Sound Generation via Cross-Modal Transcoding

Alessandro Tibo ALESSANDRO.TIBO@STUD.UNIFI.IT
DINFO, Università degli Studi di Firenze, Via di S. Marta 3, 50139 Firenze, Italy

Simone Conforti SIMONE.CONFORTI@MUSSTDESIGN.COM
MUSST Design, 8400 Winterthur ZH, Switzerland

Lorenzo Brusci LORENZO.BRUSCI@MUSSTDESIGN.COM
MUSST Design, 8400 Winterthur ZH, Switzerland

Tijn Borghuis V.A.J.BORGHUIS@TUE.NL
Technische Universiteit Eindhoven, Postbus 513, 5600 MB Eindhoven, The Netherlands

Marco Gori MARCO.GORI@UNISI.IT
DIISM, Università degli Studi di Siena, Via Roma, 56 53100 Siena, Italy

Paolo Frasconi PAOLO.FRASCONI@UNIFI.IT
DINFO, Università degli Studi di Firenze, Via di S. Marta 3, 50139 Firenze, Italy

Abstract

We focus on the problem of cross-modal data mapping in a purely unsupervised setting. Our motivating application is the generation of sound or music from complex input sensory data (such as images, video, or text). We present preliminary empirical results on text-image data association and on a realistic musical scenario where the sonic parameters of a synthesiser (such as timbre, envelope, pitch and volume) are driven by a transcoded version of input textual data. Examples of generated sound can be found at `http://ai.dinfo.unifi.it/transcoder/`

1. Introduction

Learning with multimodal data (such as images and text, or video and audio) has recently received considerable attention, particularly within the deep learning community. For example Ngiam *et al.* (2011) trained bimodal deep autoencoders to learn joint features of audio and video for classification purposes. Srivastava & Salakhutdinov (2012) have applied multimodal deep Boltzmann machines to image tagging. Other authors have worked in the context of cross-modal retrieval introducing coupled architectures for multimodal hashing (Masci et al., 2014) or for learning correlated features of two data modalities (Feng et al., 2015).

Proceedings of the Constructive Machine Learning workshop @ ICML 2015.

All these methods employ some form of supervision in the learning process, either in the form of class labels or in the form of known associations between data points belonging to different modalities. In this paper, we focus on the problem of generating sound from data collected in a different modality such as text, images, or video, in a completely unsupervised fashion. While our approach is not suitable where exact denoting and pertinent associations are required, it finds a natural application in the context of music generation, where similar representative and expressive input sensory data should originate similar representative and expressive sonic perceptual experiences, but a precise cross-modal association is unnecessary. This could find applications in areas such as sound and music compositional assistive technologies, automatic music generation in gaming, design of software instruments, generation of music contents in commercial and public platforms.

2. Cross-modal data transcoding

We are given two data sets $X_{\text{in}} \in \mathbb{R}^{n_{\text{in}} \times p_{\text{in}}}$ and $X_{\text{out}} \in \mathbb{R}^{n_{\text{out}} \times p_{\text{out}}}$ of examples, where subscripts denote input and output modalities. No association between input and output examples is known in our setting. Individual examples are assumed to be sampled independently from their unknown distributions. As it is the case in most dimensionality reduction settings, we assume that data in both modalities lie on low-dimensional manifolds of intrinsic dimensionalities $k_{\text{in}} \ll p_{\text{in}}$ and $k_{\text{out}} \ll p_{\text{out}}$. In order to map data points x_{in} and x_{out} onto their low-dimensional representations $\phi_{\text{in}}(x_{\text{in}}) \in \mathbb{R}^{k_{\text{in}}}$ and $\phi_{\text{out}}(x_{\text{out}}) \in \mathbb{R}^{k_{\text{out}}}$, we train two

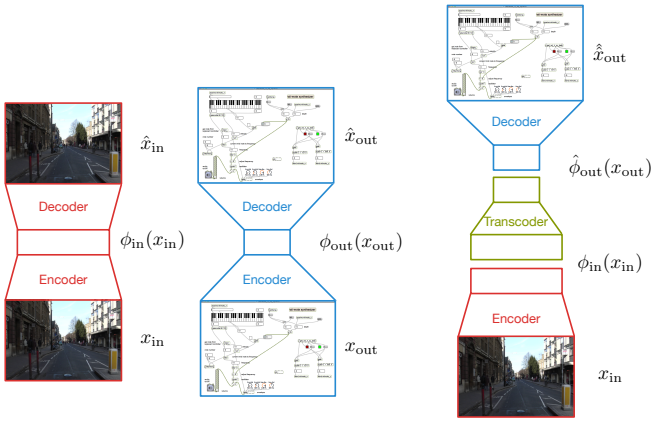

Figure 1. Left: input autoencoder; middle: output autoencoder; right: transcoding pipeline.

