Abstract—Cross-modal representation learning allows to integrate information from different modalities into one representation. At the same time, research on generative models tends to focus on the visual domain with less emphasis on other domains, such as audio or text, potentially missing the benefits of shared representations. Studies successfully linking more than one modality in the generative setting are rare. In this context, we verify the possibility to train variational autoencoders (VAEs) to reconstruct image archetypes from audio data. Specifically, we consider VAEs in an adversarial training framework in order to ensure more variability in the generated data and find that there is a trade-off between the consistency and diversity of the generated images - this trade-off can be governed by scaling the reconstruction loss up or down, respectively. Our results further suggest that even in the case when the generated images are relatively inconsistent (diverse), features that are critical for proper image classification are preserved.

Index Terms—neural networks, cross-modal learning, generative models

I. INTRODUCTION

In representation learning, a frequent goal is to find a mapping from the data space (e.g. a space of images) to a more compact latent space, which is not directly observable (an embedding space / feature space). One of the main assumptions underlying such a process is that real data can be reliably summarized by a relatively limited set of factors. These factors correspond to elements of the latent / embedding / feature vectors. The resulting reduction of dimensionality facilitates the use of latent vectors for a range of tasks, such as classification or compression. Latent vectors are also frequently used in generative models, which often utilize random latent vectors in the process of data generation.

Regardless of the actual application, most of representation learning research is focused on data from one modality (e.g. image, text, audio, etc.). For instance, image classification or generation tasks usually use data solely from the image modality. Despite the success of such applications, it seems that limiting oneself to data from one modality is not a realistic model for deep learning. Real-world objects usually generate data from more than one modality and humans perceive the world around them through more than just one sense. A significant portion of the objects we interact with can be identified based on data from multiple senses. In particular, it is usually possible to imagine an object based on only one source of sensory information (e.g. imagining a car based on the sound of its engine), which suggests the possibility of learning a common representation of an object integrating information from multiple modalities. Such a representation could then plausibly be matched to data from just one modality and the missing modalities could be inferred.

The process of learning such common representations would be facilitated by the temporal alignment of data from multiple modalities. In such a setting, sensory information from different modalities appears together and this link between multiple modalities enables learning without additional supervision.

In this work, we aim to highlight the possibility of learning shared representations from cross-modal (multi-modal) data due to the temporal alignment of datasets.

A. Main contribution

- We construct audio-image datasets and show that it is possible to generate images based on audio data, as audio features overlap with visual features to an extent that makes image generation possible.
- We discuss how the data alignment process influences the generation process. Many-to-one mappings result in data archetypes, while one-to-one mappings ensure more variability in the data.
- We analyze how scaling the reconstruction loss up or down influences the properties of the generated data.
- We quantitatively assess the degree to which transition between audio and image data preserves relevant visual features.

II. RELATED WORK

This work is related to several strands of research, which include representation learning, joint latent space learning, input reconstruction and cross-modal learning.

Representation learning is addressed by Hinton et al. [1] who discuss the possibility of representing a concept as an activity pattern in more than one computing element within a network of such elements - a distributed representation. They argue that such representations could potentially achieve content-addressable memory and generalization capabilities. Schmidhuber [2] proposes that statistical independence of parts of the representation should be a goal of unsupervised learning. Bengio et al. [3] restate this in the form of disentangled factors of variation, motivate the importance of representation learning, and provide a review of success stories involving representation learning. They explicitly hypothesize...
that representations might be partially shared for different tasks or domains and that this might be beneficial in transfer learning, domain adaptation or multi-task learning.

In reference to cross-modal learning and joint latent space learning, Weston et al. [4] train a method to learn a joint embedding space for both images and their annotations. Arandjelović and Zisserman [5] show that it is possible to jointly learn audio and visual representations from an aligned audio-visual dataset with no additional supervision. The representations from the audio and visual modalities can then be compared to localize objects potentially linked to the sound and vice versa. Xing et al. [6] combine information from visual and semantic feature spaces to improve model performance in a few-shot learning setting. Li et al. [7] present the existence of adversarial examples for cross-modal text/image data and show that adversarial training can improve the robustness of a cross-modal hashing network. Wen et al. [8] present an architecture capable of learning a shared representation for voices and faces by mapping them to their common covariates. Lu et al. [9] learn task-agnostic joint representations of images and natural language. In a video setting, Sun et al. [10] combine video data and text produced by speech-recognition methods to learn high-level linguistic features. Karam et al. [11] design a cross-domain transfer learning framework for wrist fracture detection.

In the context of input reconstruction, Ackley et al. [12] pose an encoder problem where input should be reconstructed in spite of a sparse representation, while Rumelhart et al. [13] show that it is possible to train a neural network with an intermediate bottleneck layer in an unsupervised manner to reconstruct input, an idea further developed in [14], [15], [16], [17]. Kingma and Welling [18] use a mapping from the input to the parameters of a multidimensional normal distribution in a variational autoencoder (VAE) architecture. This facilitates the process of generating new data from latent space representations without the corresponding input data. Extensions incorporating discrete latent codes have also been proposed [19], resulting in images of increasing size [20].

