A NOTE ON DEEP VARIATIONAL MODELS FOR UNSUPERVISED CLUSTERING

Rui Shu & Curtis Langlotz
Biomedical Informatics
Stanford University
{ruishu,langlotz}@stanford.edu

James Brofos
The MITRE Corporation
jbrofos@mitre.org

ABSTRACT

Recently, the Gaussian Mixture Variational Autoencoder (GMVAE) has been introduced to handle unsupervised clustering (Dilokthanakul et al., 2016). However, the existing formulation requires the introduction of the free bits term into the objective function in order to overcome the effects of the uniform prior imposed on the latent categorical variable. By considering our choice of generative and inference models, we propose a simple variation on the GMVAE that performs well empirically without modifying the variational objective function.

1 INTRODUCTION

Motivated by the success of the variational autoencoder (VAE) in representation learning and semi-supervised classification (Kingma & Welling, 2013; Chen et al., 2016; Kaae Sønderby et al., 2016; Jimenez Rezende et al., 2014), recent attention has been given to the application of VAE to unsupervised clustering. In this setting, observed data X is generated as a stochastic function of both a latent class variable Y and a continuous latent random variable Z. To perform this task, Dilokthanakul et al. (2016) proposed using a Gaussian Mixture Variational Autoencoder (GMVAE). The authors further highlighted the importance of incorporating a free bits term (Kingma et al., 2016) to overcome over-regularization in some settings.

In this paper, we address the following questions. What is the behavior of models with discrete latent variables when trained on unlabeled data using the standard variational objective function—the evidence lower bound (ELBO)? Is it possible to achieve successful unsupervised clustering with the standard ELBO objective function?

We perform our analysis on variants of the M2 model (Kingma et al., 2014) and a variant of the GMVAE and show that (i) the discrete latent variable may not necessarily be used in the generative model as desired and (ii) by introducing very simple modifications to the model, we can encourage the meaningful incorporation of the discrete latent variables into the generative model.

2 A VARIANT OF THE GAUSSIAN MIXTURE VAE

In our experiments, we consider a variant of the GMVAE where we instead use the following sequential sampling scheme

$$y \sim \text{Cat}(1/K) \tag{1}$$
$$z \sim \mathcal{N}(\mu_{z|y}(y), \text{diag}(\sigma^2_{z|y}(y))). \tag{2}$$
$$x|z \sim \text{Ber}(\mu_{x|z}(z)), \tag{3}$$

Unlike the original GMVAE, whose marginal distribution in Z_1 is a mixture of arbitrary distributions, our formulation explicitly ensures that the marginal distribution of Z is a mixture of K factorized Gaussians. We reason that having a marginal distribution as a mixture of Gaussians is desirable due to the unimodality of each Gaussian component. This ensures that samples generated from the same Y tend to be closer in the space of Z.

For inference, we factorize using top-down inference and as well as the precision-weighted merging scheme proposed by Kaae Sønderby et al. (2016) to perform the inference of $q(z|x,y)$,

$$q(z|x,y) \propto p(z|y)\hat{q}(z|x) \tag{4}$$

$$p(z|y) = \mathcal{N}(z|\mu_{z|y}, \text{diag}(\sigma^2_{z|y})) \tag{5}$$

$$\hat{q}(z|x) = \mathcal{N}(z|\hat{\mu}_{z|x}, \text{diag}(\hat{\sigma}^2_{z|x})) \tag{6}$$

$$\sigma^{-2}_{z|x,y} = \sigma^{-2}_{z|y} + \sigma^{-2}_{z|x} \tag{7}$$

$$\mu_{z|x,y} = \frac{\mu_{z|y}\sigma^{-2}_{z|y} + \hat{\mu}_{z|x}\hat{\sigma}^{-2}_{z|x}}{\sigma^{-2}_{z|y} + \hat{\sigma}^{-2}_{z|x}} \tag{8}$$

In comparison to directly parameterizing $q(z|x,y) = \mathcal{N}(z|\mu(x,y), \text{diag}(\sigma^2(x,y)))$, precision-weighted merging simplifies the inference model by sharing $p(z|y)$ across the inference and generative models. By virtue of the weighting scheme, this also has the added desirable property of encouraging the sampled $z \sim q(z|x,y)$ to be near the cluster region defined by $p(z|y)$.

