LQVAE + LASS hybrid

January 19, 2024

Michele Pieroni

Abstract

Music source separation is a critical task in audio processing, involving the extraction of individual sources of instruments from a mixed audio signal. This project compares the source separation performance of two models: *Linearly Quantized Variational Autoencoder* (**LQ-VAE**) and *Latent Autoregressive Source Separation* (**LASS**). To ensure a fair comparison, both models were retrained using a subset of the publicly available Slakh dataset.

Furthermore, the project explores a novel hybrid approach by incorporating the technique of counting occurrences in the model codebook, as in LASS, into a model with a loss function analogous to LQ-VAE.

1. Introduction

Music source separation is a critical issue in audio signal processing, the objective is to isolate a mixture of multiple signals into their respective sources. The significance of solving this problem extends to various use cases in audio processing tasks: improving audio quality in restoration, increasing accuracy in speech recognition systems, or solving the 'cocktail party'(Haykin & Chen, 2005) problem.

Deep learning methods have been extensively studied to tackle this issue. By leveraging their capacity to capture essential features of the data they provide a considerable advantage over traditional methods. Generative deep models have tried to address the challenge of dealing with the intricate high-level semantics of raw audio by transitioning to a lower-dimensional space: redundant information will be lost while still keeping the essential features. This is the main idea behind the Vector Quantised-Variational AutoEncoder (**VQ-VAE**) (Van Den Oord et al., 2017): rather than relying on parametrized distributions, the VQ-VAE creates a mapping from the data to a discrete latent representation.

One model that employs this approach is OpenAI's Jukebox (Dhariwal et al., 2020). Its structure is built on a hierarchical VQ-VAE and an autoregressive Sparse Transformer architecture (Child et al., 2019). Both LASS (Postolache et al., 2023) and LQ-VAE (Mancusi et al., 2021) propose a novel approach by leveraging the useful properties of this architecture. The distinction between them lies in the computation of the VQ-VAE loss function and the implementation of Bayesian inference. This project aims to demonstrate and compare which of the two returns superior separation performance and, finally, present a hybrid model that combines the contributions of both approaches: impose a post-quantization linearization loss on the VQ-VAE, enforcing an algebraic structure on the latent domain, and express the likelihood function in the Bayesian posterior by employing discrete conditionals.

2. Background

This section will briefly describe the core components behind LQ-VAE and LASS. The first and most important concept behind both methods is the VQ-VAE.

2.1. VQ-VAE

Given an input sequence $\mathbf{x} \in \mathbb{R}^n$, x is mapped to a latent space via an encoder: $E_\theta : \mathbb{R}^n \mapsto \mathbb{R}^{C \times S}$ where C denotes the number of latent channels and S the length of the latent sequence. Then the bottleneck is applied: $B_\theta : \mathbb{R}^{C \times S} \mapsto [K]^S$, this function maps the latent vectors to the index of the closest latent code. Latent codes consist of a single set of learnable vectors $\mathbb{L} = \{e_k\}_{k=1}^K \subset \mathbb{R}^C$. Finally the decoder $D_\theta : [K]^S \mapsto R^n$ reconstructs the sequence back to the data domain.

2.2. Latent autoregressive priors

Since the mapping from the audio domain to a latent sequence $\mathbf{z}$ has been established, it is possible to estimate p_{data} through an autoregressive prior $p_\phi(\mathbf{z})$. Sequences of latent codes can be sampled through ancestral sampling and then decoded to generate new data examples. This process

Email: Michele Pieroni <pieroni.1704202@studenti.uniroma1.it>.

Deep Learning and Applied AI 2023, Sapienza University of Rome, 2nd semester a.y. 2022/2023.

is learned by the Transformers within the Jukebox architecture.

2.3. LQ-VAE

The goal is to separate the sources $\mathbf{x}_1, \mathbf{x}_2$ given the mixture $\mathbf{y}$, achieved through the utilization of the Bayesian posterior that can be locally expressed in the latent domain as:

$$p(\mathbf{z}_s|\mathbf{z}_{<s}, \mathbf{y}^{\text{latent}}_{\leq s}) \propto p_\phi(\mathbf{z}_s|\mathbf{z}_{<s}) p(\mathbf{y}^{\text{latent}}_{\leq \mathbf{s}}|\mathbf{z}_{\leq s}) \quad (1)$$

