Imitating Bach: Training a Diffusion Model to Compose Fugues

Andreas Kalavas

National Technical University of Athens

Motivation

- Experiment with Diffusion Models, and find strengths and limitations.
- Train an existing Model [1], with specific dataset (Fugues by Bach).
- I like music 🙂.

Dataset

A new method is outlined for generating binary image datasets from MIDI files. For the creation of the datasets, augmentation techniques are also being utilized.

Augmentation Techniques:

- **Overlapping** the images, so the ending of one image aligns with the beginning of the next one.
- **Transposing** the pieces, to facilitate the spread of melodies across a broader range of pitches.
- **Reducing the Tempo** of the pieces, to generate additional segments without introducing an excessive volume of new information into the model.

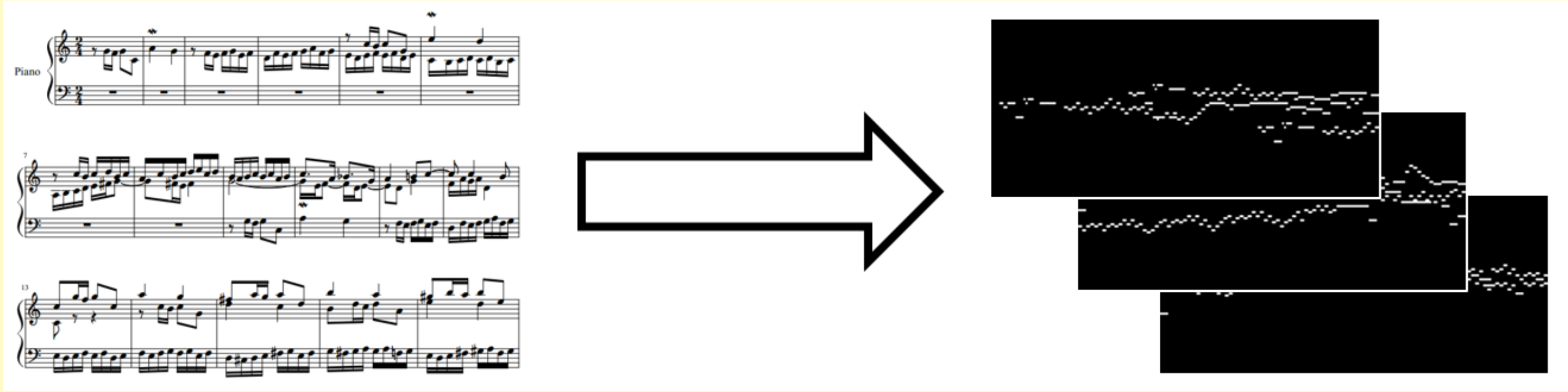


Figure 1. Dataset Creation, from MIDI files to binary images.

Training Dataset

For the training of the Model, the 24 fugues of the first Well-Tempered Clavier book by Bach were used, and yielded a dataset of 2262 binary images of size 192x88 pixels.

Model

Efficient Forward Process

A forward process that solely depends on the input x_0. This results in efficiently computing the output at step t, thus enabling the model to generate them dynamically.

$$a_t = 1 - b_t \qquad \bar{a}_t = \prod_{s=1}^{t} a_s \qquad q(x_t|x_0) = B(x_t; \bar{a}_t x_0 + (1-\bar{a}_t)0.5)$$

Sampling Algorithms

The model is tested on the tasks of Unconditional Generation and Infilling. The algorithms used are shown below.

```
Algorithm 1: Generating new samples
Input: A piano roll sampled from a
  binomial distribution $x_T$
for $t = T, T-1, \ldots, 1$ do
    $\hat{x}_0 = \text{UNet}(x_t)$
    $\delta = x_T \bigoplus \hat{x}_0$
    mask $\sim B(\delta\beta_t)$
    $x_{t-1} = \hat{x}_0 \bigodot (1-\text{mask}) + x_t \bigodot \text{mask}$
end
return $x_0$
```

```
Algorithm 2: Infilling samples
Input: A fixed region $r$, a piano roll
  sampled from binomial distribution $x_T$
fix r in $x_T$
for $t = T, T-1, \ldots, 1$ do
    $\hat{x}_0 = \text{UNet}(x_t)$
    $\delta = x_T \bigoplus \hat{x}_0$
    mask $\sim B(\delta\beta_t)$
    $x_{t-1} = \hat{x}_0 \bigodot (1-\text{mask}) + x_t \bigodot \text{mask}$
    fix r in $x_{t-1}$
end
return $x_0$
```

