Unsupervised music source separation with deep generative priors

March 8, 2020

Contents

1 Proposal **1**
1.1 Abstract 1
1.2 Research Question 1
1.3 Related works 1
1.3.1 Deep Latent-Variable Models 1
1.3.2 Sound 3
1.3.3 Source separation 3
1.4 Methodology 3
1.4.1 Datasets 3
1.5 Planning 3

Chapter 1

Proposal

1.1 Abstract

1.2 Research Question

Source separation is the task of finding a set of latent sources $\boldsymbol{s} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n]$ to an observed mix of those sources $\boldsymbol{m}$. The induced model proposes a mixing function $\boldsymbol{m} = f(\boldsymbol{s})$ which might just be a liner mixing $\boldsymbol{m} = \boldsymbol{A} \cdot \boldsymbol{s}$. The task is to find the inverse model $f^{-1}(\cdot)$ which retrieves $\boldsymbol{s}$.

> Can we learn an sound source separation model in an unsupervised manner. Unsupervised relating to missing pairing of sources to mixes.

$$p(s'|m) \cdot p(s) \tag{1.1}$$

1.3 Related works

In this chapter we discuss previous research in supervised and semi-supervised source separation.

1.3.1 Deep Latent-Variable Models

amortized inference [7]

[15]

We have an observed set of data $\boldsymbol{x} \in \mathcal{D}$ for which there exists an unknown data probability distribution $p^*(\boldsymbol{x})$. In our directed graphical model we introduce an approximate model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ with model parameters $\boldsymbol{\theta}$. Learning now means finding the values for $\boldsymbol{\theta}$ that give the closest approximations of the true underlying process:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx p^*(\boldsymbol{x}) \tag{1.2}$$

The model $p_{\boldsymbol{\theta}}$ has to be complex enough to be able to fit the data distribution while being simple enough to be learnable. Every model comes with *inductive biases* making a replication of the data distribution impossible.

In the following described models we assume the sampled data points to be *independent and identically distributed* samples drawn from $\mathcal{D}$. Therefore we can write the data log-likelihood as:

$$\log p_{\boldsymbol{\theta}}(\mathcal{D}) = \sum_{\boldsymbol{x} \in \mathcal{D}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{1.3}$$

The maximum likelihood estimation of our model parameters maximizes this criterion.

Latent-variable models we introduce *latent variables*. Latent variables are part of the directed graphical model but not observed.

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} \tag{1.4}$$

Following the *variation principle* [10] we introduce the *inference model* $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \log \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \tag{1.5}$$

$$= \log \int \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \tag{1.6}$$

$$\geq -KL[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})] \tag{1.7}$$

The VAE framework

VAE [14][22]

β-VAE [8] - intoduces βas controlling hyperparameter in the VAE objective - constraint that controls the capacityof the latent space - gives trade off between reconstruction quality and representation simplicity - similar to information bottleneck [1]

VQ-VAE [27]

Flow based models

$$\mathbf{z} \sim p_{\boldsymbol{Z}}(\boldsymbol{z}) \tag{1.8}$$

$$\mathbf{x} = f^{-1}(\boldsymbol{z}) \tag{1.9}$$

change of variable

NICE [4] - coupling layer - triangular shape

Normalizing Flow [21]
RealNVP [5]
Glow [13] - invertible 1x1 convs - ActNorm - zero init

[25] introduced WaveNet an autoregressive generative model for raw (*time-domain*) audio. WaveNet closely similar to the earlier PixelCNN [26] but adapted for the audio domain. Unomoidified Cnns are unsuitable to the application to raw audio because of the form of data. as digital audio is ampled at a extremely high sample rate commonly 16kHz up to 44kHz the features of interest lie at scale of stringly different magnitudes. On the one hand recognizing phase, frequency of a wave might require features at those ms scales on the other hand the modelling of speech or music audio happens at the scale of seconds or minutes. As such a generative model for this domain has to cpature those different time sclaes. The wvaenet accomplishes this by using dilated convolutions a common tool in signal processing [6]. A dilated convolutions uses a kernel with an inner stride. Using a stack of dialted convolutions increases the recpetive field of the features without increasing the comutional complexity.

