Maurice Frank

Unsupervised music source separation with deep generative priors

Contents

Proposal

Abstract

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.

$$p(s'|m) \cdot p(s) \tag{1}$$

Related works

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

Deep Latent-Variable Models

We have an observed set of data $\boldsymbol{x} \in \mathcal{D}$ for which there exists an unknown data probability distribution $p^*(\mathcal{D})$. We introduce an approximate model with density[1] $p_{\boldsymbol{\theta}}(\mathcal{D})$ and model parameters $\boldsymbol{\theta}$. Learning or modelling means finding the values for $\boldsymbol{\theta}$ which will give the closest approximation of the true underlying process:

[1] We write density and distribution interchangeably to denote a probability function.

$$p_{\boldsymbol{\theta}}(\mathcal{D}) \approx p^*(\mathcal{D}) \tag{2}$$

The model $p_{\boldsymbol{\theta}}$ has to be complex enough to be able to fit the data density while little enough parameters to be learnable. Every choice for the form of the model comes will *induce* biases[2] about what density we can model.

[2] called *inductive biases*

In the following described models we assume the sampled data points $\boldsymbol{x}$ to be drawn from $\mathcal{D}$ *independent and identically distributed*[3]. Therefore we can write the data log-likelihood as:

[3] meaning the sample of one datum does not depend on the other data points

$$\log p_{\boldsymbol{\theta}}(\mathcal{D}) = \sum_{\boldsymbol{x} \in \mathcal{D}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{3}$$

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

To form a latent-variable model we introduce *latent variable*[4]. The data likelihood now is the marginal density of the joint latent density:

[4] Latent variables are part of the directed graphical model but not observed.

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} \tag{4}$$

Typically we introduce a factorization of the joint. Most commonly:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \tag{5}$$

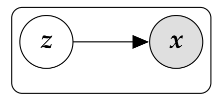

Figure 1: The graphical model with the simple introduced latent variable $\boldsymbol{z}$. Observed variables are shaded.

This corresponds to the graphical model in which $\boldsymbol{z}$ is generative parent node of the observed $\boldsymbol{x}$, see fig. 1.

If the latent is small, discrete it might be possible to directly marginalize over it. In this case

Following the *variation principle*[5] we introduce the *inference model* $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$.

[5] Michael I. Jordan et al. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233.

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \log \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} && (6)\\
&= \log \int \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} && (7)\\
&\geq - \mathrm{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})] && (8)
\end{aligned}$$

The VAE framework

VAE[6,7]

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[9]

VQ-VAE[10]

[6] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (2014). arXiv: 1312.6114 [cs, stat].

[7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: (2014). arXiv: 1401.4082 [cs, stat].

[8] Irina Higgins et al. "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: (2016).

[9] Christopher P. Burgess et al. "Understanding Disentangling in Beta-VAE". In: (2018). arXiv: 1804.03599 [cs, stat].

[10] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. "Neural Discrete Representation Learning". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 6306–6315.

Flow based models

$$\begin{aligned}
\mathbf{z} &\sim p_Z(z) && (9)\\
\mathbf{x} &= f^{-1}(z) && (10)
\end{aligned}$$

change of variable

NICE[11] - coupling layer - triangular shape

Normalizing Flow[12]

RealNVP[13]

Glow[14] - invertible 1x1 convs - ActNorm - zero init

[15] introduced WaveNet an autoregressive generative model for raw (*time-domain*) audio. WaveNet closely similar to the earlier PixelCNN[16] 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.[17] 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[18] - dilated convs - global conditioning - -law encoding[19] - slow cause autoreg (better with[20]) -

PixelCNN++[21]

Sound

NSynth[22]

In[23]

FloWaveNet[24]

Source separation

WaveNet for Speech denoising[25]

WaveNet-VAE unsupervised speech rep learning[26]

Wave-U-Net[27]

DeMucs[28]

Source Sep in Time Domain[29]

[11] Laurent Dinh, David Krueger, and Yoshua Bengio. "NICE: Non-Linear Independent Components Estimation". In: (2015). arXiv: 1410.8516 [cs].

[12] Danilo Jimenez Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: (2016). arXiv: 1505.05770 [cs, stat].

[13] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density Estimation Using Real NVP". In: (2017). arXiv: 1605.08803 [cs, stat].

[14] Diederik P. Kingma and Prafulla Dhariwal. "Glow: Generative Flow with Invertible 1x1 Convolutions". In: (2018). arXiv: 1807.03039 [cs, stat].

[15] Aäron van den Oord et al. "WaveNet: A Generative Model for Raw Audio". In: (2016). arXiv: 1609.03499 [cs].

[16] Aäron van den Oord et al. "Conditional Image Generation with PixelCNN Decoders". In: (2016). arXiv: 1606.05328 [cs].

[17] P. Dutilleux. "An Implementation of the Algorithme à Trous to Compute the Wavelet Transform". In: *Wavelets*. Ed. by Jean-Michel Combes, Alexander Grossmann, and Philippe Tchamitchian. Inverse Problems and Theoretical Imaging. Berlin, Heidelberg: Springer, 1990, pp. 298–304.

[18] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[19] *Recommendation G. 711. Pulse Code Modulation (PCM) of Voice Frequencies*. 1988.

[20] Tom Le Paine et al. "Fast Wavenet Generation Algorithm". In: (2016). arXiv: 1611.09482 [cs].

