MidiMe: Personalizing a MusicVAE model with user data

Monica Dinculescu
Google
noms@google.com

Jesse Engel
Google
jesseengel@google.com

Adam Roberts
Google
adarob@google.com

Abstract

We introduce an approach to quickly train a small personalized model to control a larger pretrained latent variable model. To show its application for creative interactions, we implement the model in TensorFlow.js as a standalone application, so that training happens in real-time, in a browser, closest to the user.

1 Motivation

One of the areas of interest for music generative models is to empower individual expression. But how can a creator personalize a machine learning model to make it their own?

Training a custom deep neural network model like Music Transformer[3], MusicVAE[4] or SketchRNN[2] from scratch requires significant amounts of data (millions of examples) and compute resources (specialized hardware like GPUs/TPUs) as well as expertise in hyperparameter tuning. Without sufficient data, models are either unable to produce realistic output (underfitting), or they memorize the training examples and are unable to generalize to produce varied outputs (overfitting) – it would be like trying to learn all of music theory from a single song.

We introduce a new model for sample-efficient adaptation to user data, based on prior work by Engel et al[1]. We can quickly train this small, personalized model to control a much larger, more general pretrained latent variable model. This allows us to generate samples from only the portions of the latent space we are interested in without having to retrain the large model from scratch. In fact, training this model is so fast that we can do it in a couple of seconds in the browser, using Magenta.js[5].

We demonstrate this technique in an online demo[1] that lets users upload their own MIDI files (melodies or multi-instrument songs) and generate samples that sound like their input. Screenshots of this demo can be seen in Figure 1.

2 Constraining MusicVAE samples

MusicVAE[4] is a hierarchical variational autoencoder that learns a summarized representation of musical qualities as a latent space. It encodes a musical sequence into a latent vector, which can then later be decoded back into a musical sequence. Because the latent vectors are regularized to be similar to a standard normal distribution, it is also possible to sample from the distribution of sequences, generating realistic music based on a random combination of qualities.

However, even though MusicVAE's latent space is significantly smaller than the space of all possible note sequences, it was still trained to approximate many different musical sequences. This means that without knowing exactly which latent space a particular style or genre is encoded to, conditional sampling is not exactly possible. Even if you could encode a particular style to a latent vector, it's

[1] https://midi-me.glitch.me/

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

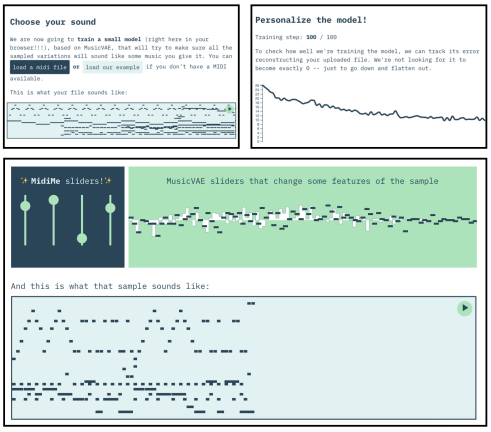

Figure 1: [Top left] The user uploads a MIDI file, visualizes its latent features, and listens to it. At this point (not shown), the user can also generate samples from the pretrained model, and explore its latent space by moving sliders representing its different dimensions. [Top right] After uploading the input, the user can train the personalized MidiMe model directly in the browser. [Bottom] After training a MidiMe, the user can control a much smaller latent space (4 dimensions instead of MusicVAE's 256). The generated samples at this point will preserve many qualities of the input while still introducing some variation. In this example, you can see the peaks of the original melody combined with non-memorized measures.

hard to determine how to further modify each of the 256 dimensions to sample more melodies in that style.

To solve this, we train a *smaller* VAE on MusicVAE's latent space. This smaller VAE (called "MidiMe") learns a compressed representation of the already encoded latent vectors. The intuition is that if 256 dimensions were enough to summarize the musical features of all training examples, then a much smaller number (like 4) can be enough to summarize the summarized musical feature of a particular input. MidiMe works both on single-instrument melodies, as well as multi-instrument trios. We illustrate this approach in Figure 2.

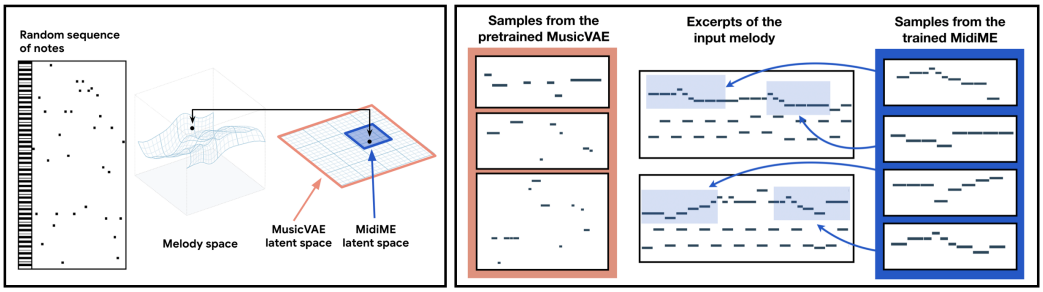

Figure 2: [Left] MidiMe is a Variational Autoencoder trained on the latent space of MusicVAE. [Right] Unlike the pretrained MusicVAE which has no knowledge of the input, samples from MidiMe resemble the structure of the input melody, without actually memorizing it.

Since MidiMe learns a subset of an existing MusicVAE latent space, its reconstruction or generative qualities depend on the MusicVAE model you start with. For example, a MusicVAE model trained with many free bits (a constraint on the minimum amount of information per group of latent variables) might be very good at reconstructing its input. The MidiMe model trained on its latent space will also learn to reconstruct the training data well, and most of its generated samples will sound similar to the training data (like the examples in Figure 2). In contrast, a model with fewer free bits is optimized to generate very plausible samples, but not good reconstructions – the MidiMe model trained on this will generate samples that don't sound identical to the training data, but have more varied motifs and musical patterns than the MusicVAE samples (like the example in Figure 1).

3 Ethical implications

As with most Generative Models such as GPT-2 or GANs, MidiMe blurs the lines between real and fake samples. A MusicVAE model could be personalized to generate realistic sounding variations of a song. However, these kind of "deepfakes" are not new to the musical community, where remixing and reusing is a fairly common practice. While it could be used for harmful copycats, MidiMe can

also be seen as a creative tool that empowers musicians to generate new samples that are based on their repertoire. This approach was recently employed by YACHT in their latest album[2], where the band used MusicVAE to interpolate between songs in their catalog. This was a tedious process, which could be streamlined by a model like MidiMe.

References

[1] Jesse Engel, Matt Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. In *International Conference on Learning Representations (ICLR)*, 2018.

[2] David Ha and Douglas Eck. A neural representation of sketch drawings. In *ICLR 2018*, 2018. 2018.

[3] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, and Douglas Eck. Music transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*, 2018.

[4] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 4364–4373, 2018.

[5] Adam Roberts, Curtis Hawthorne, and Ian Simon. Magenta.js: A javascript api for augmenting creativity with deep learning. In *Joint Workshop on Machine Learning for Music (ICML)*, 2018.

[2] `http://bit.ly/yacht-arstechnica`