Translating Visual Art into Music

Maximilian Müller-Eberstein
University of Amsterdam
edu@personads.me

Nanne van Noord
University of Amsterdam
n.j.e.vannoord@uva.nl

Abstract

The Synesthetic Variational Autoencoder (SynVAE) introduced in this research is able to learn a consistent mapping between visual and auditive sensory modalities in the absence of paired datasets. A quantitative evaluation on MNIST as well as the Behance Artistic Media dataset (BAM) shows that SynVAE is capable of retaining sufficient information content during the translation while maintaining cross-modal latent space consistency. In a qualitative evaluation trial, human evaluators were furthermore able to match musical samples with the images which generated them with accuracies of up to 73%.

1. Introduction

Art is experienced as a flow of information between an artist and an observer. Should the latter be visually impaired however, a barrier appears. One way to overcome this obstacle might be to translate visual art, such as paintings, from an inaccessible sensory modality into an accessible one, such as music.

Our research builds upon single-modality generative models for images [2, 3] and music [8] as well as on multimodal models which leverage corresponding audio-visual data in order to learn denser information representations [4] or to make visual information more accessible [1]. Furthermore, generative models have been used to measure the expressiveness of image-based audio generation tools for the visually impaired [5] and as such they offer a solid basis for our approach.

While previous approaches make use of paired audio-visual ground truth data, these are unavailable for visual art and music, making an unsupervised approach necessary. The main focus of this research will therefore be to learn a consistent cross-modal mapping in the absence of paired data such that similar images produce similar music. Evaluation metrics must be able to quantitatively reflect consistency and retained information content while a qualitative evaluation with humans must measure whether such consistencies are actively perceived.

To summarize, the main contributions of this research are as follows:

- With the Synesthetic Variational Autoencoder (SynVAE), we introduce an unsupervised cross-modal architecture for translating data from one sensory modality into another consistently without the need for subjectively paired ground truth datasets (see Section 2.1).

- In a series of experiments on generating music from the MNIST [6] and Behance Artistic Media (BAM) [9] datasets, we compare a variety of mutual information metrics in order to establish a quantitative basis for evaluating such cross-modal models.

- In a qualitative study based on these quantitative metrics, we evaluate the human perception of the cross-modal translation consistency and lay out a framework for avoiding subjective biases within this process (see Sections 2.2 and 3).

2. Methodology

Leveraging the well established visual β-VAE architecture (VisVAE) [3] and the auditive MusicVAE [8], we are able to construct the Synesthetic VAE (SynVAE) (see Section 2.1). Additionally, evaluating SynVAE with respect to cross-modal consistency requires a diverse set of methods which are outlined in Section 2.2.

2.1. Synesthetic Variational Autoencoder

Translating information across the audio-visual modal boundary requires a synesthetic approach. By using multiple single-modality models which have unrestricted access to high quality data in their respective domains, we are able to remove the need for subjective correlations of images and music and are able to construct a fully unsupervised architecture which is outlined in Figure 1.

Initially, the visual encoder $p_{\text{venc}}(\boldsymbol{z}_v|\boldsymbol{x})$ creates a 512 dimensional latent representation $\boldsymbol{z}_v$ from the original image $\boldsymbol{x}$. It provides the initial state of the pre-trained MusicVAE decoder $p_{\text{adec}}(\boldsymbol{a}|\boldsymbol{z}_v)$ which then produces a melodic output $\boldsymbol{a}$ using its hierarchical architecture. These two components
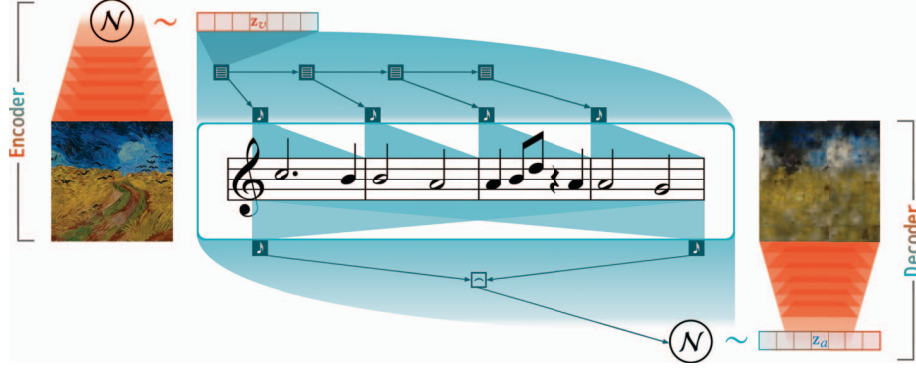
Figure 1. Synesthetic VAE Architecture. An image is first encoded into a latent vector z_v by the VisVAE encoder, before being decoded into music by the MusicVAE decoder. During training, the music is subsequently re-encoded into z_a by the MusicVAE encoder and reconstructed into the output image by the VisVAE decoder.

