**Narrative2Music: Generating Emotion-Aligned Music for Sentences**

**Introduction:** We introduce Narrative2Music, a novel dataset of emotionally aligned text-music pairs, and demonstrate its application in a Transformer model for generating emotionally aligned music for sentences. Previous to this work, an existing dataset of emotionally aligned text-music pairs was not available.

**Data:** We constructed a dataset of 1,078 text-music pairs by matching sentences from the GoEmotions dataset [1] with music files in EMOPIA [2]. EMOPIA has 1,078 symbolic music files in MIDI format labeled with an arousal-valence quadrant (see Fig. 1). GoEmotions has 58k sentences labeled with one of 27 different emotions, from which we pulled 1,078 sentences and mapped their labels to arousal-valence quadrants (Fig.1). We used GiantMIDI-Piano (10,855 MIDI files) [3] for pre-training.

**Models:** We demonstrated the utility of our Narrative2Music dataset by developing an encoder-decoder Transformer model for generating emotionally aligned music for text. For the encoder, we used pre-trained ModernBERT [4]. We used a Transformer decoder which we first pre-trained on GiantMIDI-Piano. We then fine-tuned the encoder-decoder model on our Narrative2Music dataset.

**Evaluation:** We generated music for 80 unseen sentences and calculated several objective metrics on the generated files for each quadrant. We calculated the ratio of major key to minor key files (capturing valence) and average note velocity, note density, and note length (capturing arousal). We then evaluated whether the trained model replicated EMOPIA's distribution of metrics in each quadrant.

**Results:** Figures 2 and 3 show valence and arousal metrics for EMOPIA and our generated files after epoch 0 (pre-training) and 200 (fine-tuning). Fine-tuning with Narrative2Music improved alignment of all metrics, though major key ratio and note velocity saw more room for improvement.

**Conclusions:** The promising performance of our Transformer model for generating emotionally aligned music for sentences indicates the effectiveness of our Narrative2Music dataset. We are currently testing longer narratives than single sentences and more epochs of fine-tuning to improve performance. A human-rater subjective evaluation of the generated files is also in progress.

**References:** [1] Demszky et al., 2020. GoEmotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547.

[2] Hung et al., 2021. Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. arXiv preprint arXiv:2108.01374

[3] Kong et al., 2020. Giantmidi-piano: A large-scale midi dataset for classical piano music. arXiv preprint arXiv:2010.07061.

[4] Warner, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv preprint arXiv:2412.13663

Figure 1. Arousal-Valence Quadrants

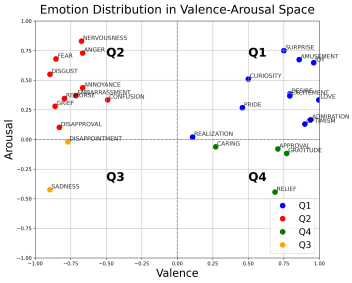


Figure 2. Valence Metric across Quadrants

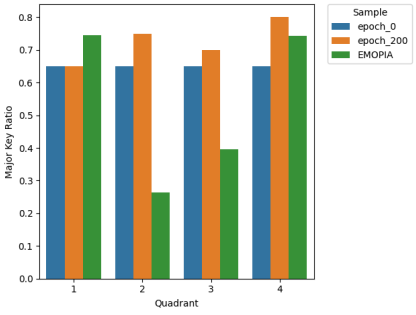


Figure 3. Arousal Metrics across Quadrants