A Sparsity-promoting Dictionary Model for Variational Autoencoders

Mostafa Sadeghi and Paul Magron

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

INTERSPEECH'22

Poster number: 237

Overview

- Structuring the latent space in variational autoencoders (VAEs) is addressed.
- A simple yet effective sparsity-promoting dictionary model is presented for the latent variables (SDM-VAE).
- The proposed methodology is tuning-free, relying on a zero-mean Gaussian latent prior distribution with learnable variances.
- Experiments on speech generative modeling demonstrate the advantage of the proposed approach over competing techniques.

Unsupervised representation learning

- Automatically extracting useful information from unlabeled data, in the form of a *feature* or *representation* vector.
- Could be useful for various downstream tasks or for learning distributions of data, e.g. using variational autoencoders (VAEs).
- To learn more efficient and interpretable representations, it's important to somehow impose some meaningful structures.
- One such powerful constraint is obtained through a *sparsity* assumption.

We focus on sparsity promoting latent models for VAEs.

Background on VAEs

Generative modeling:

1. Let $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ denote a set of training data with $\mathbf{s}_i \in \mathbb{R}^n$.
2. Define the latent variables $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ with $\mathbf{z}_i \in \mathbb{R}^m$, $m \ll n$.
3. Model the joint distribution $p(\mathbf{s}, \mathbf{z}) = p(\mathbf{s}|\mathbf{z}) \cdot p(\mathbf{z})$ as follows:

$$\begin{cases} p_\theta(\mathbf{s}|\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{z}), \mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{z}))), \\ p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{cases} \tag{1}$$

- $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- $\mathbf{I}$: the identity matrix of appropriate size.
- $\boldsymbol{\mu}_\theta(.)$ and $\boldsymbol{\sigma}_\theta(.)$: non-linear functions implemented via some deep neural networks (DNNs) with parameters θ.

Parameter estimation:

To learn θ, the intractable posterior $p_\theta(\mathbf{z}|\mathbf{s})$ is approximated with a parametric Gaussian distribution, which is called the encoder [1]:

$$q_\psi(\mathbf{z}|\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_\psi(\mathbf{s}), \mathrm{diag}(\boldsymbol{\sigma}^2_\psi(\mathbf{s}))), \tag{2}$$

with $\boldsymbol{\mu}_\psi$ and $\boldsymbol{\sigma}_\psi$ implemented by some DNNs with parameters ψ.

Using variational inference, maximize a lower bound of $\ln p_{\boldsymbol{\theta}}(\mathbf{s})$:

$$\mathcal{L}(\Phi; \mathbf{s}) = \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{s})}[\log p_\theta(\mathbf{s}|\mathbf{z})] - \mathcal{D}_{\mathrm{KL}}(q_\psi(\mathbf{z}|\mathbf{s}) \| p(\mathbf{z})), \tag{3}$$

where $\Phi = \{\theta, \psi\}$, and $D_{\mathrm{KL}}(. \parallel .)$ is the Kullback–Leibler divergence.

- Optimize $\mathcal{L}(\Phi; \mathbf{s})$ over Φ using a gradient-based solver together with the application of the reparametrization trick.

Structuring the latent space

The standard normal prior distribution is not effective, as it might not efficiently capture the underlying distribution of the data. Several alternative distributions have been introduced that promote sparsity in the latent space (increasing interpretability, reducing the risk of overfitting):

- Mixture of two Gaussian distributions with one having a very small variance [2].
- A Spike-and-Slab distribution for the prior & encoder [3].
- A *deterministic* dimension selector function that deactivates some dimensions of the latent vector [4].

