Connective Fusion: Learning Transformational Joining of Sequences with Application to Melody Creation

Taketo Akama
Sony Computer Science Laboratories (CSL), Tokyo
taketo.akama@sony.com

Sony CSL

Introduction

We present

- *Connective Fusion*, a sequence generation scheme by transformational joining of two sequences for creative purposes.
- Our model connects and fuses possibly unrelated sequences in a coherent way.
- This transformation can be applied iteratively to gradually fuse the input sequences. (Fig1, left)
- The *style latent space* is simultaneously learned, allowing users to control how the two sequences are merged. (Fig1, right)

Connective Fusion in music is useful for

- providing users with musical ideas based on not necessarily polished ideas on hand.
- creating novel musical flows by combining and fusing musical fragments of different characteristics in a coherent way.

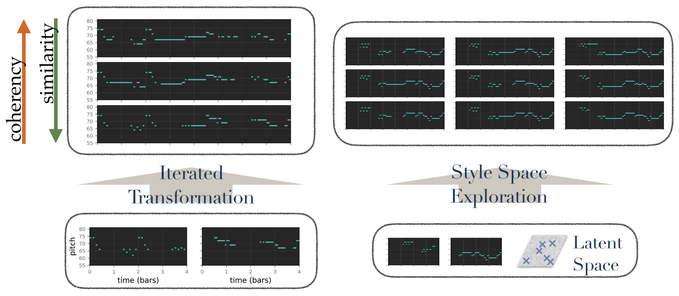

Figure 1: Illustration of Application Examples

Model Training

- Train variational auto-encoder (VAE) to obtain $q_{\hat{\theta}}$ and $p_{\hat{\theta}}$. (Fig.2, Pre-train)
- Using $q_{\hat{\theta}}$, train mapping G_ψ with adversarial learning. (Fig.2, Train)

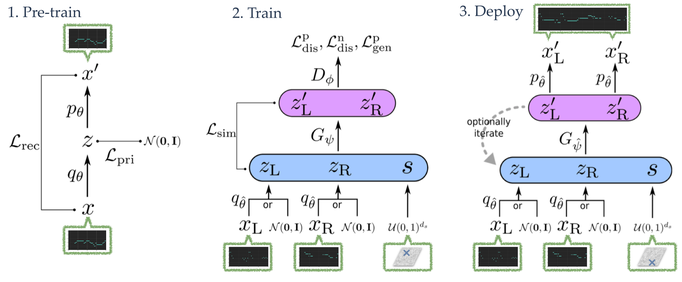

Figure 2: Model Schematic

Model Deployment

- Using $q_{\hat{\theta}}$, $p_{\hat{\theta}}$, and $G_{\hat{\psi}}$, conduct generative transformation. (Fig.2, Deploy)
- Iterated transformation is done by feeding the transformed latent vectors back into the input.
- Style space exploration is done by altering the input s.

Loss Function Details

We use short hand $\mathcal{L}^{\mathrm{p}}_{D_\phi}(z_{\mathrm{L}}, z_{\mathrm{R}}) = -\log D_\phi(z_{\mathrm{L}}, z_{\mathrm{R}})$, $\mathcal{L}^{\mathrm{n}}_{D_\phi}(z_{\mathrm{L}}, z_{\mathrm{R}}) = -(1 - \log D_\phi(z_{\mathrm{L}}, z_{\mathrm{R}}))$, and $\mathcal{U} = \mathcal{U}(0,1)^{d_s}$. $\mathcal{D}_z$ and $\mathcal{D}'_z$ denote paired and unpaired dataset of latent vectors, respectively. The loss function for the discriminator is

$$\mathcal{L}_{\mathrm{dis}} = \mathcal{L}^{\mathrm{p}}_{\mathrm{dis}} + \mathcal{L}^{\mathrm{n}}_{\mathrm{dis}}, \tag{1}$$

where

$$\mathcal{L}^{\mathrm{p}}_{\mathrm{dis}} = \mathop{\mathbb{E}}_{(z_{\mathrm{L}}, z_{\mathrm{R}}) \sim \mathcal{D}_z} \left[\mathcal{L}^{\mathrm{p}}_{D_\phi}(z_{\mathrm{L}}, z_{\mathrm{R}}) \right] \tag{2}$$

