29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021

Unsupervised Blind Source Separation with Variational Auto-Encoders - Slides

Julian Neri[1] Philippe Depalle[1] Roland Badeau[2]

[1]McGill University, CIRMMT, Montréal, Canada
[2]LTCI, Télécom Paris, Institut Polytechnique de Paris, France

August 26, 2021

McGill CIRMMT NSERC CRSNG TELECOM ParisTech

Overview

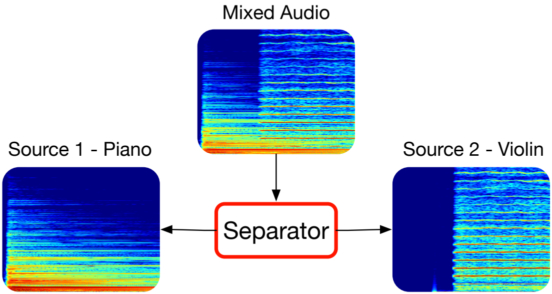
Source code and data are available on my website,

`http://www.music.mcgill.ca/~julian/vae-bss`

Introduction

Unsupervised blind source separation: estimate the underlying sources from a single-channel mixture signal

- without knowing the true number of sources, and
- without clean target source signals for model training.



Applications: re-mixing, hearing aids, dictation, computer vision.

- Ex: Interact with an audio recording in real-time.

Introduction

Supervised

Requires the true underlying source files

- Mature enough to separate real musical recordings consisting of vocals, drum, bass, piano.

Unsupervised

Learns solely from the mixtures

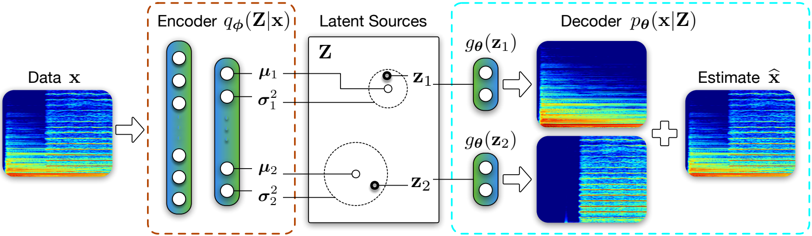
- Harder than supervised, requiring strong prior information:
  - Make assumptions about (model) the sources.
  - Learn prior source information from data.

Examples: NMF [Fevotte et al., 2009], DNN [Halperin et al., 2019], [Wisdom et al., 2020].

Methodology - Overview

We address the problem with a new variational auto-encoder (VAE) [Kingma and Welling, 2014, Neri et al., 2021].

- **Encoder** disentangles (separates) latent sources
- **Decoder** independently generates each source signal



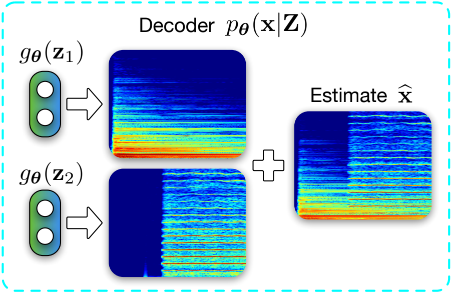
Methodology - VAE Decoder (Generative Network)

Gaussian Prior $$p(\mathbf{Z}) = \prod_{k=1}^{K} p(\mathbf{z}_k) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{z}_k|\mathbf{0}, \mathbf{I}\right)$$

K = number of sources

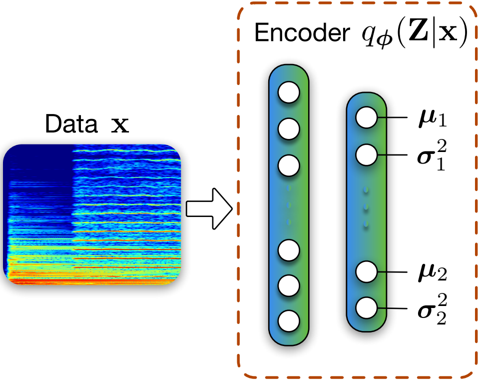
$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_K \end{bmatrix}$$

$$\widehat{\mathbf{x}} = \sum_{k=1}^{K} \widehat{\mathbf{s}}_k = \sum_{k=1}^{K} g_{\boldsymbol{\theta}}(\mathbf{z}_k)$$



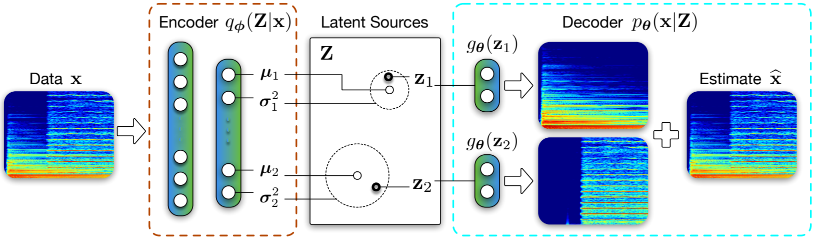
Laplace Likelihood $$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}) = \prod_{i=1}^{D_x} \frac{1}{2b} \exp\left(-\frac{|x_i - \widehat{x}_i|}{b}\right)$$

