IMU2Music: Learning to Generate Music from IMU Sensor Readings

Julian de Gortari Briseno

Overall Project Goals

Main goal: design a neural network that translates IMU readings of a person playing an instrument into actual soundwaves

Motivation

- Neural networks can perform cross-modality sequence translation tasks
- Previous attempts at mapping human motion to sound, disregarding IMU
- Can IMU replace microphones for music recording?

Specific Aims and Deliverables

1. Collect a dataset that captures both IMU readings and the audio recording of people playing an instrument.
2. Design a set of neural networks that translate these IMU readings into an audio representation.
3. Test and compare the performance of these neural networks.

Technical Approach

Dataset of violin recordings with IMU readings (~47 minutes)

Universitat Pompeu Fabra Barcelona

TELMI
Technology Enhanced Learning of Musical Instrument Performance.

Technical Approach

Music consists of recurrent elements and there is reference to elements way too in the past => use transformer

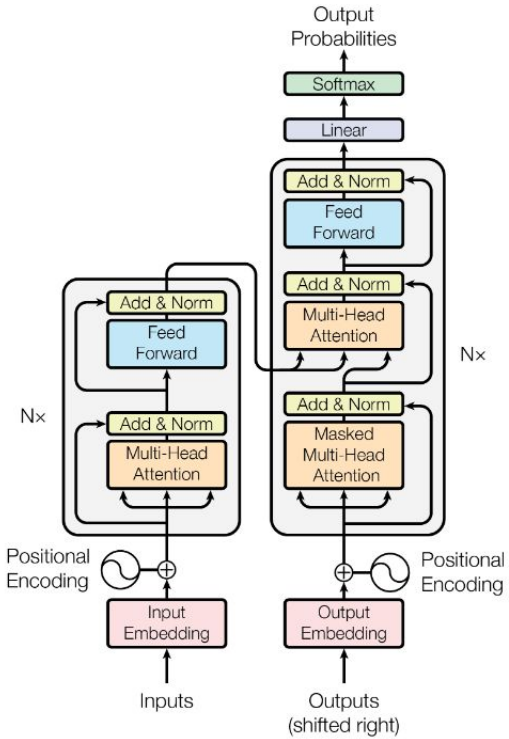


Figure 1: The Transformer - model architecture.

Image: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)

Technical Approach

- For decoder we use Musical Instrument Digital Interface (MIDI) representation.
- Includes timing, pitch and velocity (hardness of played note).
- Easier to learn.

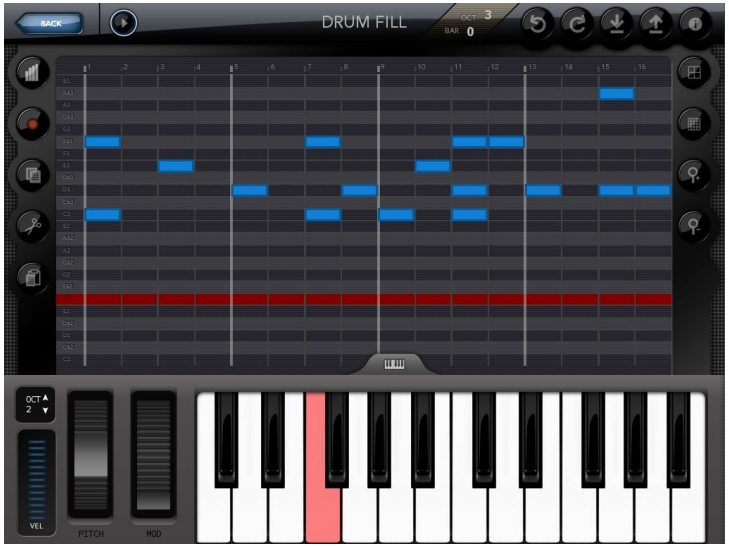


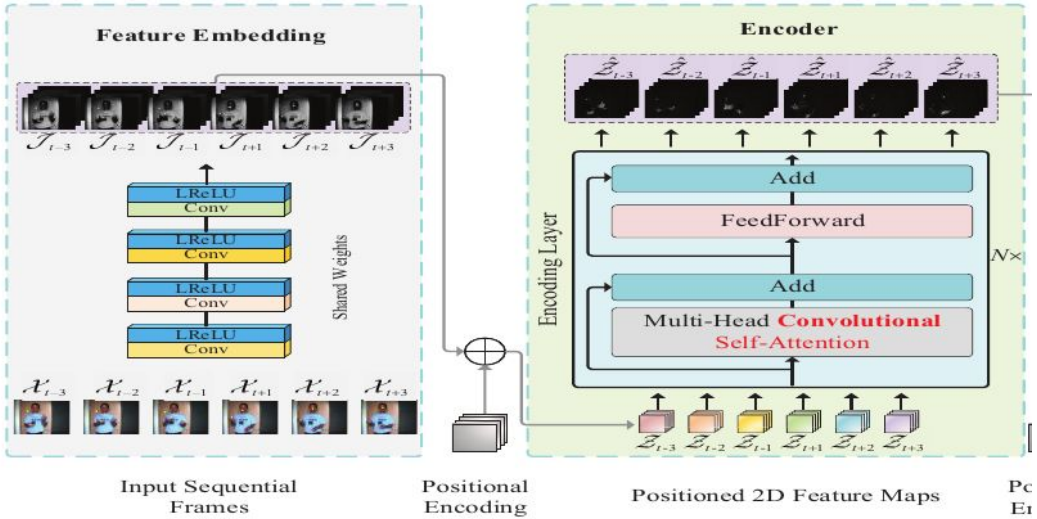
Image: https://static.kvraudio.com/i/b/gms3.jpg

Technical Approach

What to use for the encoder part?