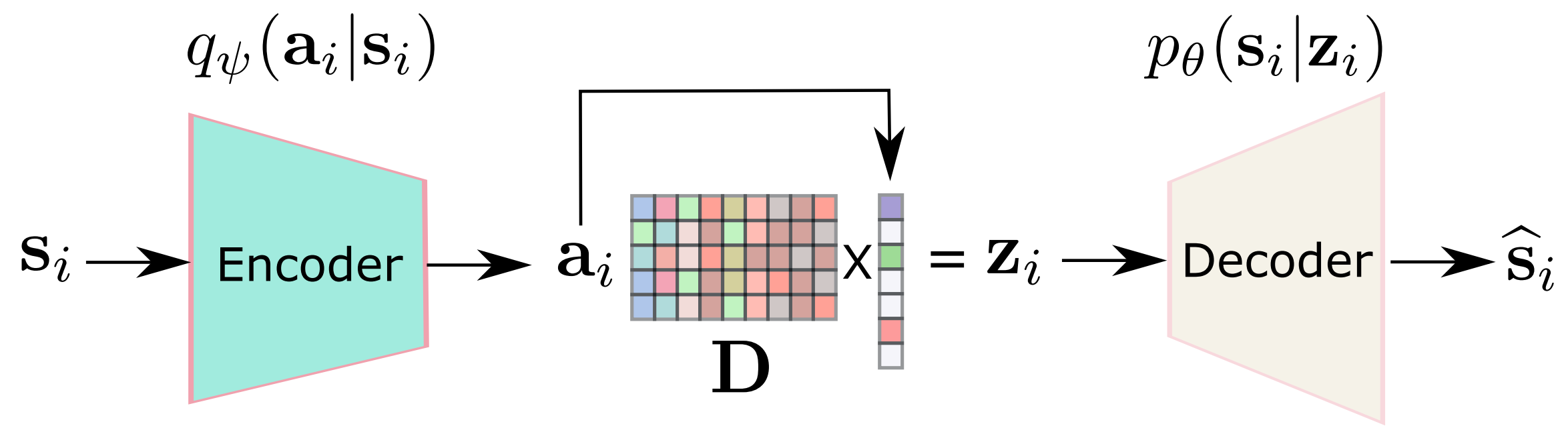
These approaches though being effective, they often involve several user-defined hyperparameters and non-Gaussian distributions or discrete latent variables, which complicates the learning process.

Proposed methodology: SDM-VAE

Main idea: A sparse dictionary model for each latent code $\mathbf{z}_i$ as follows:

$$\mathbf{z}_i = \mathbf{D}\mathbf{a}_i, \quad \forall i, \tag{4}$$

- $\mathbf{D} \in \mathbb{R}^{m \times k}$: a (potentially overcomplete) dictionary with unit-norm columns.
- $\mathbf{a}_i \in \mathbb{R}^k$: a sparse representation vector.



☞ Keeping the reconstruction quality intact, while promoting interpretability.

Sparsity-promoting prior:

$$p(\mathbf{a}_i; \boldsymbol{\gamma}_i) = \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\gamma}_i)). \tag{5}$$

- $\boldsymbol{\gamma}_i$: a vector of learnable variances.

With an inverse Gamma hyperprior on $\boldsymbol{\gamma}_i$ in a Bayesian setting, the prior

$$p(\mathbf{a}_i) = \int p(\mathbf{a}_i|\boldsymbol{\gamma}_i) \cdot p(\boldsymbol{\gamma}_i) \mathrm{d}\boldsymbol{\gamma}_i, \tag{6}$$

becomes a Student's t-distribution, which promotes sparsity due to its sharp peak at zero.

Inference

$$\mathcal{L}(\Phi, \boldsymbol{\gamma}; \mathbf{s}) = \mathbb{E}_{q_\psi(\mathbf{a}|\mathbf{s})}[\log p_\theta(\mathbf{s}|\mathbf{z})] - \mathcal{D}_{\mathrm{KL}}(q_\psi(\mathbf{a}|\mathbf{s}) \| p(\mathbf{a}; \boldsymbol{\gamma})), \quad \text{with} \quad \mathbf{z}_i = \mathbf{D}\mathbf{a}_i \quad \forall i.$$

☞ Alternating minimization approach to update Φ and $\boldsymbol{\gamma}$:

1. Update $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} \leftarrow \underset{\boldsymbol{\gamma}}{\mathrm{argmin}} \ \mathcal{D}_{\mathrm{KL}}(q_\psi(\mathbf{a}|\mathbf{s}) \| p(\mathbf{a}; \boldsymbol{\gamma})) \tag{7}$$

2. Update Φ:

$$\Phi \leftarrow \arg\max_{\Phi} \ \mathcal{L}(\Phi, \boldsymbol{\gamma}; \mathbf{s}) \tag{8}$$

For $\boldsymbol{\gamma}_i$, we obtain the following closed-form solution:

$$\boldsymbol{\gamma}_i = \mathbb{E}_{q_\psi(\mathbf{a}_i|\mathbf{s})}[\mathbf{a}_i^2] = \mu^2_\psi(\mathbf{s}_i) + \sigma^2_\psi(\mathbf{s}_i), \quad \forall i. \tag{9}$$