Architecture

Within the steps of the Diffusion Model, a UNet model [2] is incorporated.

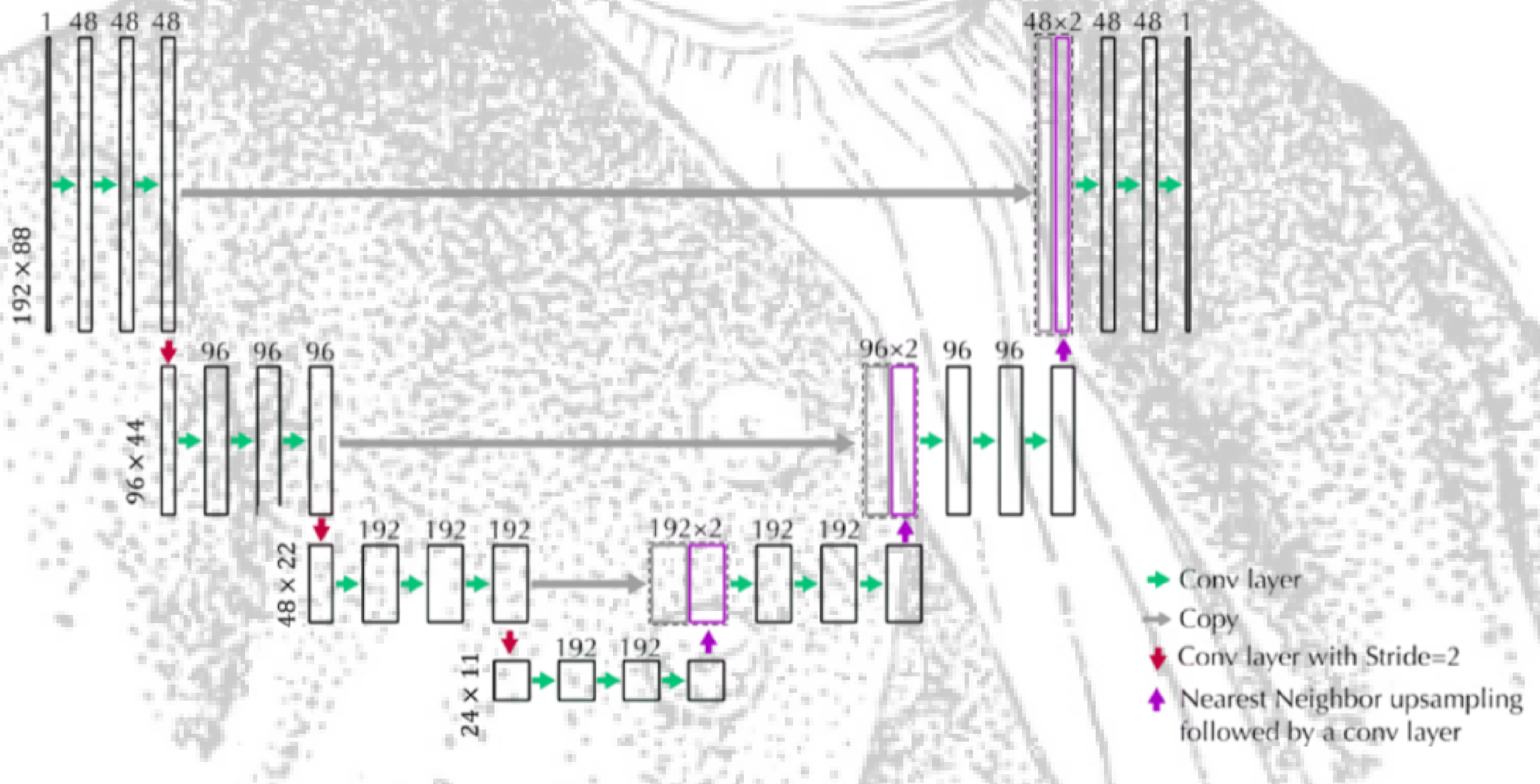
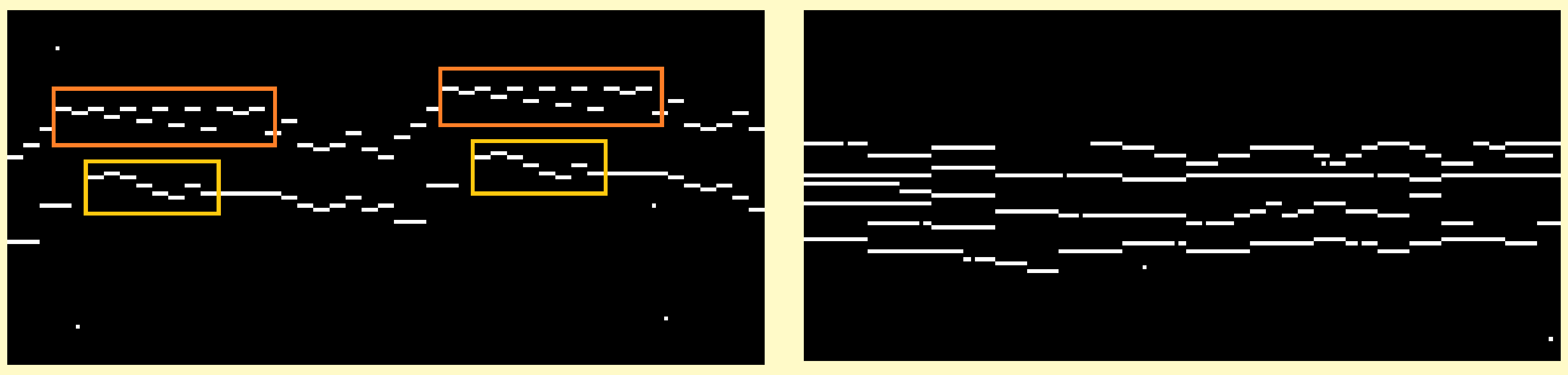