3 A VARIANT OF THE M2 MODEL

Given the success of semi-supervised learning with VAEs, we re-visited the M2 model (Kingma et al., 2014) and evaluated its performance in unsupervised clustering. Surprisingly, the M2 model fails to perform as an unsupervised clustering algorithm despite its success in semi-supervised learning. The failure of the M2 model to perform unsupervised clustering is intriguing, especially since the M2 model and the GMVAE are intimately related. In particular, we note that M2 can also be interpreted as having a Gaussian mixture marginal distribution in the space of the first hidden layer of the decoder $\mu(y,z)$. This is because the first hidden layer h of the decoder is computed as $h = W_y y + W_z z$, where W_y and W_z are the weights of the first hidden layer. This implies that the marginal distribution of H is potentially a mixture of K Gaussians. However, because W_z is not a square matrix and is unconstrained during optimization, we cannot guarantee that the resulting Gaussian mixture has non-degenerate, factorized Gaussian components. To test whether this has an effect on the model, we modify the M2 model by constraining W_z to be an identity matrix during training. We denote this as the M2-Modified model.

4 EXPERIMENTS

We evaluate the performances of the M2, M2-Modified and GMVAE models on permutation-invariant MNIST, training using the standard variational objective function and with $K = 10$.

4.1 "UNSUPERVISED" CLASSIFICATION

To determine if the clusters are formed according to class labels, we rely on the the prediction protocol in Dilokthanakul et al. (2016). In addition to measuring the classification accuracy, we also look at the empirical conditional entropy value $H(Y|X) = \frac{1}{N}\sum_i H(q(y|x^{(i)}))$ and the ELBO.

Figure 1 shows that the M2 model exhibits starkly different behavior than M2-Modified and GMVAE. The M2 model assigns near-equal probability mass to all K clusters when conditioned on some sample x, resulting in a classification score close to 10%. However, simply by constraining the M2 model, the M2-Modified model achieves significantly better clustering performance. The GMVAE, which explicitly introduces Z as being a mixture of Gaussians, achieves the best clustering behavior.

4.2 ADMISSIBILITY OF NEAR-OPTIMAL CLASSIFIERS

It is not immediately clear why the M2 model learns an uninformative classifier. We consider two possibilities: either the M2 model gets stuck at a local optimum, or learning a proper classifier is not beneficial to improving M2's ELBO. To determine the cause, we remove the classifier $q(y|x)$ from all models and provide the label information Y during training.

Providing the label information is equivalent to achieving an optimal classifier when training the models in an unsupervised fashion. In Table 1, we show that providing the label information does

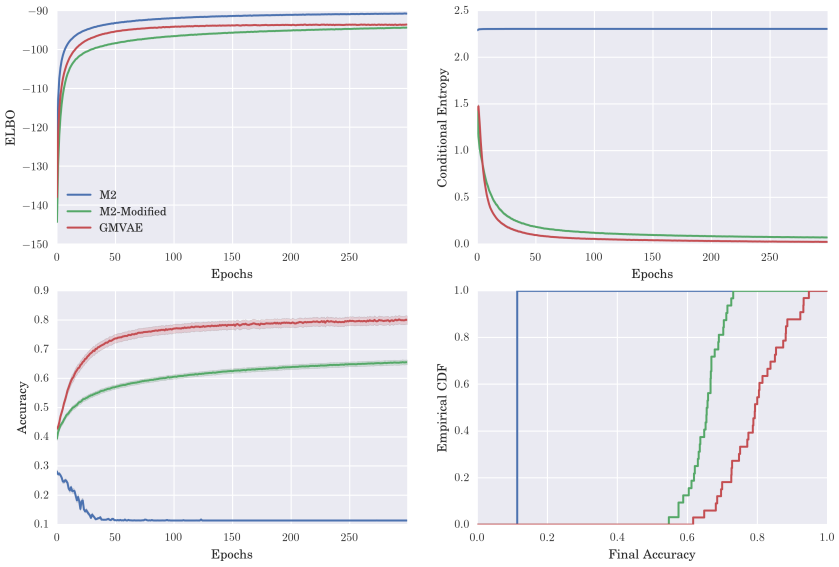

Figure 1: Comparison of the test set ELBO, conditional entropy, and accuracy achieved by the M2, M2-Modified, and GMVAE models.

