Visual Features and Style Classification with Convolutional Autoencoders

Matt Chiu, Assistant Professor of Music Theory,
Baldwin Wallace University, mchiu@bw.edu

Format preference (oral or poster):
Poster preference and interested in submitting an extended paper to JNMR (if accepted).

Abstract

There is a considerable amount of computational work dedicated to studying, differentiating, and classifying musical styles. Many of these papers analyze the behavior of symbolic representations or audio representations of music (e.g., Chiu and Temperley, 2024; Conklin & Witten, 1995; McKay & Fujinaga, 2006; Pons, Lidy, & Serra, 2017). This paper approaches style through *optical music representation* (OMR), or the way music *looks*. That is, this paper studies the visual aspects of style by building a *deep convolutional autoencoder* (CAE)—a type of neural network used primarily in computer vision (Masci et al., 2011).

We first build a dataset of piano sheet music and songs comprised of different styles and time periods. The dataset contains 6 musical periods/styles (Baroque, Classical, Romantic, Impressionist, Ragtime, Pop/Rock), where each style had at least 3 composers/artists, and a minimum of 30 pages for each composer/artist. There are a total of 1285 pages in the dataset, each tagged with composer and style in the metadata. For each style, 8 random samples (448x448 pixels) are extracted from each page, yielding a dataset of 4,320 snapshots.

A semi-supervised CAE (Figure 1) with a classification head is trained on this dataset, leveraging the unsupervised generalizability of autoencoders with the guided aim of classifying pieces, composers, and styles (heavily weighting style classification). After the model is trained, we can study the features of the model, investigating which visual features are most prioritized. In the model, certain neurons respond to specific higher-order features, including chord inversions, key signatures, scalar passages, note density, text, etc. Clustering samples shows that samples are grouped primarily by texture—including sparse, polyphonic, and homophonic textures. Instances of false positives/negatives for composer classification are inconsistent, suggesting that the generic hand shapes in scores make it difficult to differentiate between one composer/artist at a local level.

Word Count: 299

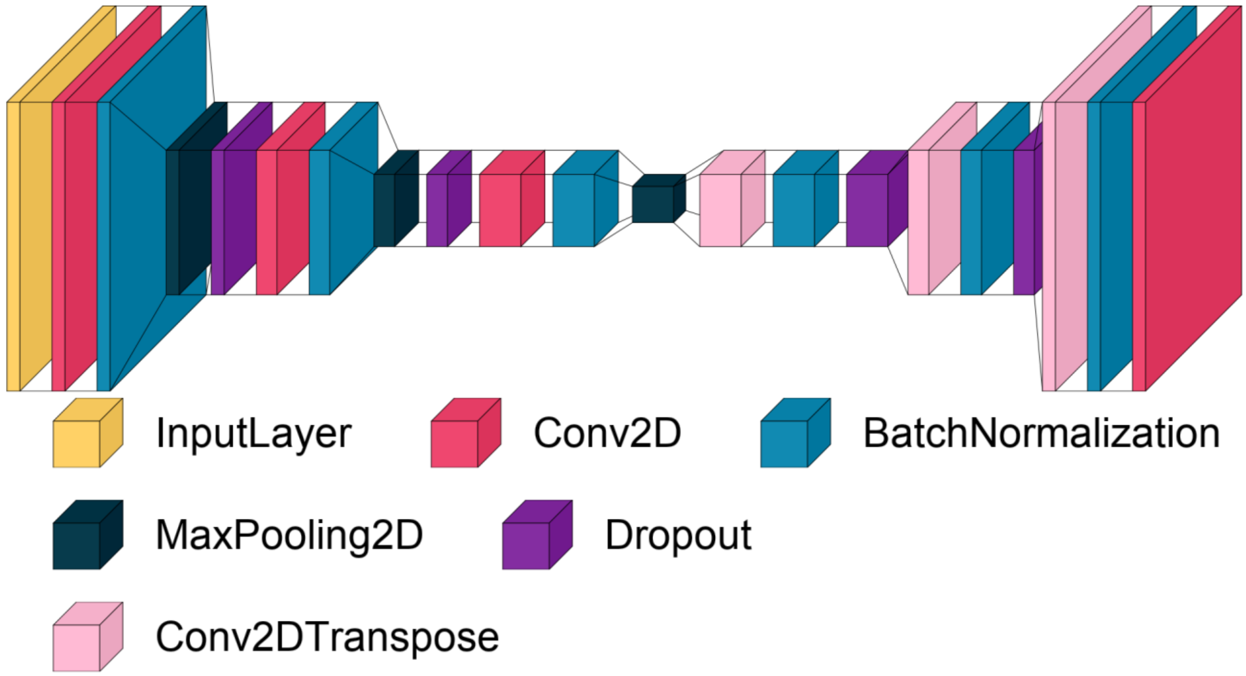

Figure 1. Autoencoder Convolutional Neural Network Architecture (classification heads not included).

References

Balke, S., Achankunju, S. P., & Müller, M. (2015). Matching musical themes based on noisy OCR and OMR input. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 703–707. https://doi.org/10.1109/ICASSP.2015.7178060

Byrd, D., & Simonsen, J. G. (2015). Towards a standard testbed for optical music recognition: Definitions, metrics, and page images. *Journal of New Music Research, 44*(3), 169–195. https://doi.org/10.1080/09298215.2015.1045424

Chiu, M., & Temperley, D. (2024). Melodic Differences Between Styles: Modeling Music with Step Inertia. *Music and Science* 7. https://doi.org/10.1177/20592043231225731

Conklin, D., & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research, 24*(1), 51–73. https://doi.org/10.1080/09298219508570672

Hajic Jr, J., Pecina, P., et al. (2017). The MUSCIMA++ dataset for handwritten optical music recognition. In *International Conference on Document Analysis and Recognition (ICDAR)* (pp. 39–46). https://doi.org/10.1109/ICDAR.2017.16

Herremans, D., Chuan, C. H., & Chew, E. (2017). A functional taxonomy of music generation systems. *ACM Computing Surveys (CSUR), 50*(5), 1–30. https://doi.org/10.1145/3108242

Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks* (pp. 52–59). Springer. https://doi.org/10.1007/978-3-642-21735-7_7

McKay, C., & Fujinaga, I. (2006). Musical genre classification: Is it worth pursuing and how can it be improved? In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR 2006)*.

Pons, J., Lidy, T., & Serra, X. (2017). Experimenting with musically motivated convolutional neural networks. In *14th International Workshop on Content-Based Multimedia Indexing* (pp. 1–6). https://doi.org/10.1109/CBMI.2016.7500246

Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A. R. S., Guedes, C., & Cardoso, J. S. (2012). Optical music recognition: State-of-the-art and open issues. *International Journal of Multimedia Information Retrieval, 1*(3), 173–190. https://doi.org/10.1007/s13735-012-0004-6

Sigtia, S., Kumar, S., Mauch, M., Benetos, E., & Dixon, S. (2016). *An end-to-end neural network for polyphonic piano music transcription. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(5), 927–939. https://doi.org/10.1109/TASLP.2016.2535164