Figure 2. UNet Architecture. Taken from [1]

Results

Unconditional Generation

In these generated samples we can witness the ability of the model in generating from single voice to multi voice piano rolls. The first piano roll is like the beginning of a fugue, in the second we can see that the model repeated motifs, and in the last we observe the polyphonic texture.

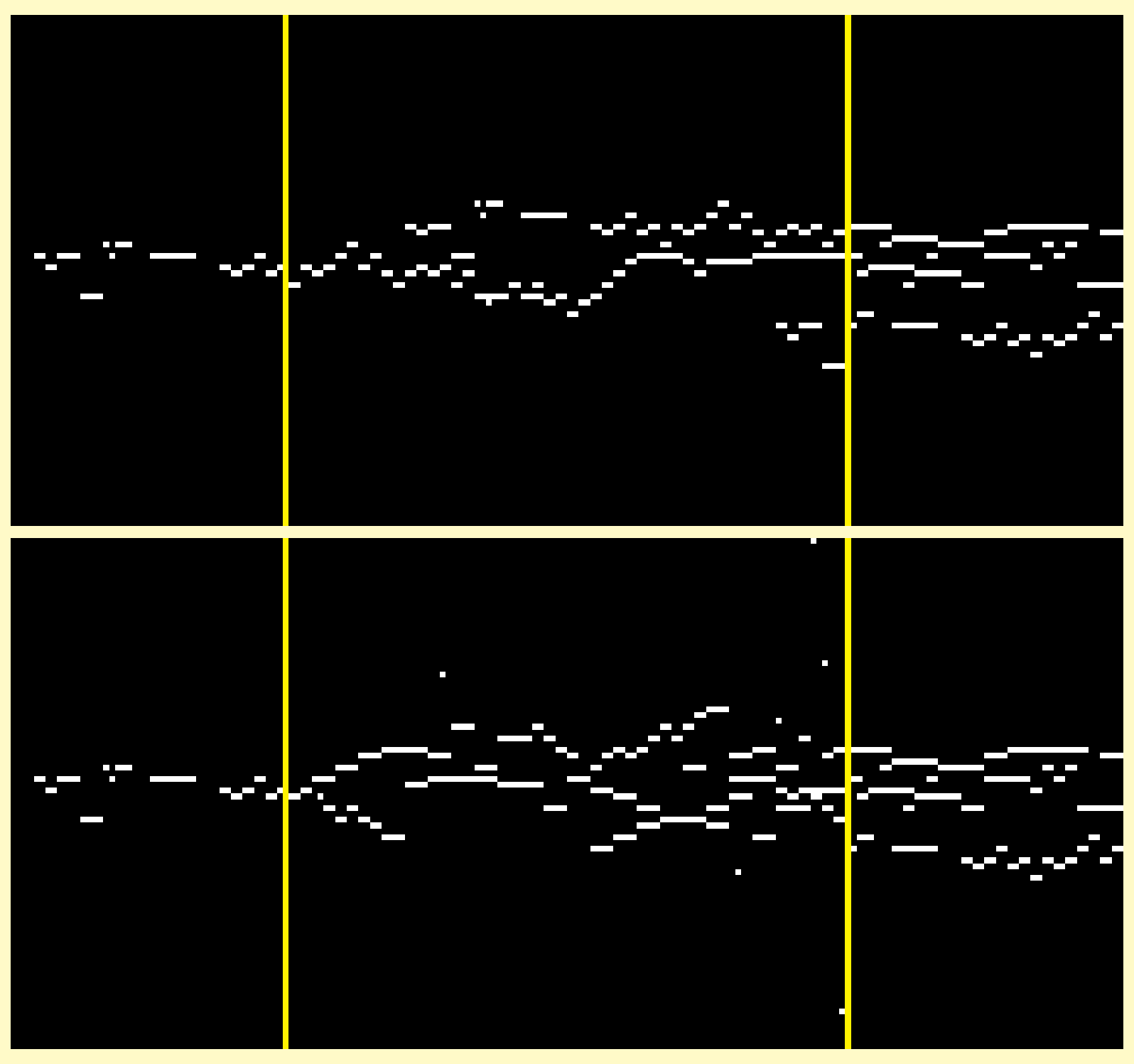


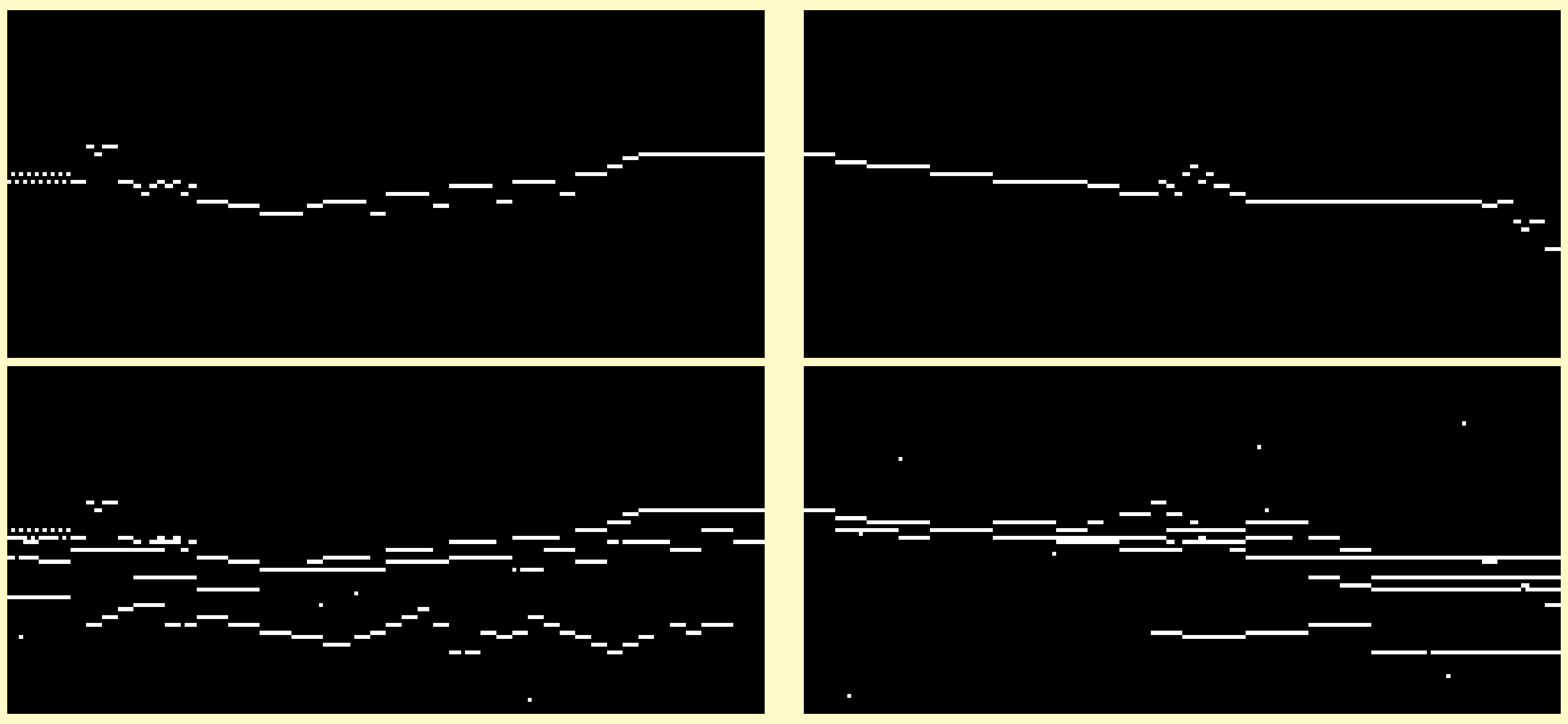
Infilling

Our model was tested in three ways in the task of infilling:

- Infill in the time domain by fixing the beginning and the ending of the image (top right).
- Infill with the top voice given by fixing the melody and all the area above (bottom left).
- Infill with the middle voice given by just fixing the melody (bottom right).

For each case, the original and the infilled samples are shown.





Limitations

- No efficient statistical metrics for evaluation [3].
- Heavily compacted pieces in time axis.
- Overfit on the only two - voiced fugue (Fugue no. 10 in E minor).
- Sources - needed 130 hours for a dataset of 2262 images.
- Loss of many musical elements.
- Questionable comprehension of the model, as it sees images.

GitHub

Conclusion

In our work, we introduced a novel dataset and outlined a method for generating binary image datasets from MIDI files and vice versa. Subsequently, we employed this dataset to train an existing Diffusion Model with UNet architecture, and then utilizing the trained model to generate samples. The model underwent testing in the tasks of unconditional generation and infilling, yielding results that exhibited similarity to the input dataset. Additionally, we shared our reflections on our work and addressed certain limitations inherent in this approach to music generation.

References

[1] L. Atassi, "Generating symbolic music using diffusion models," University of California, 2023. arXiv:2303.08385

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[3] M. Plasser, S. Peter, and G. Widmer, "Discrete Diffusion Probabilistic Models for Symbolic Music Generation," IJCAI, 2023. arXiv:2305.09489

andreaskalavas@gmail.com