deep autoencoders. Stacked RBMs (Hinton & Salakhutdinov, 2006) or denosing autoencoders (DAEs) (Vincent et al., 2010) are possible approaches for this step. We then learn the mapping between $\phi_{\text{in}}(x_{\text{in}})$ and $\phi_{\text{out}}(x_{\text{out}})$ using a third *transcoding* network, as illustrated in Figure 1. The transcoder's weights are learned using the Stress-1 loss, commonly used in multi-dimensional scaling (Borg & Groenen, 2005):

$$J(W_T) = \frac{\sum_{i<j}^{n_{\text{in}}} \left(D_{\text{in}}^{(ij)} - D_{\text{out}}^{(ij)}\right)^2}{\sum_{i<j}^{n_{\text{in}}} \left(D_{\text{in}}^{(ij)}\right)^2} \tag{1}$$

where

$$D_{\text{in}}^{(ij)} = \left\| \phi_{\text{in}}(x_{\text{in}}^{(i)}) - \phi_{\text{in}}(x_{\text{in}}^{(j)}) \right\| \tag{2}$$

$$D_{\text{out}}^{(ij)} = \left\| \hat{\phi}_{\text{out}}(x_{\text{out}}^{(i)}; W_T) - \phi_{\text{in}}(x_{\text{out}}^{(j)}) \right\| \tag{3}$$

3. Music synthesizer

To demonstrate the relationship between text analysis and sound generation through transcoding, we developed an application in the visual programming language Max MSP (Puckette, 2002). Its core is an additive synthesizer consisting of 32 sine wave oscillators which are individually controllable in frequency and amplitude. Such a synthesizer is capable of generating a wide variety of timbres, with the first oscillator producing the fundamental frequency and the others contributing subsequent partials to the sound signal (Roads, 1996). In our system, the fundamental pitch is defined by the notes of a tempered scale and the ratio between the frequencies of the oscillators is adjustable, it can be harmonic as well as inharmonic. The synthesizer has 34 parameters: the fundamental frequency, the ratio between the oscillator frequencies, and an amplitude for each of the oscillators. Hence, every sound it can produce is completely characterized by an array of 34 values.

4. Experimental results

A proper assessment of the proposed methodology should involve human subject in order to test their ability to discern similarities in the musical output due to similarities in the input signal, and also to verify that the result is emotionally effective. In the field of cross-modal assocation studies (Deroy & Spence, 2013), several approaches exist in the literature for this purpose (Parise & Spence, 2012; Chan & Dyson, 2015). While such a study is beyond the scope of the current preliminary contribution, in this Section we also want to provide a quantitative empirical evaluation of the algorithms, using real data, but in a domain that does not human judgement to measure performance. In particular, we aim measure how often similar input signals are mapped into similar output signals, using output modality data on which performance can be objectively measured in terms of (multiclass) classification accuracy.

4.1. Mapping text to digit images

In this experimental evaluation, we use a supervised input data set $(X_{\text{in}}, y_{\text{in}})$ and an output data set X_{out} where class labels in y_{in} are used for performance assessment only and not during training. The goal is to obtain a pseudo classification accuracy measuring how often input examples of the same class are mapped into output data points of the same class. Since output data points are genuinely new, measuring this kind of accuracy would require human labeling on the generated data. To avoid this step we use handwritten digits as the output modality, given that classifiers with human-level accuracy are available for this type of data. We used 60000 training examples from the MNIST data set to train the output autoencoder. The input modality in this experiment is text and we used 5954 training documents of ten distinct classes from the 20-newsgroups dataset. The input and output deep autoencoders both consisted of stacked RBMs, fine-tuned by backpropagation (Hinton & Salakhutdinov, 2006). Geometries were 784-1000-500-250-4 for MNIST and 2000-500-250-125-6 for newsgroups. The transcoder geometry was 6-30-30-4 using ReLU units in the hidden layers. The overall transcoding pipeline is thus a network with 10 hidden layers. For testing, we fed the whole encoding-transcoding-decoding pipeline with 3963 test documents, generated synthetic characters (samples shown in Fig. 2) and classified them using a state-of-the-art convolutional network (Goodfellow et al., 2013). A heatmap showing the contingency matrix of predicted MNIST classes vs. true newsgroup classes is shown in Fig. 3. The associated multiclass accuracy (using max-weigthed bipartite matching to associate classes of the two modalities) is 47.8%. Interestingly, several "errors" occur with documents of similar categories that cannot be easily distinguished without a supervision signal.