Model	Accuracy	ELBO	ELBO-Y	Δ ELBO
M2	0.114 ± 0.000	$\mathbf{-90.79 \pm 0.02}$	$\mathbf{-90.75 \pm 0.02}$	0.04
M2-Modified	0.655 ± 0.008	-94.40 ± 0.03	-94.11 ± 0.03	0.29
GMVAE	$\mathbf{0.800 \pm 0.015}$	-93.61 ± 0.11	-90.96 ± 0.04	**2.64**

Table 1: Comparison of the test set ELBO achieved by the models when trained in an unsupervised manner versus the ELBO (denoted as ELBO-Y) achieved by the models when the label information Y is provided. For convenience Δ = ELBO-Y $-$ ELBO is provided, as is the average accuracy achieved by the models from unsupervised training along with the standard errors.

not improve the ELBO for the M2 model. On the other hand, GMVAE benefits significantly when the label information is provided. The improved ELBO resulting from the incorporation of label information demonstrates that the variational objective function encourages the GMVAE to learn a good classifier. In contrast, M2's ability to achieve the best ELBO despite not learning a good classifier suggests that it is exploiting the discrete latent variable Y in an alternative fashion. It remains an open question how exactly M2 is using its discrete latent variable.

Given M2's poor performance as an unsupervised clustering algorithm, it is worth considering how M2 has been successfully applied to semi-supervised classification of MNIST digits (Kingma et al., 2014). In semi-supervised learning, the M2 objective function contains both a term for labeled data and one for unlabeled data. In theory, the weighting of the terms should be proportional to the size of the labeled and unlabeled data sets. In practice, however, implementations of M2 for semi-supervised classification require significant up-scaling of the labeled term (Kingma et al., 2014; Maaløe et al., 2016). We believe that the necessity of up-scaling the labeled term is in part attributable to M2's poor performance as a stand-alone unsupervised clustering algorithm. Given its superior unsupervised clustering performance, the GMVAE may potentially serve as a more attractive alternative for semi-supervised classification.

Ultimately, the fact that such simple changes to the generative and inference models can significantly alter the model behavior signifies the complex interplay with the choice of generative model and variational family. Future lines of work that better explore this dynamic may lead to better insights and improve our ability to learn meaningful representations using deep variational models.

ACKNOWLEDGMENTS

James Brofos' affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.

Approved for Public Release; Distribution Unlimited, Case Number 17-0795.

REFERENCES

X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational Lossy Autoencoder. *ArXiv e-prints*, November 2016.

N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan. Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders. *ArXiv e-prints*, November 2016.

D. Jimenez Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *ArXiv e-prints*, January 2014.

C. Kaae Sønderby, T. Raiko, L. Maaløe, S. Kaae Sønderby, and O. Winther. Ladder Variational Autoencoders. *ArXiv e-prints*, February 2016.

D. P Kingma and M. Welling. Auto-Encoding Variational Bayes. *ArXiv e-prints*, December 2013.

D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-Supervised Learning with Deep Generative Models. *ArXiv e-prints*, June 2014.

Diederik P. Kingma, Tim Salimans, and Max Welling. Improving variational inference with inverse autoregressive flow. *CoRR*, abs/1606.04934, 2016. URL `http://arxiv.org/abs/1606.04934`.

L. Maaløe, C. Kaae Sønderby, S. Kaae Sønderby, and O. Winther. Auxiliary Deep Generative Models. *ArXiv e-prints*, February 2016.