Methodology - VAE Encoder (Inference Network)



Approximate Posterior $q_\phi(\mathbf{Z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{Z}|\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x})\mathbf{I}\right)$

Methodology - VAE Training



$$\mathcal{L}(\boldsymbol\theta, \phi; \mathbf{X}) = \sum_{n=1}^{N} \underbrace{\langle \ln p_{\boldsymbol\theta}(\mathbf{x}^{(n)}|\mathbf{Z})\rangle_{q_\phi}}_{\text{Expected log-likelihood}} - \underbrace{D_{KL}\left(q_\phi(\mathbf{Z}|\mathbf{x}^{(n)})\|p(\mathbf{Z})\right)}_{\text{KL divergence}}$$

Expected log-likelihood: reconstruction error between $\mathbf{x}$ and $\widehat{\mathbf{x}}$
KL divergence: regulates posterior, helps source disentanglement

Architecture

Table: Encoder input, hidden, and output layer units (dimensions). MNIST: handwritten digit images. MUMS: audio spectrograms.

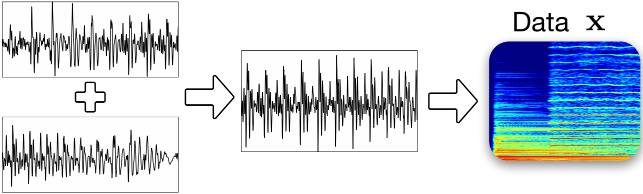
Dataset	Input D_x	Hidden Layers L1	L2	L3	L4	L5	Output $2 \times D_Z$
MNIST	784	700	600	500	400	300	$2 \times 20K$
MUMS	32768	2560	2048	1536	1024	512	$2 \times 64K$

VAE Mask (VAEM) After model training (at run-time), makes sum of sources exactly equal to the input mixture.

$$\check{\mathbf{s}}_k = \widehat{\mathbf{s}}_k \odot (\mathbf{x} \oslash \widehat{\mathbf{x}})$$

Evaluation - Datasets

- Data was generated by randomly sampling and adding source signals.
- Sources were never mixed between testing and training datasets.
- For audio, each mixture was transformed into a spectrogram.
- Data normalized to values between 0 and 1.



MNIST : 32 $\times$ 32 images of handwritten digits
MUMS : 256 $\times$ 128 spectrograms of audio (5.5 kHz $\times$ 1.5 s)

Evaluation - Baseline Methods, and Ideal Masks

Baseline / State-of-the art

- NMF - nonnegative matrix factorization [Fevotte et al., 2009]
- GLO - generative latent optimization [Halperin et al., 2019]
- MixIT - mixture invariant training [Wisdom et al., 2020]

Ideal masks (for audio)

- IRM - ideal ratio mask [Stöter et al., 2018]
- IBM - ideal binary mask [Stöter et al., 2018]

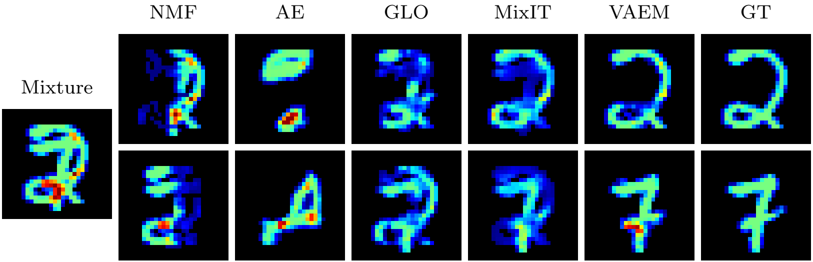
Ours

- AE - auto-encoder
- VAE - variational auto-encoder
- VAEM - variational auto-encoder with masking

Evaluation - MNIST Handwritten Digits

Two Model Sources

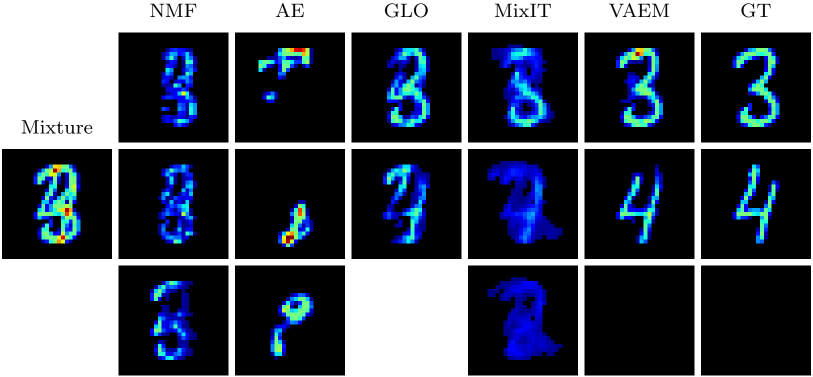
	NMF	AE	GLO	MixIT	VAEM
PSNR	17.22	15.96	24.63	25.69	**26.69**
SSIM	0.50	0.43	0.86	0.88	**0.93**



