Generating music with sentiment using Transformer-GANs

Pedro L. T. Neves, Jose E. F. Novo Junior, João B. Florindo

State University of Campinas

Introduction

In summary, we propose a Machine Learning model based on Generative Adversarial Networks (GANs) capable of generating musical excerpts related to specific emotions quantified by Arousal and Valence.

Background

GANs - GANs [1] can learn to produce fake data that is similar to data originated from a real distribution through the adversarial interaction between two networks.

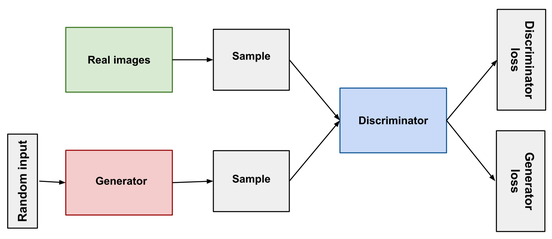

Fig. 1: Schematics of the GAN training loop.

Russell's Circumplex Model - A model where the human perception of emotions is represented as a product of two independent neurophysiological systems, valence (deactivation-activation) and arousal (unpleasantness-pleasantness).

GANs for discrete sequences - Due to the non-differentiability of the sampling process of a categorical distribution, GANs usually do not work well with discrete data. One strategy to solve this problem is to approximate the categorical distribution through the Gumbel-Softmax distribution [3].

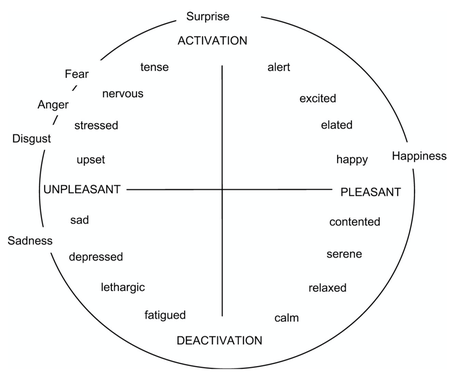

Fig. 2: Russel's Circumplex Model

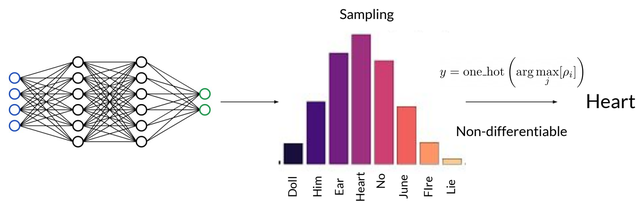

Fig. 3: Sampling of categorical distribution

$$y = \text{one\_hot}\left(\arg\max_j[g_i + \log \pi_i]\right)$$

$$y = \text{softmax}((1/\tau)(h + g))$$

Fig. 4: Gumbel-Softmax approximation

Methods

Both the Generative and Discriminator networks are Transformers [4]. Each model is made up of 6 Attention blocks.

Generator - The Generator network fulfills two roles: First, it is in charge of predicting each item of each sequence of the actual dataset based on the previous elements. The second objective of the network is to generate sequences that resemble those coming from the real set.

Discriminator - Each sequence received by the network is separated into subsequences of a predetermined size, which are then compressed into unique feature vectors. The Discriminator network produces two outputs: a global map associated with the whole sequence and a local map associated with each subsequence.