make up the overall synesthetic encoder $p_{\text{senc}}(a|x)$ which, during inference, can be used to perform the audio-visual translation.

In order to obtain a differentiable loss formulation, the audio output is subsequently re-encoded as z_a using the pre-trained, bidirectional MusicVAE encoder $p_{\text{aenc}}(z_a|a)$ and then passed through VisVAE's decoder $p_{\text{vdec}}(x'|z_a)$ to produce an image reconstruction x'. These two components make up the synesthetic decoder $p_{\text{sdec}}(x'|a)$.

Both the visual and auditive latent vectors z_v and z_a in this model are sampled from multivariate Normal distributions that share the same regularizing prior distribution $\mathcal{N}(\mathbf{0}, I)$. This actively encourages the latent spaces to follow a similar and consistent shape across modalities when compared to an unregularized training procedure and is therefore at the core of our unsupervised approach.

Since this architecture requires an expressive musical latent space with high variability, the weights of the MusicVAE model remain fixed throughout the entire training process. Additionally, its pre-conditioning on the $\mathcal{N}(\mathbf{0}, I)$ prior alleviates us from the need to enforce a regularizing KL constraint on the frozen auditive components. Furthermore, audio reconstruction need not and cannot be measured due to the general absence of an audio-visual ground truth. The differentiable basis $\mathcal{L}_{\text{syn}}$ for the optimisation process therefore only consists of a KL constraint on the visual encoder, in addition to the difference of the synesthetic decoder's reconstruction with the original image:

$$\mathcal{L}_{\text{syn}} = - \mathbb{E}_{a \sim p_{\text{senc}}(a|x)}[\ln p_{\text{sdec}}(x|a)] \\ + \beta \, \text{KL}(p_{\text{venc}}(z_v|x) \, \| \, p_{\text{prior}}(z_v)) \quad (1)$$

with β controlling the balance between reconstruction quality and adherence to the canonical prior. In the synesthetic case, this carries additional importance since the actively trained visual components cannot stray too far from

the prior without risking to land in undefined music space.

Extending this approach to further modalities is low friction. For instance, switching the auditive and visual components results in a SynVAE which encodes music into corresponding visuals. Regardless of the modality pair, as long as the encoders are regularized with matching prior distributions and the central generative component is capable of producing realistic results across its latent space, the cross-modal translation is likely to succeed. Since advances in single-modality models are being made constantly, this architecture could keep on improving in parallel to such generative models provided they offer consistent latent spaces.

2.2. Evaluation

Apart from the quantitative metrics already present in the loss formulation $\mathcal{L}_{\text{syn}}$ (*i.e.* MSE for reconstructions' information retention and KL divergence as a proxy for latent space consistency), labels of the visual datasets can be used to measure how well latent representations encode semantic similarity. By collecting the nearest neighbours of each encoded data point, it is possible to calculate the class precision at rank n. Since the higher inner-class visual variance of more complex datasets may not be well reflected however, we attempt to bridge this gap with reconstruction classification accuracy. Using the same architectures as for the visual encoders, but with a final softmax layer, classification networks are trained and tested on the images reconstructed by the autoencoders.

We are also strongly interested in the degree to which the visual and auditive latent spaces within the synesthetic model overlap. In absence of paired images and audio, we make use of Data-Efficient Mutual Information Neural Estimation (DEMINE) [7] in order to approximate a lower bound on the mutual information $I(\mathbf{Z}_v; \mathbf{Z}_a)$ of corresponding visual and auditive latent vectors.

