1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–5

Informed Priors for Deep Representation Learning

Judith Bütepage BUTEPAGE@KTH.SE
KTH Royal Institute of Technology

Jiawei He JHA203@SFU.CA
Simon Fraser University

Cheng Zhang CHENG.ZHANG@MICROSOFT.COM
Microsoft Research, Cambridge

Leonid Sigal LSIGAL@CS.UBC.CA
University of British Columbia

Greg Mori MORI@SFU.CA
Simon Fraser University

Stephan Mandt MANDT@UCI.EDU
University of California, Irvine

Abstract

In this work, we discuss the possibility of using informed priors for deep generative models. We use word embedding as an example to form such prior. These informed priors aid the learning of the latent variables towards the desired representation instead of superimposing a certain structure. We exemplify the usefulness of method on the problems of image classification and retrieval and use it as an image similarity metric.

1. Introduction

Amortized inference (Kingma and Welling, 2014; Zhang et al., 2017) combines the representational power of deep neural networks with the advantages of Bayesian models, such as generative abilities, uncertainty incorporation and prior specification. Recent work has focused on encoding additional information such as multiple modalities (Serban et al., 2016; Suzuki et al., 2017; Wang et al., 2016; He et al., 2018) and how to integrate incomplete label knowledge in a semi-supervised fashion (Kingma et al., 2014). These approaches propose new model structures to incorporate additional information. However, in the case of representation learning, it is preferable to weakly supervise the representation, instead of changing the model structure. This can be achieved by constructing informed priors that guide the latent variables towards the desired representation. In this work, we describe how to construct informed multivariate Gaussian priors with full covariance matrices.

We exemplify our method on the problem of multi-label image annotation, i.e., we are presented with an image and a number of correlated descriptive terms. Our aim is to learn a latent representation that incorporates the semantic information of the annotations. In order to construct a prior, that encodes the annotations in a continuous space with semantic structure, we can make use of word embeddings, such as Word2Vec (Mikolov et al., 2013). Word embeddings have been shown to represent relationships between words in semantically coherent manner (Mikolov et al., 2013). By applying priors that represent the statistics of a given embedding, we hope to learn a structured latent space. Our method can be used to

annotate images, to retrieve images for a given set of words and to compute the similarity of images based on their semantic content.

2. Methodology

We introduce our framework which builds on variational autoencoders (VAEs) (Kingma and Welling, 2014). This model class assumes that an observation $\mathbf{x} \sim p(\mathbf{x} \,|\, \mathbf{z})$ depends on a latent variable $\mathbf{z} \sim p(\mathbf{z})$ which is often chosen to consist of independent Gaussian units.

Full Covariance Matrix. The independence assumption between the latent variables can lead to poor approximations when the true posterior $p(\mathbf{z} \,|\, \mathbf{x})$ comprises dependencies. A natural step is therefore to extend the vanilla VAE to utilize a full covariance matrix. As a covariance matrix Σ is always symmetric and positive semi-definite, the covariance matrix can be decomposed as $\Sigma = LL^T$ where L is a lower diagonal matrix. Given this decomposition, we can have the inference network infer a lower diagonal matrix L and apply the reparameterization trick.

$$\mathbf{z} = g(\epsilon, \mu(\mathbf{x}), \sigma(\mathbf{x})) = \mu(\mathbf{x}) + L(\mathbf{x})\,\epsilon\,. \tag{1}$$

A Semantic Prior. In order to encode structure in the latent space, for each observation $\mathbf{x}^{(i)}$ we require an informed prior $p(\mathbf{z}^{(i)}) \sim \mathcal{N}(\mu_{\mathbf{z}}^{(i)}, \Sigma_{\mathbf{z}}^{(i)})$. Specifically, we require that the mean and covariance encode semantic information about a given image.

In this work, we employ the semantic structure of word embeddings to construct structured priors. Naturally, images contain semantic meaning in form of objects, events and spatial relations. Here, we focus on object labels. Let an image $\mathbf{x}^{(i)}$ have a set of M_i labels $\mathbf{C}^{(i)} = \{c^{(i,1)}, c^{(i,2)}, ...c^{(i,M_i)}\}$ and let $\mathbf{W}^{(i)} = \{\mathbf{w}^{(i,1)}, \mathbf{w}^{(i,2)}, ... \mathbf{w}^{(i,M_i)}\}$ be the corresponding word vectors, where each vector $\mathbf{w}^{(i,j)}$ has the same dimensionality as the latent variable $\mathbf{z}^{(i)}$. We would like our model to associate all of these vectors with image $\mathbf{x}^{(i)}$ following the generative process:

