Assuming that obtaining tighter bounds to MI is good, it is of interest to get closer to the D + R = H line.
- It can be shown that a "perfect" m(z) and a "perfect" d(x|z) will take us to the D + R = H line.

$$H \equiv -\int dx \, p^*(x) \log p^*(x) \tag{3}$$

$$D \equiv -\int dx \, p^*(x) \int dz \, e(z|x) \log d(x|z) \qquad (4)$$

$$R \equiv \int dx \, p^*(x) \int dz \, e(z|x) \log \frac{e(z|x)}{m(z)} \tag{5}$$

Distortion-Rate based loss



- Low distortion implies better encoding and decoding of data.
- Therefore, a good loss function must minimize *R* while also keeping *D* as low as possible.

$$H \equiv -\int dx \, p^*(x) \log p^*(x) \tag{3}$$

$$D \equiv -\int dx \, p^*(x) \int dz \, e(z|x) \log d(x|z) \qquad (4)$$

$$R \equiv \int dx \, p^*(x) \int dz \, e(z|x) \log \frac{e(z|x)}{m(z)} \tag{5}$$

Distortion-Rate based loss

Consider the following optimization problem and its equivalent form:

 $egin{aligned} \min_{e,d,m} D \ s.\,t.\,R \leq \epsilon \end{aligned}$

$$egin{aligned} \min_{e,d,m} D + eta R \ \min_{e(z|x),m(z),d(x|z)} \int dx \, p^*(x) \int dz \, e(z|x) \ \left[-\log d(x|z) + eta \log rac{e(z|x)}{m(z)}
ight]. \end{aligned}$$

eta=1 Corresponds to the standard VAE loss

Experiments: Toy problem



Figure 2. Toy Model illustrating the difference between fitting a model by maximizing ELBO (b) vs minimizing distortion for a fixed rate (c). **Top (i):** Three distributions in data space: the true data distribution, $p^*(x)$, the model's generative distribution, $g(x) = \sum_z m(z)d(x|z)$, and the empirical data reconstruction distribution, $d(x) = \sum_{x'} \sum_z \hat{p}(x')e(z|x')d(x|z)$. **Middle (ii):** Four distributions in latent space: the learned (or computed) marginal m(z), the empirical induced marginal $e(z) = \sum_x \hat{p}(x)e(z|x)$, the empirical distribution over zvalues for data vectors in the set $\mathcal{X}_0 = \{x_n : z_n = 0\}$, which we denote by $e(z_0)$ in purple, and the empirical distribution over z values for data vectors in the set $\mathcal{X}_1 = \{x_n : z_n = 1\}$, which we denote by $e(z_1)$ in yellow. **Bottom:** Three $K \times K$ distributions: (iii) e(z|x), (iv) d(x|z) and (v) $p(x'|x) = \sum_z e(z|x)d(x'|z)$.

Experiments: Binary MNIST

	autoe	encoder v	syntact	ic encoder	semanti	ic encoder	autod	ecoder	
	$\beta = (R, D) =$	0.1 (156.0, 4.8)	$\beta = (R, D) =$	0.9 (120.3, 8.1)	$\beta = (R, D) =$	1.0 (6.2, 74.1)	$\beta = 1.1$ (<i>R</i> , <i>D</i>) = (0.0, 80.6)		
	ELBO =	- 162.0	ELBO =	- 128.0	ELBO =	- 80.3	ELBO =	- 80.6	
data	sample	average	sample	average	sample	average	sample	average	
2	2	2	2	2	2	2	2	3	
0	0	0	0	0	Ø	0	¥	ß	
7	1	1	1	1	1	1	4	à	
8	8	8	8	8	8	8	3	\$	

(a) Reconstructions from -+v with $\beta = 0.1 - 1.1$.

autoe -+ β= (R,D) = ELBO =	oncoder 0.1 (156.0, 4.8) - 162.0	syntact -+ β = (R, D) = ELBO =	ic encoder v 0.9 = (120.3, 8.1) = - 128.0	semanti -+ β = (R, D) = ELBO =	v 1.0 (6.2, 74.1) - 80.3	autodecoder -+v $\beta = 1.1$ (R, D) = (0.0, 80.6) ELBO = - 80.6			
sample	average	sample	average	sample	average	sample	average		
1	1	5	5	ŝ	3	3	3		
2	3	1	1	6	ŀ	v	8		
5	5	; ;	9	7	7	5	ß.		
3	3	9	9	6	6	6	Ø		

(b) Generations from -+v with $\beta = 0.1 - 1.1$

	syntactic encoder				autodecoder				syntactic encoder		semantic encoder		autode		ecoder		
	$\beta = 1.0$		$\beta =$	-v 1.0	$\beta = 1.0$		$\beta = 1.0$			$\beta = 1.0$		$\beta = 1.0$		$\beta = 0.4$		$\beta = 1.1$	
	(R, D) = (25.4, 59.3) ELBO = -84.7		4,59.3) (R,D) = (25.8,57.7) .7 ELBO = -83.4		(R, D) = (0.0, 81.0) ELBO = -81.0		(R, D) = (0.0, 80.7) ELBO = -80.7			(R, D) = (25.8, 57.7) ELBO = -83.4		(R, D) = (8.3, 71.9) ELBO = -80.2		(R, D) = (2.8, 78.2) ELBO = -81.0		(R, D) = (0.0, 80.6) ELBO = -80.6	
data	sample	average	sample	average	sample	average	sample	average	data	sample	average	sample	average	sample	average	sample	average
2	2	2	2	2	Θ	3	5	茱	2	2	2	2	2	O	Q.	2	3
0	0	0	0	0	4	Ŧ	0	Ŕ	0	0	0	0	Ø	0	0	¥	ß
7	1	1	7	1	7	Ì.	8	督	1	1	1	7	1	8	*	4	ð
8	8	8	8	8	Ŧ	Å	8	Ċ.	8	8	8	8	8	Q	Ą	3	Ş

References

- Alemi, Alexander, et al. "Fixing a broken ELBO." *International Conference on Machine Learning*. PMLR, 2018.
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).