Studying the Propagation of Information in VAE Decoders

Yannick Rudolph*[1,2], Samuel G. Fadel[2], Sebastian Mair[3], and Ulf Brefeld[2]

[1]SAP SE
[2]Leuphana University of Lüneburg
[3]Uppsala University

1 Motivation

A common approach to generative modeling introduces latent variables $\mathbf{z}$ and defines a joint generative model $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, where $p(\mathbf{x})$ refers to the data distribution and $p(\mathbf{z})$ is a known prior distribution. In variational autoencoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014) the joint model allows us to optimize a lower bound $\mathcal{L}$, called ELBO, on the data log-likelihood. The ELBO consists of a reconstruction error involving the decoder $p_\theta(\mathbf{x}|\mathbf{z})$ and of the Kullback-Leibler (KL) divergence between the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$. Convenient choices for the amortized variational posterior and the prior are Gaussians with diagonal covariance matrices, since given these choices it is straightforward to compute the KL divergence in closed-form.

Approaches to increase the power of VAEs focus on more expressive prior distributions and/or variational posteriors. A recent idea is to use implicit rather than prescribed distributions for the variational posterior (Mescheder et al., 2017; Huszár, 2017). Since the KL can no longer be calculated in closed-form, it is estimated via density ratio estimation: during training, KL estimation is reduced to a classification problem in an inner loop.

In this abstract, we propose to use implicit KL estimates to answer the question: How does information, i.e. the mutual information $\mathcal{I}(\mathbf{x}, \mathbf{z})$ between $\mathbf{x}$ and $\mathbf{z}$, propagate within a VAE decoder?

2 Idea

Let $f_1, \ldots, f_L$ be the *deterministic* transformations composing an L-layer decoder. We further denote the posterior, prior and latent representations as $q_{\phi,0}$, p_0, and $\mathbf{z}_0$, respectively. Our key observation is: within the decoder of a VAE, intermediate representations $\mathbf{z}_k = f_k(\mathbf{z}_{k-1})$ for $k = 1, \ldots, L$ can be considered random variables. Via samples from the prior p_0 and posterior $q_{\phi,0}$, we thus observe *implicit* distributions $p_1, \ldots, p_L$ and $q_{\phi,1}, \ldots, q_{\phi,L}$ within the decoder. These distributions relate to priors and variational posteriors respectively, which all now also (partially) depend on θ. With this, we arrive at ELBOs $\mathcal{L}_1, \ldots, \mathcal{L}_L$ that are of the same form as the *standard* ELBO $\mathcal{L}$, but that are wrt. implicit distributions from within the decoder:

$$\begin{aligned}\mathcal{L}_k(\mathbf{x}) =& \mathbb{E}_{q_{\phi,k}(\mathbf{z}_k|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}_k)\right] \\ &- \mathrm{KL}\left(q_{\phi,k}(\mathbf{z}_k|\mathbf{x}) \| p_k(\mathbf{z}_k)\right).\end{aligned}$$

We further observe that expected log-likelihoods are independent of which representation $\mathbf{z}_k$ is used: it holds that $\mathbb{E}_{q_{\phi,k}(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}_k)\right]$ has the same value for all $k = 0, \ldots, L$, as all higher level latent variables deterministically depend on $q_{\phi,0}$ and θ.

We note that having access to implicit KL terms in the decoder allows us to have different lower bounds on $p(\mathbf{x})$. We thus gain insight in the propagation of information in the decoder.

3 Methodology

Using density ratios $r_k(\mathbf{z}_k) = \frac{q_{\phi,k}(\mathbf{z}_k|\mathbf{x})}{p_k(\mathbf{z}_k)}$, the intermediate KL divergences $\mathrm{KL}\left(q_{\phi,k}(\mathbf{z}_k|\mathbf{x}) \| p_k(\mathbf{z}_k)\right)$ can be rewritten as

$$\mathbb{E}_{q_{\phi,k}(\mathbf{z}_k|\mathbf{x})}\left[\log \frac{q_{\phi,k}(\mathbf{z}_k|\mathbf{x})}{p_k(\mathbf{z}_k)}\right] = \mathbb{E}_{q_{\phi,k}(\mathbf{z}_k|\mathbf{x})}\left[\log r_k(\mathbf{z}_k)\right].$$

For $k > 0$, where sampling is possible but densities are not explicitly known, implicit distributions

*Corresp. author: yannick.rudolph@stud.leuphana.de

can be used. The problem of density ratio estimation is reduced to a binary classification problem (Bickel et al., 2007).