- gated convs -pixelcnn -lstm[9] - dilated convs - global conditioning - -law encoding [19] - slow cause autoreg (better with [17]) -

PixelCNN++ [23]

1.3.2 Sound

NSynth [11]
In [18]
FloWaveNet [12]

1.3.3 Source separation

WaveNet for Speech denoising[20]
WaveNet-VAE unsupervised speech rep learning[2]
Wave-U-Net[24]
DeMucs[3]
Source Sep in Time Domain[16]

1.4 Methodology

1.4.1 Datasets

ToyData

MusDB

1.5 Planning

Bibliography

[1] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in Beta-VAE," 04/10/2018. arXiv: 1804.03599 [cs, stat] (cit. on p. 2).

[2] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2041–2053, 12/2019. arXiv: 1901.08810 (cit. on p. 3).

[3] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed," 09/03/2019. arXiv: 1909.01174 [cs, eess, stat] (cit. on p. 3).

[4] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear Independent Components Estimation," 04/10/2015. arXiv: 1410.8516 [cs] (cit. on p. 2).

[5] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP," 02/27/2017. arXiv: 1605.08803 [cs, stat] (cit. on p. 3).

[6] P. Dutilleux, "An Implementation of the algorithme à trous to Compute the Wavelet Transform," in *Wavelets*, J.-M. Combes, A. Grossmann, and P. Tchamitchian, Eds., ser. Inverse Problems and Theoretical Imaging, Berlin, Heidelberg: Springer, 1990, pp. 298–304 (cit. on p. 3).

[7] S. Gershman and N. Goodman, "Amortized inference in probabilistic reasoning," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 36, 2014 (cit. on p. 1).

[8] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," 11/04/2016 (cit. on p. 2).

[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11/01/1997 (cit. on p. 3).

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 11/01/1999 (cit. on p. 2).

[11] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," 02/23/2018. arXiv: 1802.08435 [cs, eess] (cit. on p. 3).

[12] S. Kim, S.-g. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet : A Generative Flow for Raw Audio," 05/20/2019. arXiv: 1811.02155 [cs, eess] (cit. on p. 3).

[13] D. P. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," 07/10/2018. arXiv: 1807.03039 [cs, stat] (cit. on p. 3).

[14] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," 05/01/2014. arXiv: 1312.6114 [cs, stat] (cit. on p. 2).

[15] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," 07/24/2019. arXiv: 1906.02691 [cs, stat] (cit. on p. 1).

[16] F. Lluís, J. Pons, and X. Serra, "End-to-end music source separation: Is it possible in the waveform domain?," 06/28/2019. arXiv: 1810.12187 [cs, eess] (cit. on p. 3).

[17] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang, "Fast Wavenet Generation Algorithm," 11/28/2016. arXiv: 1611.09482 [cs] (cit. on p. 3).

[18] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," 10/30/2018. arXiv: 1811.00002 [cs, eess, stat] (cit. on p. 3).

[19] *Recommendation G. 711. Pulse Code Modulation (PCM) of voice frequencies*, 1988 (cit. on p. 3).

[20] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," 01/31/2018. arXiv: 1706.07162 [cs] (cit. on p. 3).

[21] D. J. Rezende and S. Mohamed, "Variational Inference with Normalizing Flows," 06/14/2016. arXiv: 1505.05770 [cs, stat] (cit. on p. 3).

[22] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic Backpropagation and Approximate Inference in Deep Generative Models," 05/30/2014. arXiv: 1401.4082 [cs, stat] (cit. on p. 2).

[23] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications," 01/19/2017. arXiv: 1701.05517 [cs, stat] (cit. on p. 3).

[24] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," (Paris, France), 09/23/2018 (cit. on p. 3).

[25] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," 09/12/2016. arXiv: 1609.03499 [cs] (cit. on p. 3).

[26] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional Image Generation with PixelCNN Decoders," 06/16/2016. arXiv: 1606.05328 [cs] (cit. on p. 3).

[27] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 6306–6315 (cit. on p. 2).