Abstract

Learning a mapping between two unrelated domains—such as image and audio, without any supervision—is a challenging task. In this work, we propose to use a distance-preserving generative adversarial model to translate images of human faces into an audio domain. The audio domain is defined by a collection of musical note sounds recorded by 10 different instrument families (NSynth [5]) and a distance metric where the instrument family class information is incorporated together with a mel-frequency cepstral coefficients (MFCCs) feature. To enforce distance-preservation, a loss term that penalizes the difference between pairwise distances of the faces and the translated audio samples is used. Further, we discover that the distance preservation constraint in the generative adversarial model leads to reduced diversity in the translated audio samples, and propose the use of an auxiliary discriminator to enhance the diversity of the translations while using the distance preservation constraint. We also provide a visual demonstration of the results (video demo) and numerical analysis of the fidelity of the translations.

1. Introduction

There has been a lot of work in attempting to find a meaningful mapping between two different domains without any supervision—i.e. only unordered and unpaired samples from the two domains are given. However, most of them are often limited to a task of mapping between two visual domains such as image-to-image translation [6, 9, 20], where the learner takes an image in one domain and maps it into another pixel domain with its content or style changed to be similar to that of the target samples. In this work, we focus on finding a mapping between two highly unrelated domains—image and audio. More precisely, we want to find a meaningful mapping that links images of human faces into musical sounds, while the measure of similarity between face images is consistent with the one between translated audio samples so that visually dis/similar face images can be mapped into audio samples that sound perceptually dis/similar to each other. This could be used to aid those with a visual impairment by encoding the visual information (e.g. facial appearance) into the audio domain, which allows them to perceive this information using their ears [15, 11].

1Typically GANs take as input a latent vector sampled from a Gaussian or uniform distribution. For our task, we are using the pretrained feature embedding space as the input latent.
criminators are simultaneously trained with the generator, whose task is to predict whether the generated output is real or fake. See the overview of our model’s structure in Fig. 1.

**Metric Preservation Constraint:** To enforce the distance preservation constraint, we add a metric loss term (Eq. 1) to the GAN’s adversarial loss, which computes the difference of pairwise distances of the face features and features of the generated audio samples (‘Pairwise distance loss’ block in Fig. 1).

\[
L_{\text{metric}} = \frac{1}{Z} \sum_{i<j} \frac{||f(x_i) - f(x_j)||_2 - \mu_x}{\sigma_x} - \frac{||\phi(y_i) - \phi(y_j)||_2 - \mu_y}{\sigma_y}
\]

where \( Z \) is the number of possible (unordered) pairs of samples in the training mini-batch, \( f(x) \) is the face embedding feature of the input face image \( x \), \( \phi(y) \) denotes our audio embedding feature of the translated audio output \( y \). The standardization parameters, \( (\mu_x, \sigma_x) \) and \( (\mu_y, \sigma_y) \), are derived from source (VGGFace2) and target (NSynth) datasets respectively and are the mean and standard deviation of pairwise distances of samples from within each respective dataset.

**Audio Metric:** The audio feature is computed by a combination of a pretrained audio feature embedding model and MFCCs (The ‘Audio feature embedding’ block in Fig. 1): We first train a feature embedding model for audio that maps raw audio waveforms into a compact Euclidean space where the distance directly corresponds to a measure of audio similarity. This is done by using the triplet loss function [18] that aims to separate the positive sample from the negative sample by a distance margin, where the positive and negative are determined by the annotated instrument family class label of the NSynth dataset. As a result, audio samples that have the same musical instrument timbre are mapped into features with small distances and vice versa. To further incorporate perceptual notions of distance into the audio metric, the Mel-frequency cepstral coefficients (MFCCs) [19] of the audio are computed and concatenated with the learned audio features. Fig. 2 (b) shows the audio metric space of NSynth samples. For the visualization, we mapped them into a 2D space by tSNE [10]. Each blob represents an NSynth sample’s feature vector and is color-coded by the instrument family (bass, brass, flute, guitar, keyboard, mallet, organ, reed, string, and vocal). For a pitch classifier and a family classifier, the distance preservation loss and then the second discriminator to the main discriminator in a way that it further encourages the generative model to output audio samples that fit into the distribution of the target audio dataset. Note that, unlike the main discriminator, we attach this second discriminator on top of the hybrid MFCC-learned audio embedding feature (Fig. 1 ‘Audio feature embedding’ block) to predict the real/fake by processing the audio features rather than the raw audio signals. As we will see in the next section, this helps the generative model output audio samples that are spread over the distribution of real audio samples in our audio metric space (Fig. 2 (b) third row). Furthermore, our additional discriminator has a much simpler and more economical architecture than the main discriminator, consisting of five fully connected layers with dimension 128, 64, 32, 16, and 1.