Density ratio estimation For each $\mathbf{x}$ in the training data, let samples from the prior distribution $p_k(\mathbf{z}_k)$ be labeled as 0 and samples from the variational posterior distribution $q_{\phi,k}(\mathbf{z}_k|\mathbf{x})$ be labeled as 1. The task of the classifier $D_k(\mathbf{z}_k, \mathbf{x})$ is to output the probability that a sample $\mathbf{z}_k$ is from the variational posterior. Assuming an equal class prior, the optimal classifier is

$$D_k^\star(\mathbf{z}_k, \mathbf{x}) = \frac{q_{\phi,k}(\mathbf{z}_k|\mathbf{x})}{q_{\phi,k}(\mathbf{z}_k|\mathbf{x}) + p_k(\mathbf{z}_k)}.$$

Hence, the KL divergences $\mathrm{KL}\left(q_{\phi,k}(\mathbf{z}_k|\mathbf{x})\|p_k(\mathbf{z}_k)\right)$ can be approximated by

$$\mathbb{E}_{q_{\phi,k}(\mathbf{z}_k|\mathbf{x})}\left[\log D_k(\mathbf{z}_k, \mathbf{x}) - \log(1 - D_k(\mathbf{z}_k, \mathbf{x}))\right].$$

Without explicit access to the densities, the classifiers D_k have to be learned from the labeled samples and the expectation has to be estimated via Monte Carlo sampling.

4 Preliminary experiment

In a preliminary experiment, we estimate KL divergences and mutual information in a trained and in an untrained decoder on synthetic data.

Metrics Classification allows us to estimate the KL divergence for every intermediate representation $\mathbf{z}_k$ $(k = 0, \ldots, L)$. For $k = 0$, where we have access to densities, we validate the estimate against the closed-form KL.

Next to $\mathrm{KL}\left(q_{\phi,k}(\mathbf{z}_k|\mathbf{x})\|p_k(\mathbf{z}_k)\right)$, the KL divergence between the *aggregated* variational posterior $q_{\phi,k}(\mathbf{z}) = \mathbb{E}_{p(\mathbf{x})} q_{\phi,k}(\mathbf{z}_k|\mathbf{x})$ and the prior $p_k(\mathbf{z}_k)$ is also of interest. To estimate $\mathrm{KL}\left(q_{\phi,k}(\mathbf{z}_k)\|p_k(\mathbf{z}_k)\right)$ we also resort to density ratio estimation similar as described above, but with classifiers $D_k(\mathbf{z}_k)$ that are independent of $\mathbf{x}$. With both divergences, we can further estimate the mutual information (MI)

$$\begin{aligned}\mathcal{I}(\mathbf{x}, \mathbf{z}_k) =& \mathbb{E}_{p(\mathbf{x})}\left[\mathrm{KL}\left(q_{\phi,k}(\mathbf{z}_k|\mathbf{x})\|p_k(\mathbf{z}_k)\right)\right] \\ &- \mathrm{KL}\left(q_{\phi,k}(\mathbf{z}_k)\|p_k(\mathbf{z}_k)\right).\end{aligned}$$

For details on the aggregated variational posterior and mutual information wrt. VAEs see Hoffman and Johnson (2016) and Dieng et al. (2019).

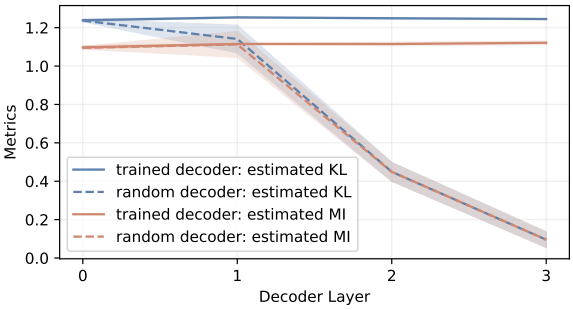

Figure 1: Estimated metrics in the decoder (means and standard errors are over five runs)

Data and model details We train a VAE on data sampled from a two-component Gaussian mixture model in two dimensions. We use a diagonal Gaussian for $q_\phi(\mathbf{z}|\mathbf{x})$, an isotropic Gaussian for $p_\theta(\mathbf{x}|\mathbf{z})$ and standard Gaussian for $p(\mathbf{z})$. The architectures for the variational posterior and decoder are $2 - 16 - 16 - 2 \cdot 16$ and $16 - 16 - 16 - 2$, respectively. All layers are linear and have GELU (Hendrycks and Gimpel, 2016) activations except the output layers. We estimate quantities in intermediate layers before the activations. Classifiers for the density ratio estimation are also fully-connected neural networks.

Results We estimate the metrics in the VAE decoder once after training the VAE and once after randomly re-initializing the decoder. For results see Figure 1. We observe that the trained decoder preserves information, while the random decoder loses information that is present in latent space.

5 Conclusion and outlook

In this abstract, we leveraged ideas for training VAEs with implicit variational posteriors to study the propagation of information in VAE decoders. Our key observation was to introduce KL divergences in terms of intermediate representations of the decoder.

While experimenting, we noticed that the propagation of information in randomly initialized decoders differs for different decoder architectures. We deem it interesting, whether the preservation of information within those decoders can serve as an indicator for the difficulty of training them.

References

Bickel, S., Brückner, M., and Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *International Conference on Machine learning*, pages 81–88.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. (2019). Avoiding Latent Variable Collapse with Generative Skip Models. In *International Conference on Artificial Intelligence and Statistics*, pages 2397–2405.

Hendrycks, D. and Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.

Hoffman, M. D. and Johnson, M. J. (2016). ELBO surgery: yet another way to carve up the variational evidence lower bound. In *(Workshop) Advances in Neural Information Processing Systems*.

Huszár, F. (2017). Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

Mescheder, L., Nowozin, S., and Geiger, A. (2017). Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. In *International Conference on Machine Learning*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286.