$$\mathcal{L}^{\mathrm{n}}_{\mathrm{dis}} = \mathop{\mathbb{E}}_{z_{\mathrm{L}} \sim \mathcal{D}'_z} \mathop{\mathbb{E}}_{z_{\mathrm{R}} \sim \mathcal{D}'_z} \left[\mathcal{L}^{\mathrm{n}}_{D_\phi}(z_{\mathrm{L}}, z_{\mathrm{R}}) \right] + \mathop{\mathbb{E}}_{z_{\mathrm{L}} \sim p(z)} \mathop{\mathbb{E}}_{z_{\mathrm{R}} \sim p(z)} \left[\mathcal{L}^{\mathrm{n}}_{D_\phi}(z_{\mathrm{L}}, z_{\mathrm{R}}) \right] + \mathop{\mathbb{E}}_{z_{\mathrm{L}} \sim \mathcal{D}'_z} \mathop{\mathbb{E}}_{z_{\mathrm{R}} \sim \mathcal{D}'_z} \mathop{\mathbb{E}}_{s \sim \mathcal{U}} \left[\mathcal{L}^{\mathrm{n}}_{D_\phi}(G_\psi(z_{\mathrm{L}}, z_{\mathrm{R}}, s)) \right] + \mathop{\mathbb{E}}_{z_{\mathrm{L}} \sim p(z)} \mathop{\mathbb{E}}_{z_{\mathrm{R}} \sim p(z)} \mathop{\mathbb{E}}_{s \sim \mathcal{U}} \left[\mathcal{L}^{\mathrm{n}}_{D_\phi}(G_\psi(z_{\mathrm{L}}, z_{\mathrm{R}}, s)) \right]. \tag{3}$$

The loss function for the generator is

$$\mathcal{L}_{\mathrm{gen}} = \mathcal{L}^{\mathrm{p}}_{\mathrm{gen}} + \lambda \mathcal{L}_{\mathrm{sim}}, \tag{4}$$

where

$$\mathcal{L}^{\mathrm{p}}_{\mathrm{gen}} = \mathop{\mathbb{E}}_{z_{\mathrm{L}} \sim \mathcal{D}'_z} \mathop{\mathbb{E}}_{z_{\mathrm{R}} \sim \mathcal{D}'_z} \mathop{\mathbb{E}}_{s \sim \mathcal{U}} \left[\mathcal{L}^{\mathrm{p}}_{D_\phi}(G_\psi(z_{\mathrm{L}}, z_{\mathrm{R}}, s)) \right] + \mathop{\mathbb{E}}_{z_{\mathrm{L}} \sim p(z)} \mathop{\mathbb{E}}_{z_{\mathrm{R}} \sim p(z)} \mathop{\mathbb{E}}_{s \sim \mathcal{U}} \left[\mathcal{L}^{\mathrm{p}}_{D_\phi}\left(G_\psi(z_{\mathrm{L}}, z_{\mathrm{R}}, s)\right) \right] \tag{5}$$

$$\mathcal{L}_{\mathrm{sim}} = \mathop{\mathbb{E}}_{z_{\mathrm{L}} \sim \mathcal{D}'_z} \mathop{\mathbb{E}}_{z_{\mathrm{R}} \sim \mathcal{D}'_z} \mathop{\mathbb{E}}_{s \sim \mathcal{U}} \left[\mathcal{L}_{\mathrm{dist}}(z_{\mathrm{L}}, z_{\mathrm{R}}, s) \right] + \mathop{\mathbb{E}}_{z_{\mathrm{L}} \sim p(z)} \mathop{\mathbb{E}}_{z_{\mathrm{R}} \sim p(z)} \mathop{\mathbb{E}}_{s \sim \mathcal{U}} \left[\mathcal{L}_{\mathrm{dist}}(z_{\mathrm{L}}, z_{\mathrm{R}}, s) \right] \tag{6}$$

with

$$\mathcal{L}_{\mathrm{dist}}(z_{\mathrm{L}}, z_{\mathrm{R}}, s) = \frac{1}{d_z} \left\| \frac{1}{\bar{\sigma}^2_z} \log\left(1 + \left(z'_{\mathrm{L}} - z_{\mathrm{L}}\right)^2\right) \right\|_1 + \frac{1}{d_z} \left\| \frac{1}{\bar{\sigma}^2_z} \log\left(1 + \left(z'_{\mathrm{R}} - z_{\mathrm{R}}\right)^2\right) \right\|_1$$
$$(z'_{\mathrm{L}}, z'_{\mathrm{R}}) = G_\psi(z_{\mathrm{L}}, z_{\mathrm{R}}, s). \tag{7}$$