- Normal transformer encoder with relative positional representation
- Use a ConvTransformer encoder

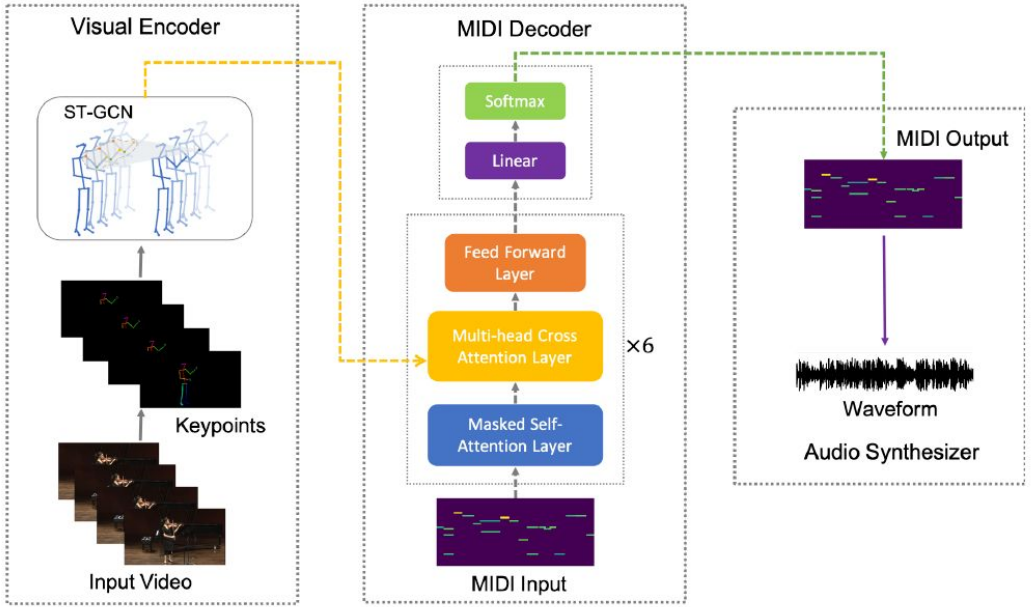
Play with feature embedding



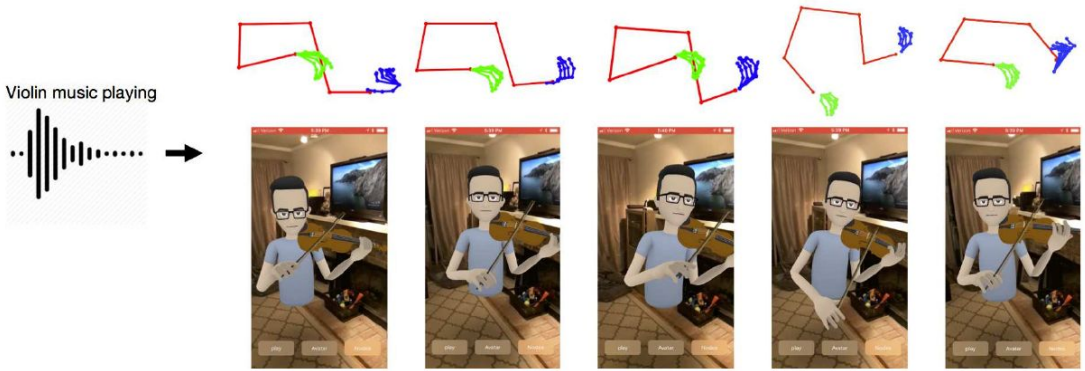
Zhouyong Liu and Shun Luo and Wubin Li and Jingben Lu and Yufan Wu and Chunguo Li and Luxi Yang (2020). ConvTransformer: A Convolutional Transformer Network for Video Frame Synthesis. https://arxiv.org/abs/2011.10185

Related Work

Chuang Gan and Deng Huang and Peihao Chen and Joshua B. Tenenbaum and Antonio Torralba (2020). **Foley Music: Learning to Generate Music from Videos**. https://arxiv.org/abs/2007.10984



Related Work



Eli Shlizerman, Lucio Dery, Hayden Schoen, Ira Kemelmacher-Shlizerman (2018). **Audio to Body Dynamics**. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7574-7583. https://openaccess.thecvf.com/content_cvpr_2018/html/Shlizerman_Audio_to_Body_CVPR_2018_paper.html

Current Status

- Dataset collected and processed
- Transformer code adapted to process IMU data
- Original transformer model trained, ready to be used for transfer learning

Next Steps

- Include ConvTransformer encoder code
- Test different network architectures
- Do some feature engineering