$$p(\mathbf{x}^{(i)} \,|\, \mathbf{z}^{(i)}), \quad p(\mathbf{z}^{(i)} \,|\mu_{\mathbf{z}}^{(i)}, \Sigma_{\mathbf{z}}^{(i)}), \quad \mu_{\mathbf{z}}^{(i)} \sim Uniform(\mathbf{W}^{(i)}),$$

where the mean $\mu_{\mathbf{z}}^{(i)}$ is uniformly drawn from the set of word vectors and the covariance matrix $\Sigma_{\mathbf{z}}^{(i)}$ encodes the covariance structure in $\mathbf{W}^{(i)}$. An alternative to drawing $\mu_{\mathbf{z}}^{(i)}$ uniformly would be to set it equal to the mean word vector $\widetilde{\mathbf{w}}^{(i)}$. This would however impose additional assumptions on the decomposition of the data, e.g., that each object contributes equally to a scene.

Since most images are only associated with a few categories, the M_i word vectors might not suffice to compute a full-rank covariance matrix. Therefore, we make use of a standard result to counteract shrinkage of eigenvalues and set the final covariance matrix to $\Sigma_{\mathbf{z}}^{(i)} = (1-\lambda)\hat{\Sigma}_{\mathbf{z}}^{(i)} + \lambda\mathbb{I}$, where $\hat{\Sigma}_{\mathbf{z}}^{(i)} = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(\mathbf{w}^{(i,j)} - \widetilde{\mathbf{w}}^{(i)})(\mathbf{w}^{(i,j)} - \widetilde{\mathbf{w}}^{(i)})^T$ and $0 \leq \lambda \leq 1$ is a parameter controlling the influence of the matrices.

A limiting factor of using this prior is that we need to invert $\Sigma_{\mathbf{z}}^{(i)}$ for every data point to compute the KL divergence between the approximate distribution and the prior. However, this is feasible as it only needs to be done once for each dataset, and the dimensionality of the latent space is commonly small.

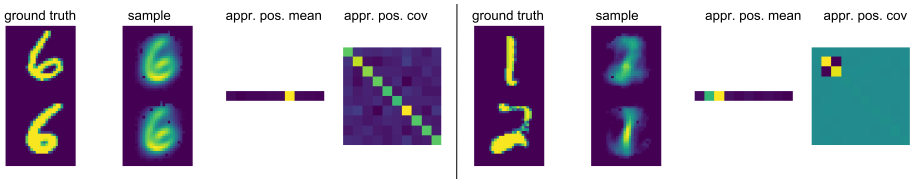


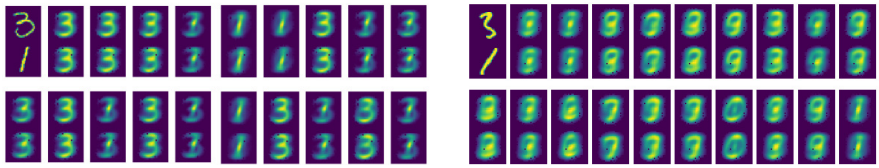
Figure 1: Examples of ground truth, a sample from $p_\theta(\mathbf{x} \mid \mathbf{z})$, the approximate posterior mean $\mu(\mathbf{x})$ and covariance matrix $\Sigma(\mathbf{x})$. The model was trained with $\lambda = 0.01$.

3. Experiments

In this section we demonstrate the mechanisms of our proposed method. We evaluate on a toy example that contains random pairs of handwritten digits from the MNIST. We chose one-hot encodings for the digits as the embedding of the annotation. On this toy dataset, we present the learned posterior structure and the samples of the model given an image. We also show how to classify, retrieve and compute the distance between images based on our model. Secondly, we evaluate the classification capability of our approach on the CIFAR-100 (Krizhevsky and Hinton, 2009) and ImageNet (Krizhevsky et al., 2012) datasets for which we use word2vec embeddings.

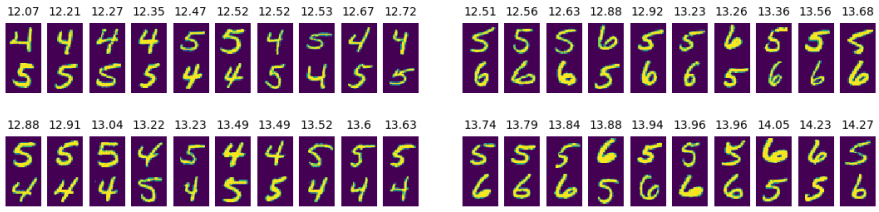
Classification: To classify an image, we can compute the likelihood of all word vectors under the posterior and rank them. When concerned with the ordinary MNIST, the classification accuracy achieves 99.5 % using the one-hot encoding. On the test set of the Paired MNIST dataset, the two most likely numbers are taken to be the determined classes. Here, we achieve a classification rate of 95.5 %. In the case of CIFAR-100, the classification based on a word2vec encoding (the proposed informed embedding) achieves an accuracy of 42.5 %, while a one-hot encoding achieves an accuracy of 12.5 %. Note that we do not perform dimensionality reduction on the word2vec embedding. On ImageNet, the classification accuracy is 27.1 % based on word2vec embeddings and 8.3 % based on one-hot encodings.