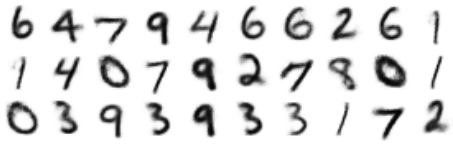

Figure 2. Synthetic digits generated from newsgroups documents.

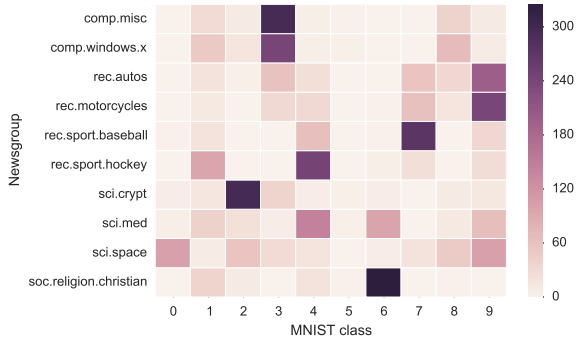

Figure 3. Transcoding newsgroup documents into digits.

4.2. Music generation

In this experiment we test the ability of the model to map text (from the same subset of the Newsgroups data set used in §4.1) to sounds. For sounds, we used a denoising autoencoder with 34 input nodes and and 4 nodes in the hidden layer of the top the encoder. For training, we manually assembled a data set of 100 examples each consisting of a carefully selected assignment to the 34 parameters of the synthesizer. Ten distinct categories of sound settings were designed, each starting from a clearly recognizable main sound with a range of timbral variations. These preserve a number of characteristics of the main sound, such as the fundamental frequency, allowing all sounds in a group to be perceived as belonging to the same sound category. Ten examples for each category were included in the data set.

Connecting the encoder for the newsgroup documents to the decoder for the sound settings through a transcoding network, we obtain a pipeline mapping documents to sound settings. When text documents are fed to the pipeline, they generate novel assignments to the synthesizer's parameters which are finally used to produce sounds. On the known sound examples, we measured the discriminability of the 4-dimensional codes using a linear SVM and obtained a multiclass accuracy of 60%. No performance measure is available in this case for the sound samples generated by the whole pipeline.

The mapping lacks a temporal dimension: one document produces one sound. In order to present experimental results in way that is musically more interesting, sound files were generated from sequences of documents. The synthesizer plays the sounds corresponding to the documents in sequence, for about 10 second each, while slowly morphing from one sound to the next by interpolating between the values of the parameters in the sound settings generated from subsequent documents.

5. Conclusions

Cross-modal transcoding is a general unsupervised approach for generating new data and can be in principled applied to arbitrary pairs of modality types. Our preliminary results show the viability of the method. An obvious direction for future work is to take into account the inherent temporal structure of both input and output signals which should enable more interesting applications such as the generation of music from video.

References

Borg, I. and Groenen, P.J.F. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

Chan, Z.Y. and Dyson, B. The effects of association strength and cross-modal correspondence on the development of multimodal stimuli. *Attention, Perception, & Psychophysics*, 77 (2):560–570, 2015.

Deroy, O. and Spence, C. Why we are not all synesthetes (not even weakly so). *Psychonomic Bulletin & Review*, 20(4):643–664, 2013.

Feng, F., Li, R., and Wang, X. Deep correspondence restricted Boltzmann machine for cross-modal retrieval. *Neurocomputing*, 154(0):50 – 60, 2015.

Goodfellow, Ian J, Warde-Farley, David, Mirza, Mehdi, Courville, Aaron, and Bengio, Yoshua. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

Hinton, G.E. and Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Masci, J., Bronstein, M., Bronstein, A., and Schmidhuber, J. Multimodal similarity-preserving hashing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4):824–830, 2014.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Multimodal deep learning. In *Proc. 28th Int. Conf. on Machine Learning (ICML-11)*, pp. 689–696, 2011.

Parise, C. and Spence, C. Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research*, 220(3-4):319–333, 2012.

Puckette, M. Max at seventeen. *Computer Music Journal*, 26(4): 31–43, 2002.

Roads, C. *The computer music tutorial*. MIT press, 1996.

Srivastava, N. and Salakhutdinov, R. Multimodal learning with deep Boltzmann machines. In *NIPS 25*, pp. 2222–2230. 2012.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res*, 11:3371–3408, 2010.