The original paper's authors opted to represent the (logarithmic) likelihood as:

$$-\frac{1}{2\sigma^2}\|\mathbf{y}^{\text{latent}}_s - \mathbf{B}(\frac{1}{2}\mathbf{e}_{z_1} + \frac{1}{2}\mathbf{e}_{z_2})\|_2^2 \quad (2)$$

At every step s the mixture is compared to the scaled sum of all possible latent codes. To compute this sum the LQ-VAE has to impose an algebraic structure on the latent space through its novel loss term:

$$\mathcal{L}_{\text{lin}} = \frac{1}{T}\sum_t \|\mathbf{L}\mathbf{Q}_t - \mathbf{Q}\mathbf{L}_t\|_2^2 \quad (3)$$

By minimizing this loss, the model forces the latent representation of the mixture of two sources (LQ term) to be equal to the sum of the latent representation of the individual sources (QL term).

2.4. LASS

An alternative strategy aims to address the modeling of the likelihood in Eq. (1) while preserving the model's structure. This method proposes to model the likelihood through discrete conditional probabilities, represented with a tensor $P \in \mathbf{R}^{K\times K\times K}$.

$$p(\cdot|z_1, z_2) = P_{z_1, z_2, :} \quad (4)$$

This rank-3 tensor is computed by counting occurrences of the latent mixed tokens given the latent sources' tokens. At inference time the likelihood is obtained by slicing the tensor:

$$p(y_s^{latent}|\cdot,\cdot) = P_{:,:,y_s^{\text{latent}}} \quad (5)$$

3. Method

The novel approach suggested by this project for source separation is a hybrid of LQ-VAE and LASS. The VQ-VAE is trained with the addition of Eq. (3) The likelihood at inference time is computed with Eq. (5). This blending of methods is driven by the main goal: to ascertain whether this hybrid approach can outperform the individual models in terms of performance while concurrently preserving efficiency during inference.

To guarantee a fair comparison all three models are retrained and then tested on a subset of songs from the Slakh (Manilow et al., 2019) dataset. From the thousands of songs, only the initial 300 bass and drums tracks were selected from the train section, all of them are monochannel and have a sampling rate of 22KHz. Once training of the VQ-VAEs and priors has been completed, the models are tested on chunks of 3 seconds, each extracted from one of 90 tracks selected from the initial 300. Signal to Distortion Ratios (**SDR**) (Stöter et al., 2018) is used to measure the separation performance.

The chosen sampling strategy is ancestral sample without top-*k*, a decision guided by its superior performance compared to beam search in experiments. All hyper-parameters were selected based on the previously determined configuration of LQ-VAE.

4. Results

The experiments were conducted using Google Colab, which provides competitive hardware. However, the imposed time limit constrained the scope of this research by limiting the available time for extensive model training, particularly for the autoregressive priors. A comparison of the SDR metric and separation time among all models is presented in Table 1. The complete code, some samples of qualitative results, and a more detailed presentation of SDR scores can be found at: https://github.com/Pieroni1704202/LQVAE-LASS-hybrid

Table 1. The reported scores are the average SDR between the best-generated samples and the original sources.

Model	Bass	Drums	Avg. separation time (s)
LQ-VAE	3.41	**5.41**	254.77
LASS	3.22	4.78	148.57
Hybrid	**4.72**	5.15	**148.4**

5. Discussion and conclusions

In conclusion, the comparison doesn't decisively favor one over the others. The hybrid model gains the efficiency at inference time from LASS, but its separation performance doesn't significantly surpass competitors. The flexibility of LASS is overall a great property, modifying a pre-trained model while achieving satisfactory results should not be overlooked. However, it must be acknowledged that given enough computational resources and with extensive hyperparameter tuning there could be a justification for choosing the hybrid model over its original counterparts. Furthermore, training the model on custom data may be better suited for specific tasks.

References

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Haykin, S. and Chen, Z. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

Mancusi, M., Postolache, E., Mariani, G., Fumero, M., Santilli, A., Cosmo, L., and Rodolà, E. Unsupervised source separation via bayesian inference in the latent domain. *arXiv preprint arXiv:2110.05313*, 2021.

Manilow, E., Wichern, G., Seetharaman, P., and Le Roux, J. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 45–49. IEEE, 2019.

Postolache, E., Mariani, G., Mancusi, M., Santilli, A., Cosmo, L., and Rodola, E. Latent autoregressive source separation. *arXiv preprint arXiv:2301.08562*, 2023.

Stöter, F.-R., Liutkus, A., and Ito, N. The 2018 signal separation evaluation campaign. In *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Surrey, UK*, pp. 293–305, 2018.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.