The effect of λ: To investigate the effect of λ, we show samples of a network trained on the Paired MNIST dataset with $\lambda = 0.01$ in Figure 2 a) and $\lambda = 0.99$ in Figure 2 b). While the samples in Figure 2 a) seem to follow an interpolation between *one* and *three*, the samples in Figure 2 b) can rarely be identified as *one* or *three* but resemble other numbers. Since a covariance matrix close to identity, i.e. when $\lambda = 0.99$, will sample each dimension from an independent normal distribution, the samples can end up anywhere in our constructed one-hot latent space.



a) b)

Figure 2: Samples from $p_\theta(\mathbf{x} \mid \mathbf{z})$. Trained with **a)** $\lambda = 0.01$ and **b)** $\lambda = 0.99$.



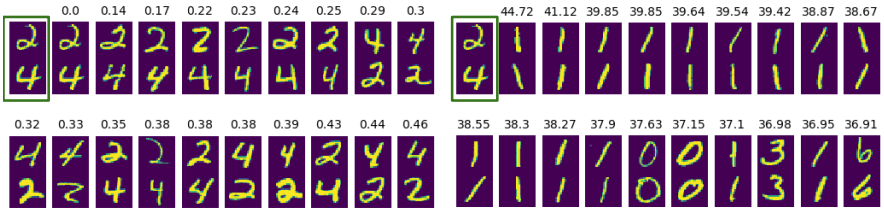
a) b)

Figure 3: Retrieved images for the combinations **a)** $4-5$ and **b)** $5-6$. The KL divergence between the prior and the posterior is marked in the title above each image.

Image retrieval: In order to retrieve images for a given set of word embeddings $\mathbf{W}^{(k)}$, we compute their mean $\mu^{(k)}$ and covariance matrix $\Sigma^{(k)}$. These parameters constitute the query distribution $p(\mathbf{z}\,|\mathbf{W}^{(k)})$. Furthermore, for all images $i \in [1,..N]$ in our database, we compute the parameters $\mu(\mathbf{x}^{(i)})$ and $\Sigma(\mathbf{x}^{(i)})$, forming the approximate posterior distribution $q_\Phi(\mathbf{z}\,|\,\mathbf{x}^{(i)})$. To retrieve R images, we compute the KL divergence $D_{KL}(p(\mathbf{z}\,|\mathbf{W}^{(k)}), q_\Phi(\mathbf{z}\,|\,\mathbf{x}^{(i)}))$ for all images and select the R images with the smallest KL divergence values.

Results for the number pairs $4-5$ and $5-6$ from the Paired MNIST dataset are shown in Figure 3. The images depict the 20 images with the lowest KL values. The retrieved images contain only the query numbers, in any order.

Image similarity: A third application of our method is to determine the similarity between images. To this end, for any two images i and j, compute the approximate posterior parameters $(\mu(\mathbf{x}^{(i)}), \Sigma(\mathbf{x}^{(i)}))$ and $(\mu(\mathbf{x}^{(j)}), \Sigma(\mathbf{x}^{(j)}))$ and calculate the KL divergence between these two distributions. In Figure 4 a) we show the closest images to the image on the top left. Only the same numbers are displayed, irrespective of their order. Thus, the KL divergence acts here a measure of semantic similarity. On the other side, in Figure 4 b), we depict the images with the highest KL divergence. All of these are pairs of identical numbers and contain mostly 0 and 1. These numbers are well distinguishable from other numbers, which means that there posterior will be distinct from the posterior of the query image.



a) b)

Figure 4: Retrieving the most **a)** similar and the most **b)** dissimilar images for the query.

References

Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. *European Conference on Computer Vision*, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, 2014.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Iulian V Serban, II Ororbia, G Alexander, Joelle Pineau, and Aaron Courville. Multi-modal variational encoder-decoders. *arXiv preprint arXiv:1612.00377*, 2016.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *ICLR Workshop*, 2017.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *arXiv preprint arXiv:1711.05597*, 2017.