[21] Tim Salimans et al. "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications". In: (2017). arXiv: 1701.05517 [cs, stat].

[22] Nal Kalchbrenner et al. "Efficient Neural Audio Synthesis". In: (2018). arXiv: 1802.08435 [cs, eess].

[23] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. "WaveGlow: A Flow-Based Generative Network for Speech Synthesis". In: (2018). arXiv: 1811.00002 [cs, eess, stat].

[24] Sungwon Kim et al. "FloWaveNet : A Generative Flow for Raw Audio". In: (2019). arXiv: 1811.02155 [cs, eess].

[25] Dario Rethage, Jordi Pons, and Xavier Serra. "A Wavenet for Speech Denoising". In: (2018). arXiv: 1706.07162 [cs].

Methodology

Datasets

ToyData

MusDB

Planning

Bibliography

Burgess, Christopher P. et al. “Understanding Disentangling in Beta-VAE”. In: (2018). arXiv: 1804.03599 [cs, stat] (cit. on p. 6).

Chorowski, Jan et al. “Unsupervised Speech Representation Learning Using WaveNet Autoencoders”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12 (2019), pp. 2041–2053. arXiv: 1901.08810 (cit. on p. 7).

Défossez, Alexandre et al. “Demucs: Deep Extractor for Music Sources with Extra Unlabeled Data Remixed”. In: (2019). arXiv: 1909.01174 [cs, eess, stat] (cit. on p. 7).

Dinh, Laurent, David Krueger, and Yoshua Bengio. “NICE: Non-Linear Independent Components Estimation”. In: (2015). arXiv: 1410.8516 [cs] (cit. on p. 6).

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. “Density Estimation Using Real NVP”. In: (2017). arXiv: 1605.08803 [cs, stat] (cit. on p. 7).

Dutilleux, P. “An Implementation of the Algorithme à Trous to Compute the Wavelet Transform”. In: *Wavelets*. Ed. by Jean-Michel Combes, Alexander Grossmann, and Philippe Tchamitchian. Inverse Problems and Theoretical Imaging. Berlin, Heidelberg: Springer, 1990, pp. 298–304 (cit. on p. 7).

Gershman, Samuel and Noah Goodman. “Amortized Inference in Probabilistic Reasoning”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36. 36. 2014.

Higgins, Irina et al. “Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework”. In: (2016) (cit. on p. 6).

Hochreiter, Sepp and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 7).

Jordan, Michael I. et al. “An Introduction to Variational Methods for Graphical Models”. In: *Machine Learning* 37.2 (1999), pp. 183–233 (cit. on p. 6).

Kalchbrenner, Nal et al. “Efficient Neural Audio Synthesis”. In: (2018). arXiv: 1802.08435 [cs, eess] (cit. on p. 7).

Kim, Sungwon et al. “FloWaveNet : A Generative Flow for Raw Audio”. In: (2019). arXiv: 1811.02155 [cs, eess] (cit. on p. 7).

Kingma, Diederik P. and Prafulla Dhariwal. “Glow: Generative Flow with Invertible 1x1 Convolutions”. In: (2018). arXiv: 1807.03039 [cs, stat] (cit. on p. 7).

Kingma, Diederik P. and Max Welling. “Auto-Encoding Variational Bayes”. In: (2014). arXiv: 1312.6114 [cs, stat] (cit. on p. 6).

Kingma, Diederik P. and Max Welling. “An Introduction to Variational Autoencoders”. In: (2019). arXiv: 1906.02691 [cs, stat].

Lluís, Francesc, Jordi Pons, and Xavier Serra. "End-to-End Music Source Separation: Is It Possible in the Waveform Domain?" In: (2019). arXiv: 1810.12187 [cs, eess] (cit. on p. 7).

Paine, Tom Le et al. "Fast Wavenet Generation Algorithm". In: (2016). arXiv: 1611.09482 [cs] (cit. on p. 7).

Prenger, Ryan, Rafael Valle, and Bryan Catanzaro. "WaveGlow: A Flow-Based Generative Network for Speech Synthesis". In: (2018). arXiv: 1811.00002 [cs, eess, stat] (cit. on p. 7).

Recommendation G. 711. Pulse Code Modulation (PCM) of Voice Frequencies. 1988 (cit. on p. 7).

Rethage, Dario, Jordi Pons, and Xavier Serra. "A Wavenet for Speech Denoising". In: (2018). arXiv: 1706.07162 [cs] (cit. on p. 7).

Rezende, Danilo Jimenez and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: (2016). arXiv: 1505.05770 [cs, stat] (cit. on p. 6).

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: (2014). arXiv: 1401.4082 [cs, stat] (cit. on p. 6).

Salimans, Tim et al. "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications". In: (2017). arXiv: 1701.05517 [cs, stat] (cit. on p. 7).

Stoller, Daniel, Sebastian Ewert, and Simon Dixon. "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation" (Paris, France). 2018 (cit. on p. 7).

Van den Oord, Aäron, Oriol Vinyals, and Koray Kavukcuoglu. "Neural Discrete Representation Learning". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 6306–6315 (cit. on p. 6).

Van den Oord, Aäron et al. "Conditional Image Generation with PixelCNN Decoders". In: (2016). arXiv: 1606.05328 [cs] (cit. on p. 7).

Van den Oord, Aäron et al. "WaveNet: A Generative Model for Raw Audio". In: (2016). arXiv: 1609.03499 [cs] (cit. on p. 7).