For Φ, we follow standard VAEs by approximating the expectation with a single sample & reparametrization: $\mathbf{a}_i = \boldsymbol{\mu}_\psi(\mathbf{s}_i) + \boldsymbol{\sigma}_\psi(\mathbf{s}_i) \odot \boldsymbol{\epsilon}_i$ then $\mathbf{z}_i = \mathbf{D}\mathbf{a}_i$ ($\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\odot$: element-wise multiplication.). Then optimize

$$\hat{\mathcal{L}}(\Phi, \boldsymbol{\gamma}; \mathbf{s}) = \log p_\theta(\mathbf{s}|\mathbf{z}) - \mathcal{D}_{\mathrm{KL}}(q_\psi(\mathbf{a}|\mathbf{s}) \| p(\mathbf{a}; \boldsymbol{\gamma})). \tag{10}$$

Application to speech modeling

Speech analysis-resynthesis:

1. Short-time Fourier transform (STFT) of speech waveform → complex-valued data $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ ($\mathbf{x}_i \in \mathbb{C}^F$).
2. Auto-encode power spectrogram $\mathbf{s} = |\mathbf{x}|^2$ by the trained VAE.
3. Reconstruct original speech waveform using estimated power spectrogram and the original phase.
4. Evaluate the reconstruction quality.

☞ A discrete cosine transform (DCT) matrix for $\mathbf{D}$ in SDM-VAE.

Experiments

- **Corpus**: TCD-TIMIT [5]: 56 English speakers (39 training, 8 validation, 9 test), 98 sentences (∼ 5 s) per speaker.
- **STFT parameters**: 64 ms sine window, 75% overlap → STFT frames of length $n = 513$.
- **Baselines**: Standard VAE [6], and variational sparse coding (VSC) [3].
- **Model architecture**: All the VAE models follow a simple architecture with a single hidden layer in both encoder & decoder [6].

Performance measures: Perceptual evaluation of speech quality (**PESQ**) [-0.5,4.5], Short-time objective intelligibility (**STOI**) [0,1], **Hoyer** metric [0,1] for sparsity.

Table 1:Reconstruction quality and sparsity measure for various VAE-based methods in terms of PESQ, STOI, and Hoyer scores.

Dimension of $\mathbf{z}$		$m = 32$			$m = 64$		
		PESQ	STOI	Hoyer	PESQ	STOI	Hoyer
VAE		3.29	0.85	0.40	3.26	0.85	0.56
VSC	$\alpha = 0.05$	3.00	0.81	0.57	3.25	0.84	0.51
	$\alpha = 0.5$	3.25	0.84	0.54	3.32	0.85	0.65
	$\alpha = 0.9$	3.25	0.84	0.47	3.26	0.85	0.60
SDM-VAE	$\mathbf{I}$	3.33	0.86	0.64	**3.45**	**0.87**	0.73
	DCT ($k = 32$)	3.37	0.86	0.66	3.28	0.84	0.66
	DCT ($k = 64$)	3.32	0.86	**0.87**	3.33	0.86	0.76

Conclusions:

▷ Advantage of sparsity as a regularizer for the model to avoid overfitting.
▷ Increasing the number of atoms in $\mathbf{D}$ improves sparsity.
▷ SDM-VAE exhibits a better and more stable performance than VSC.

References

[1] D. P. Kingma and M. Welling, "*Auto-encoding variational Bayes*," ICLR, 2014.

[2] E. Mathieu *et al.*, "Disentangling disentanglement in variational autoencoders," in Proc. International Conference on Machine Learning (ICML), June 2019.

[3] F. Tonolini *et al.*, "Variational sparse coding," in Proc. conference on Uncertainty in Artificial Intelligence (UAI), August 2020.

[4] N. Miao *et al.*,"On incorporating inductive biases into VAEs," in Proc. International Conference on Learning Representations (ICLR), May2021.

[5] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," IEEE Transactions on Multimedia, vol. 17,no. 5, pp. 603–615, 2015.

[6] S. Leglaive *et al.*, "*A variance modeling framework based on variational autoencoders for speech enhancement*," in Proc. MLSP, 2018.