Experiments

Higher coherency at a given similarity is considered to be better. To evaluate similarity, we use z-distance. To evaluate coherency, we use a reality classifier and 5 musical statistics. As depicted on the left of Fig.3, transition matrices are estimated for musical statistics. Higher correlation with test data can be seen after transformation. The right of Fig.3 suggests that sequences from different song ids are much less likely to have smaller z-distances.

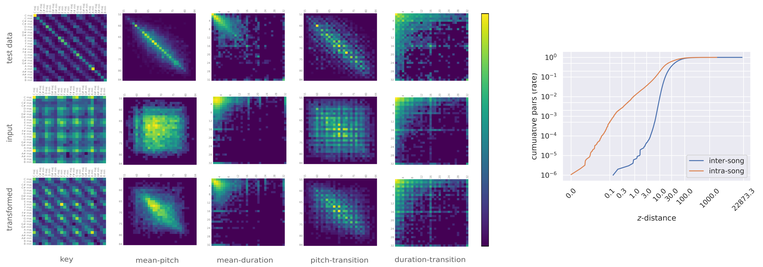

Figure 3: Visualization of musical statistics and analysis of latent space.

Fig.4 compares ours with baseline methods. In most cases, ours are over the baseline methods (LP), even though LP uses abundant computational budgets. LP consists in iterated random search in the latent space to have a higher reality classifier output.

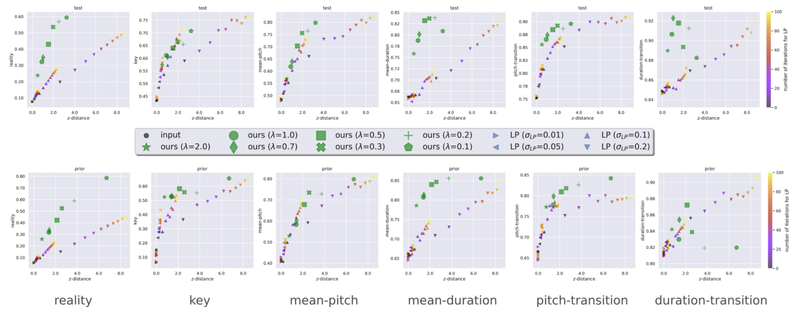

Figure 4: Evaluation of ours vs baseline methods (LP). The upper/lower rows are transformation results where inputs are randomly created pairs of test/generated (sampled from prior) data. The color gradient corresponds to the number of iterations for LP. For all metrics, higher values are better. Higher λ enforces the model to search the neighbour solution in the latent space (see Eq.4).

Fig.5 compares different number of iterations of our method for evaluating iterated transformation. Ours with larger λ (e.g., $\star$) perform well in different z-distances. Iterated transformation with larger λ allows users to explore the good balance between the coherency and the similarity.

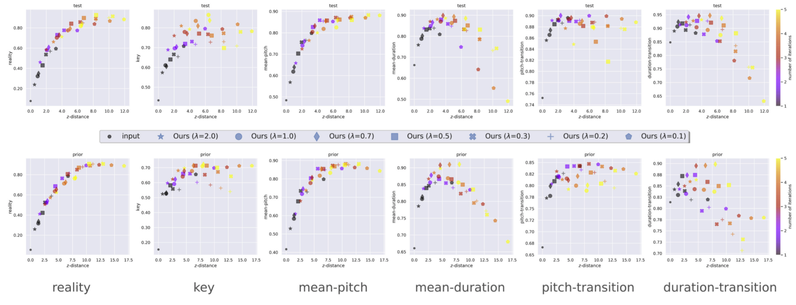

Figure 5: Evaluation of Iterated Transformation. The color gradient corresponds to the number of iterations for our method.

Related Work

- Concatenative Synthesis. Mashup.
- Learning transformation in the latent space.
- Learning transformation with adversarial learning without paired data.