

# Notes: “Improving VAE generations of multimodal data through data-dependent conditional priors”

Sun

March 23, 2020

## 1 Notes:

### 1.1 Summary:

1. Traditional variational autoencoders may struggle to generate good samples when the underlying input data has many different modalities, such as a complex mixture distribution
2. This paper presents a VAE formulation (CP-VAE) with a conditional prior that ideally learns to entirely separate different modalities
3. The latent variable in a CP-VAE model is composed of both a discrete and a continuous piece.

### 1.2 Medium level:

As always, one of the issues with the traditional VAE formulation is the huge amount of assumptions inherently made. This paper addresses a commonly looked at one; the isotropic Gaussian prior. By using both a continuous prior (a Gaussian was used) and a categorical discrete prior, the model may have an easier time distinguishing between vastly different modes via the use of a discontinuous latent space. While normally, a discontinuous latent space is highly unappealing, the sampling procedure they also provide mitigates this problem.

### 1.3 Low-ish level:

We start with the new  $p_\theta(\mathbf{x})$ :

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{c}} \int p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c}) p_\phi(\mathbf{z}|\mathbf{c}) p(\mathbf{c}) d\mathbf{z} \quad (1)$$

It is a two level hierarchical generative process; the latent space is composed of the traditional Gaussian continuous  $\mathbf{z}$  and the discrete  $\mathbf{c}$  component with the joint distribution  $p(\mathbf{z}, \mathbf{c}) = p_\phi(\mathbf{z}|\mathbf{c})p(\mathbf{c})$ . The authors assume a uniform categorical prior for  $\mathbf{c}$ .

Instead of the traditional ELBO, instead we optimize a joint KL term.

$$\text{ELBO} := \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}, \mathbf{c} | x_i)} \left[ \log p_\theta(\mathbf{x}_i | \mathbf{z}, \mathbf{c}) \right]}_{\text{① Reconstruction likelihood}} - \underbrace{D_{\text{KL}} \left[ q_\phi(\mathbf{z}, \mathbf{c} | \mathbf{x}_i) || p_\psi(\mathbf{z}, \mathbf{c}) \right]}_{\text{② Prior constraint}} \quad (2)$$

② can instead be rewritten as the sum of two separate KL terms; the categorical and continuous parts.

$$\text{②} = D_{\text{KL}} \left[ q_\phi(\mathbf{z}, \mathbf{c} | \mathbf{x}_i) || p_\psi(\mathbf{z}, \mathbf{c}) \right] = D_{\text{KL}} \left[ q_\phi(\mathbf{c} | \mathbf{x}) || p(\mathbf{c}) \right] + \mathbb{E}_{q_\phi(\mathbf{c} | \mathbf{x})} D_{\text{KL}} \left[ q_\phi(\mathbf{z} | \mathbf{c}, \mathbf{x}_i) || p_\psi(\mathbf{z} | \mathbf{c}) \right] \quad (3)$$

As such, the minimization of the initial KL term pushes both distributions towards the given prior.

$\mathbf{c}$  itself is supposed to be a pure categorical (one-hot) vector, however, considering one cannot backprop through the argmax operator, a Gumbel-softmax reparameterization is used.

In implementation, the authors made an encoder with 3 outputs:  $\mu$ ,  $\log \sigma$ , and a Gumbel-softmax representation of  $\mathbf{c}$ . The input of the decoder was then  $\mathbf{z}$  with  $\mathbf{c}$  concatenated on. Considering this implementation, one can think of  $\mathbf{c}$  as the output of a function that takes in a sample,  $\mathbf{x}$ , and ‘softly’ predicts what modality it lies in; this is learned implicitly by the model. On MNIST, using only  $\mathbf{c}$ , the model achieves about 95% accuracy without labels.

## 2 Theoretical improvements:

When compared to a vanilla VAE, CP-VAE seem to have the following positives:

1. Possible better generated recreations if the underlying data has many different roughly-grained modes
2. Possible more flexible posterior

Excluding possible pathological edge cases, there doesn’t seem to be many drawbacks. However, the authors did not test CP-VAE on any dataset of reasonable dimension, so no real conclusions can be drawn.