進捗報告

1 進捗

- 音楽生成周りの論文読み

2 Music Transfomer

Music Transfomer [1] とは, 2018 年に Google が発表した自動作曲の人工知能. 自然言語処理でよく用いられている Transformer を音楽に適用することにより, 高性能な音楽生成が可能となった.

Music Transfomer では, Attention 重みを用いることで逐次的に適切に前の音を参照しながら生成する手法を取っている. 以前に似た形の構造があれば強く Attention がかかるのでより音楽的な繰り返し構造の生成ができる.

3 Auto Foley

Auto Foley [2] とは, 2020 年にテキサス大学の研究チームが発表した, 映像に効果音をつける際の一連の作業を行うアルゴリズム及び生成モデル. AutoFoley は, あらかじめ用意された効果音を映像に当て込む. フレームから画像の特徴を抽出してこれに合った効果音を決定するモデルと, フレーム内オブジェクトのアクションを時系列的に分析するモデルによって音と映像の正確な同期を実現する. 学習にはよくあるアクションを映した短い映像と効果音を含んだ大規模データセットを利用しており, 論文が有料だったので解説記事しか読めていないが, おそらく MIT と IBM が発表している Moments in Time Dataset [3] と思われる.

4 STAIR Actions Dataset

STAIR Actions Dataset [4] とは, 人の様々なアクションを短い動画にした大規模データセットで,「誰が」「どこで」「何をしているのか」という日本語キャプションが一本の動画あたり平均 5 つ付いている.

5 テーマ決め

- どうにか, 自然言語処理を絡めた音楽生成タスクを決めたい.
- テキストキャプション付き動画 (STAIR Actions Dataset) に効果音を生成するタスクとかは実現可能でしょうか...?

参考文献

[1] 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.

[2] Sanchita Ghose and John J. Prevost. Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning. *IEEE Transactions on Multimedia*, page 1–1, 2020.

[3] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019.

[4] Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. Stair actions: A video dataset of everyday home actions. *arXiv preprint arXiv:1804.04326*, 2018.