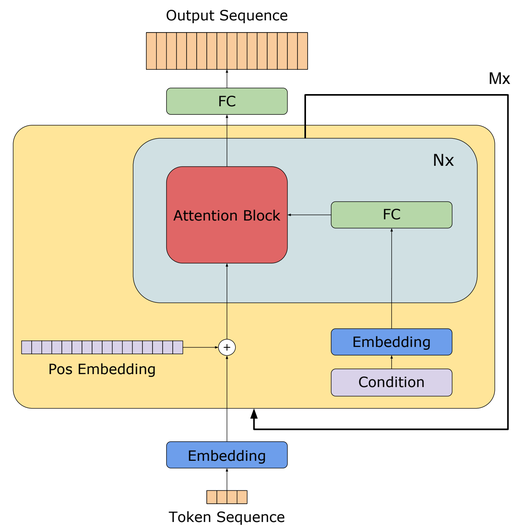

Fig. 5: Generator

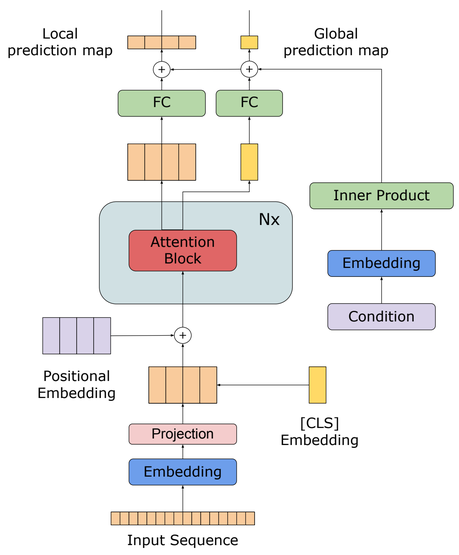

Fig. 6: Discriminator

Training We used two databases to train the models. The AILABS17k datased, that contains piano transcriptions of pop songs, and the EMOPIA dataset, which is similar, however, with the addition that the musical excerpts are separated into 4 emotional classes (4 in total, with valence and arousal being either high or low in each one).

Results

The model was compared with a non-adversarilly trained Transformer and with the Transformer model proposed in [2]. First, we did an evaluation with automatic metrics. We then carried out a survey where participants evaluated musically relevant characteristics of the synthesized material, as well as its affective content. Average scores for the Transformer-GAN, the conventional Transformer, and state-of-the-art baseline for these characteristics are given in the table below.

	PR	NPC	POLY
Real Data (EMOPIA)[2]	50.94	8.50	5.60
Baseline [2]	49.76	**8.52**	4.36
Transformer	48.79	8.65	4.37
Transformer GAN	**50.73**	9.45	**4.43**

Tab. 1: Comparison between the samples generated by ours and a state-of-art model in terms of automatic metrics. PR stands for Pitch Range, NPC is Number of Pitch Classes and POLY is Polyphony.

	H	O	S	OQ
Baseline [2]	3.32 ± 1.29	2.93 ± 1.13	3.18 ± 1.30	3.49 ± 1.04
Transformer	3.75 ± 1.24	3.22 ± 1.19	3.76 ± 1.14	3.89 ± 1.14
Transformer-GAN	3.56 ± 1.34	3.06 ± 1.21	3.38 ± 1.09	3.44 ± 1.15

Tab. 2: Comparison between the samples generated by ours and a state-of-art model in terms of human-evaluated metrics. The columns are, respectively, Human-Likeness, Originality, Structure and Overall Quality. The scale goes from 1 to 5.

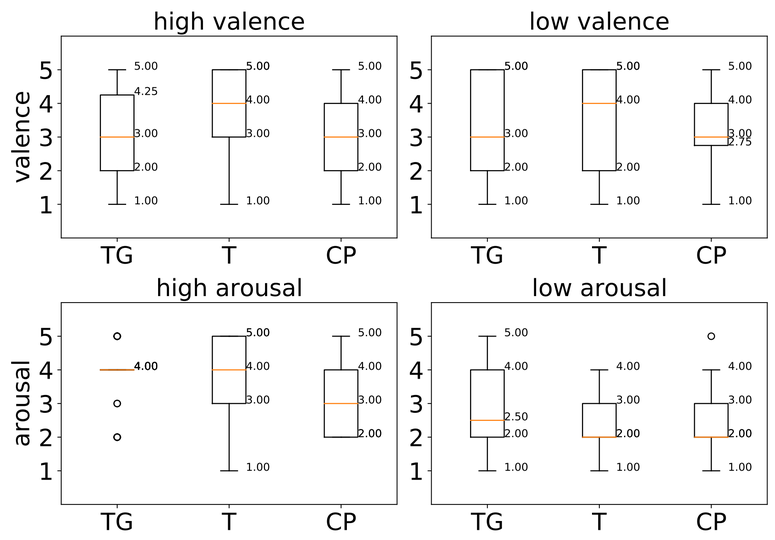

Fig. 7: Human Ratings of Valence and Arousal. TG, T and CP correspond, respectively, to the Transformer GAN, Transformer and Compound-Word Transformer Baseline.

Conclusion

- GANs achieve interesting results in the context of automatic music generation conditioned by emotion.
- The discrete nature of symbolic music makes training difficult, and current techniques used to solve these problems are limited.
- There are several open questions and possible avenues for future research in the field, such as using emotional labels in the audio domain or turning the models into tools that can be used for automatic generation of music with sentiment.

Bibliography

[1] Ian Goodfellow et al. "Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems* 3 (June 2014). DOI: 10.1145/3422622.

[2] Hsiao-Tzu Hung et al. "EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation". In: *Proc. Int. Society for Music Information Retrieval Conf.* 2021.

[3] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical Reparameterization with Gumbel-Softmax". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: https://openreview.net/forum?id=rkE3y85ee.

[4] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.