Considering SynVAE's potential application in cultural

Model	MSE	KL	P10	Acc	MI	Ql
M-Vis	16.43	24.38	0.31	0.99	-	-
M-Syn	36.66	15.63	0.28	0.96	5.03	0.73
B-Vis	273.08	56.34	0.23	0.80	-	-
B-Syn	455.16	27.89	0.25	0.77	5.16	0.71

Table 1. MSE, KL divergence, Precision@10 (P10) and classification accuracy (Acc) for VisVAE and SynVAE models on MNIST (M) and BAM (B) test sets. Mutual information lower-bound (MI) and evaluator accuracy on qualitative task (Ql) for SynVAEs.

settings such as art exhibitions, it is vital to evaluate whether humans can correlate similar images and audio in the same way as the model would predict. Since evaluating the entire corpus is not feasible and determining a sub-sample manually would introduce the curator's bias, we propose a more consistent and reproducible approach: By using VisVAE's latent space to sample images embedded closely around the centroids of their semantic classes, evaluators can be presented with a smaller subset of representative images and subsequently, their related audio. Through the accuracy with which they can identify correct audio-visual pairs, it is possible to determine whether the model's translations line up with human notions of similarity and intuitiveness.

3. Experiments

Our experiments involve the simple, but interpretable $28 \times 28 \times 1$ MNIST images [6], as well as the highly diverse BAM [9] of which we use ~180k oil and watercolour paintings, downscaled and cropped to $64 \times 64 \times 3$. The latter's four emotion labels are of especial interest for the qualitative evaluation due to their abstract nature and reliance on shared human intuition. Additionally, we highly recommend listening to selected audio-visual examples from each dataset on https://personads.me/x/synvae (full source code is also available).

Both tasks share a similar pipeline: Visual β-VAEs are trained and evaluated before they are placed into SynVAE. It too is then trained, validated and tested on the respective splits of the same data using the methodology outlined in Section 2.1. For each task, β-values in $[0.1, 2.0]$ were tested in a grid search pattern. Results for the $\beta = 0.5$ MNIST and the $\beta = 1.3$ BAM models are presented in Table 1.

From MSE and KL divergence, it is observable that passing visual information through music space results in a definite reduction in visual fidelity since VisVAEs have reconstruction errors which are close to half of that of their synesthetic counterparts. Regarding the adherence to the canonical prior however, the tables are turned: the SynVAEs have KL terms which are up to 61% smaller. This stems from the strong regularizing effect of the fixed auditive components.

Measuring our main goal of audio-visual consistency was shown to be difficult using a single quantitative met-

ric. Precision metrics based on the nearest neighbours of an embedded data point were indicative of consistency only if visual similarity was strongly correlated with labelled semantic similarity. For MNIST this was found to be the case since the monochrome nature of the dataset as well as the relatively low variance between images of the same class allow for MSE to be an appropriate surrogate for the detailed, low-level information content of an image.

For the more visually complex BAM with its exponentially larger amount of information as well as a higher in-class variance, it becomes more difficult for SynVAE to encode such lower-level details. Therefore not all visual details are retained when they are passed through SynVAE's auditive latent space and image reconstructions are limited to higher-level features such as overall colour and object placement. However, the emotion labels tend to share a larger degree of correlation with such features (*e.g.* dark images being "scary") and allow for P@10 scores of around 0.25.

Reconstruction classification provides a more flexible way to measure cross-modal consistency. It relies on semantic labels as well, but has the benefit that it is indicative of whether same-class image data is consistently encoded and decoded. Across all experiments a drop in accuracy from VisVAE to SynVAE showed that while there is a certain degree of information loss when passing through latent spaces cross-modally, overall visual consistency of semantic classes is maintained. Reconstructed MNIST digits even bear a strong enough semblance to the original images to warrant a 0.96 accuracy even after passing through music space. SynVAE's less detailed BAM reconstructions also retain sufficient higher-level information in order for multiclass emotion to be identified with 0.77 accuracy.

Independent of labelled data, DEMINE measures a relatively high amount of mutual information between corresponding visual and auditive latent vectors ($\ln(10) \approx 2.3$ nats would for instance already be sufficient to encode MNIST's class information). This shows that SynVAE does indeed learn to embed information consistently across modalities and different types of visual data. Combining these three quantitative metrics, it becomes apparent that information content is indeed being translated across modalities consistently.