Evaluation - MNIST Handwritten Digits

Three Model Sources

	NMF	AE	GLO	MixIT	VAEM
PSNR	18.27	16.30	n/a	25.41	**27.68**
SSIM	0.44	0.43	n/a	0.69	**0.94**



Evaluation - MNIST Handwritten Digits

Three Model Sources

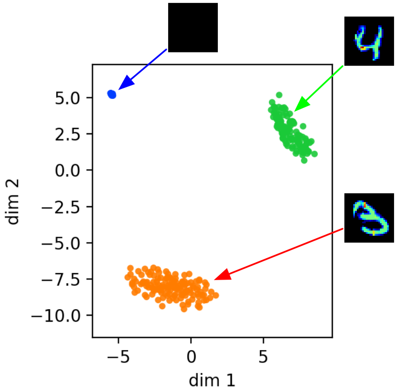


Figure: Learned Latent Space Mapped to 2D.

Evaluation - MUMS Audio Spectrograms

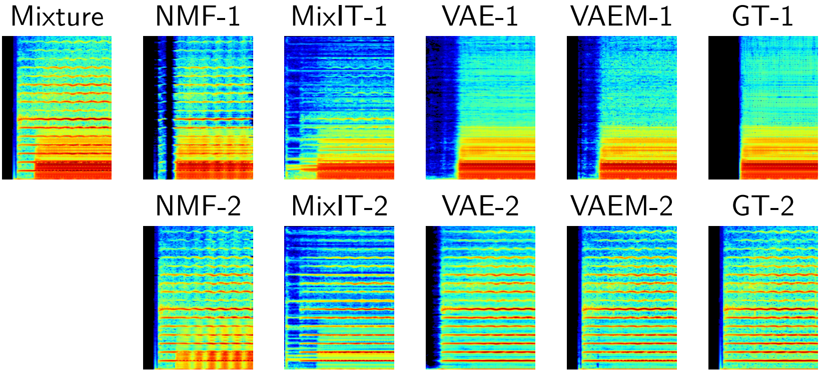
	NMF	MixIT	VAE	VAEM	IBM	IRM
SI-SDR	5.40	6.64	14.33	**17.10**	23.97	22.66
SIR	16.93	17.59	**29.92**	29.55	48.89	34.39
SAR	7.96	8.26	14.87	**18.20**	24.06	23.23

Evaluated with bss eval measures [Vincent et al., 2006] [Roux et al., 2019]:

- SI-SDR: scale-invariant signal-to-distortion ratio
- SIR: signal-to-interference ratio
- SAR: signal-to-artifact ratio

Evaluation - MUMS Audio Spectrograms

Mixture of Bassoon (GT-1) & Violin (GT-2)



Listen to more examples at http://www.music.mcgill.ca/~julian/vae-bss

Conclusion

VAE framework offers a viable solution to unsupervised blind separation of both image and audio mixtures.

- Unlike previous methods, it automatically avoids over-separation.
- Generating sources independently using the same decoder is memory efficient and crucial for separation.
- Disentangling low-dimensional encoded sources is practically effective and aligns with human perception.

→ **Future research**: model temporal evolution of sources to separate waveforms and videos of arbitrary duration.

References I

Fevotte, C., Bertin, N., and Durrieu, J. (2009).
Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis.
Neural Computation, 21:793–830.

Halperin, T., Ephrat, A., and Hoshen, Y. (2019).
Neural separation of observed and unobserved distributions.
In *Proc. 36th Int. Conf. Machine Learning (ICML)*.

Kingma, D. P. and Welling, M. (2014).
Auto-encoding variational Bayes.
In *Int. Conf. Learning Representations (ICLR)*.

Neri, J., Badeau, R., and Depalle, P. (2021).
Unsupervised blind source separation with variational auto-encoders.
In *Proc. 29th European Signal Processing Confernce (EUSIPCO)*.

References II

Roux, J. L., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019).
SDR – half-baked or well done?
In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 626–630.

Stöter, F.-R., Liutkus, A., and Ito, N. (2018).
The 2018 signal separation evaluation campaign.
In *Latent Variable Analysis and Signal Separation*, pages 293–305. Springer Int. Publishing.

Vincent, E., Gribonval, R., and Févotte, C. (2006).
Performance measurement in blind audio source separation.
IEEE Trans. Audio, Speech, Lang. Process., 14(4):1462–1469.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, J., Wilson, K., and Hershey, J. (2020).
Unsupervised sound separation using mixture invariant training.
Advances in Neural Information Processing Systems.