**3. Results**

In this section, we detail an ablation study performed on our proposed model by adding sequentially the metric preservation loss and then the second discriminator to the baseline to see the effect of each constraint. See Table 1.

The baseline model is based on WaveGAN [3]. It is trained to, given a 512-dimensional VGGFace2 feature vector, generate a raw audio signal of length 8192 that fits into the distribution of NSynth training samples. To improve the quality of the generated sounds, we employ a state-of-the-art stabilization technique, spectral normalization [12]. Further, our baseline model is conditioned by a ‘pitch’ label given by the NSynth dataset (i.e. our model is a conditional GAN similar to that described in [4]) but with conditional batch normalization layers [14] in the generative network and a projection layer at the end of the discriminator as in [13].

Table 1 shows four evaluation metric scores: 1) Pearson product momentum correlation (PC) between L2 pairwise distances of the samples in the source metric and the corresponding translated samples in the audio metric. The Pearson correlation has a value between +1 and -1, where

<table>
<thead>
<tr>
<th>Method</th>
<th>Pitch</th>
<th>Family</th>
<th>IS</th>
<th>FID</th>
<th>PC</th>
<th>SV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25.52</td>
<td>54.97</td>
<td>53.98</td>
<td>3.98, 4.09</td>
<td>0.683</td>
<td>0.0112</td>
</tr>
<tr>
<td>+metric preservation</td>
<td>46.69</td>
<td>53.18</td>
<td>90.99</td>
<td>3.98, 4.09</td>
<td>0.683</td>
<td>0.0112</td>
</tr>
<tr>
<td>+aux. discriminator</td>
<td>25.52</td>
<td>54.97</td>
<td>53.98</td>
<td>3.98, 4.09</td>
<td>0.683</td>
<td>0.0112</td>
</tr>
<tr>
<td>+metric preservation +aux. discriminator</td>
<td>46.69</td>
<td>53.18</td>
<td>90.99</td>
<td>3.98, 4.09</td>
<td>0.683</td>
<td>0.0112</td>
</tr>
</tbody>
</table>

2FID is known to be sensitive to the variety of the generated samples.
FID is zero for both classification models. For the pitch class, it is 61, and for the family class, it is 10. The minimum FID score decreased).

The translations play a much wider variety of musical sounds (so FID real clusters (See the third plot in Fig. 2 (b)) and the translations spread evenly over the metric space in a way that they are spread evenly over the audio metric space. We observed a reduced quality in the samples, i.e., FID increases from 21.84 (51.02) to 56.27 (4.59) respectively by our pre-trained pitch (family) classification model.

With the distance metric preservation constraint on the model, the Pearson Correlation (PC) score increases from 0.090 to 0.683. This demonstrates that our metric preservation loss helps the model preserve the geometric structure of the face embeddings in the target audio metric space, however, we also observed a reduced quality in the samples, i.e., FID increases from 21.84 (51.02) to 46.69 (90.00), and IS decreases from 56.27 (4.59) to 53.18 (3.98). Our visualization of the translated audio samples in the audio metric space also demonstrates this. See Fig. 2 (b). The red dots represent the translated audio clips from a subset of 20k randomly sampled face images from the VGGFace2 train set and the blobs are the NSynth real audio samples. The first two top figures correspond to the baseline and the our proposed model with metric preservation constraint on the model, the Pearson Correlation (PC) score increases from 0.090 to 0.683. This demonstrates that our metric preservation loss helps the model preserve the geometric structure of the source metric space.

We have proposed a distance preserving generative adversarial network that automatically learns an information preserving embedding between two unrelated domains of media (i.e., image and audio domain). We discovered that there is a trade-off between the variety of translations and the preservation of geometric information. To address this

---

3Note that the maximum of IS score is the number of classes which is 61 for the pitch class and 10 for the family class. The minimum score of FID is zero for the both classification models.
problem, we proposed to use the auxiliary discriminator that can support the primary discriminator, which helps to balance the performance between the diversity and the metric preservation. We demonstrate that the proposed model translates a face image from the VGGFace2 dataset into a musical sound that plays one of 10 instrument family in the NSynth dataset and the faces playing the same instrument type show a pattern on their visual appearance (e.g. skin tone, hair color etc.).

References