To assess whether human evaluators would share this notion of consistency, we conducted a classification study for MNIST and BAM with 11 and 21 participants respectively using the method described in Section 2.2. After being presented with 4 audio-visual example pairs for each of the 3 most distinct classes (as determined by VisVAE), evaluators were asked to identify which of 3 images was used to generate an audio across 20 trials.

For MNIST, evaluators achieve 0.73 accuracy ($\sigma = 0.22$) and a consistent Fleiss' kappa of 0.48 when distin-
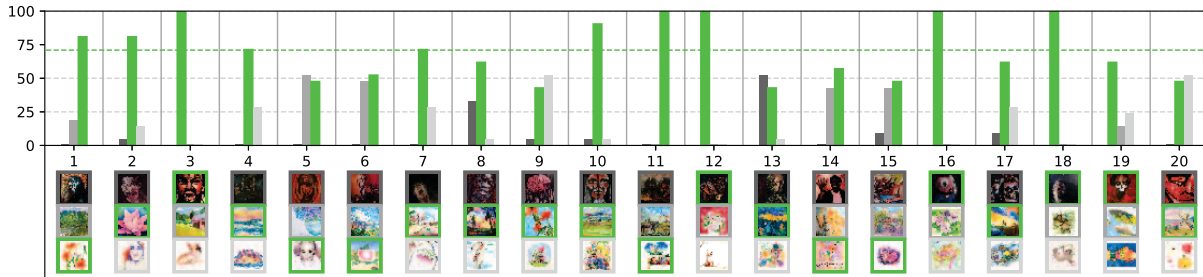
Figure 2. Percentages of evaluator choices per class ("scary", "happy+peaceful", "happy") and task on the qualitative BAM evaluation. Correct options highlighted in green. Average accuracy marked at 71%.

guishing between the digits "0", "1" and "4". In two tasks, evaluators unanimously made the correct connection. For the BAM classes "scary", "happy" and "happy+peaceful", accuracy is at a comparable 0.71 ($\sigma = 0.13$) with a Fleiss' kappa of 0.46. The results in Figure 2 further show how four "scary" images and one "happy" image were correctly identified by all evaluators. Considering that class-adherence for BAM emotions is not as easily identified as for MNIST digits, the level of agreement across tasks is remarkably high. This extends to mismatches such as in tasks 5, 6 and 13 as well, since they typically occur between the correct image and its visually closest alternative, but rarely for the most distinct incorrect option. This corroborates that the audio-visual translations of SynVAE are indeed consistent.

The high accuracy with which the evaluators were able to distinguish between the three most distinct classes of two very different datasets by ear alone, shows that low-level information, such as digits, can be conveyed audibly for simple data and high-level information, such as emotion perceived through colour and composition, can be conveyed for more complex data. This in addition to the quantitative results confirms that audio-visual consistency is not only theoretical, but also very perceivable.

4. Conclusion

As shown by our results, it can be concluded with high confidence that SynVAE is able to consistently translate a diverse range of images into the auditory domain of music through unsupervised learning mechanisms. The modular nature of SynVAE furthermore allows for this approach to be extended to any modality for which high quality single-modality datasets exist. We therefore hope that the methodology outlined in this research will provide a solid basis for evaluating unsupervised, cross-modal models, in addition to SynVAE itself enabling more intuitive and inclusive access to visual artworks across sensory boundaries.

References

[1] L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357. ACM, 2017. 1

[2] P. K. Diederik and M. Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1

[3] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017. 1

[4] W.-N. Hsu and J. Glass. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*, 2018. 1

[5] D. Hu, D. Wang, X. Li, F. Nie, and Q. Wang. Listen to the image. *arXiv e-prints*, page arXiv:1904.09115, Apr 2019. 1

[6] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1, 3

[7] X. Lin, I. Sur, S. A. Nastase, A. Divakaran, U. Hasson, and M. R. Amer. Data-efficient mutual information neural estimator. *CoRR*, abs/1905.03319, 2019. 2

[8] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck. A hierarchical latent vector model for learning long-term structure in music. *CoRR*, abs/1803.05428, 2018. 1

[9] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie. BAM! The behance artistic media dataset for recognition beyond photography. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 3