A different line of generative models has been triggered by Goodfellow et al. [21], who proposed a generative adversarial network (GAN) framework. This framework relies on game-theoretic ideas and jointly trains two neural networks with opposing objectives in order for one of the networks to learn to generate data similar but not identical to examples from the train dataset. While GANs tend to produce sharper images than VAEs, significant instabilities in adversarial training have been identified and various approaches to stabilize training have been proposed [22], [23], [24], with no approach conclusively resolving the issue so far.

There is also a growing body of research on cross-modal generative models. A generative model conditioned on linguistic features extracted from text input is trained for the text-to-speech task in [25]. Conditional GANs are used to generate images from text in [26]. The tasks of generating images from audio and generating audio from images are tackled in an adversarial setting in [27]. Hao et al. [28] extend the adversarial framework to generate audio and visual data mutually. Hsu and Glass [29] train a modified VAE to incorporate not only shared factors but also modality-dependent ones for audio/image datasets. Shi et al. [30] employ a cross-modal VAE to generate images from text and text from images. Müller-Eberstein and van Noord [31] show that, with pre-training, VAEs can be recombined to generate audio from images and that the produced audio is at least somewhat consistent with human notions of similarity between the audio and images, even in the absence of aligned datasets. Adversarial methods are applied to the task of generating facial animation based on speech data in a strategy where the motion and texture of a face are learned separately [32]. An approach to generating visually-aligned sound from videos is presented in [33], where appearance and motion features are first extracted to filter the object generating the sound from background information. Lee et al. [34] propose a GAN-based architecture which can be used to manipulate images based on audio data. GANs are also used to produce images from sound descriptions without auxiliary text data in [35].

While research on cross-modal generative models has advanced significantly, it is still very much underdeveloped. Most of this area focuses on linking the linguistic domain with the audio domain (e.g. text to speech) or providing a bridge between the linguistic and the visual domains (e.g. image captions, text-to-image generation). At present, the audio-visual branch is underrepresented within the broader cross-modal generative current. In spite of the development of cross-modal methods, relatively little research on audio-visual architectures has been published. Even less research has been conducted directly on audio-image generation. Relative to image datasets, audio-visual datasets are few and far between.

In this context we strive to align audio and image datasets in order to facilitate the process of audio-image generation. We also propose methods to extract audio features and utilize their subset, the audio-visual features, to produce images consistent with the ones observed in the data space. These methods differ from the standard supervised conditional variational autoencoder or conditional generative adversarial network setups [21]. Our work is a potential starting point for research on the possibility of utilizing variational autoencoders in an adversarial setting for audio-visual generation. It is also the first study we are aware of concerned with many-to-one vs. one-to-one mappings for cross-modal dataset alignment and the effects of these mappings on the generated data. Our work points to a new potential area of research related to precise methods of set alignment and how they can impact the data generation process. In particular, many-to-one mappings may be preferable for some tasks, and vice versa.

### III. Generative Models

For two variables \( x, y \), a generative model is one which describes the joint probability distribution \( p(x, y) \). In machine learning applications, the model can be written as \( p(x, z) \), where \( x \) comes from the data space, e.g. the space where
we can observe the data, and \( z \) comes from the \textit{latent space}, a directly unobservable space of features.

**A. Variational autoencoders (VAEs)**

VAEs [18] extend the autencoder architecture. Let us consider the generative model \( p_\theta(z)p_\theta(x|z) \), and \( q_\phi(z|x) \) be the approximation of the true intractable posterior \( p_\theta(z|x) \). We may refer to \( q_\phi(z|x) \) as a probabilistic encoder and to \( p_\theta(x|z) \) as a probabilistic decoder. Let \( p_\theta(z|x) \) and \( q_\phi(z|x) \) be neural networks parametrized by \( \theta \) and \( \phi \), respectively. Let \( z \) be a random variable and \( z = g_\theta(x, \epsilon) \), where \( \epsilon \) is an auxiliary variable with an independent marginal and \( g_\theta \) is a function parametrized by \( \theta \). We can sample from the posterior \( z^{(i)} \sim q_\phi(z|x^{(i)}) \), calculating \( z^{(i)} = g_\theta(x^{(i)}, \epsilon) = \mu^{(i)} + \sigma^{(i)} \odot \epsilon \), where \( \epsilon \sim \mathcal{N}(0, I) \). The loss function for a given data example can be expressed as:

\[
\mathcal{L}(x^{(i)}, \theta, \phi) = -E_{z \sim q_\phi(z|x^{(i)})} \left[ \log p_\theta(x^{(i)}|z) \right] + KL(q_\phi(z|x^{(i)}) || p(z)) \tag{1}
\]

where \( KL \) is the Kullback–Leibler divergence.

In practice, the process of training a VAE consists of passing data through the encoder, which maps to parameter vectors \( \mu \) and \( \sigma \). These vectors are then combined with a random \( \epsilon \) vector sampled from the standard normal distribution via the equation \( z^{(i)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon \). Thanks to the reparametrization trick, the model is differentiable and can be trained with backpropagation. It is worth noting that the loss formulation consists of two terms: the \textit{reconstruction error} and the KL divergence. While the reconstruction loss (the first term of the overall loss formulation) encourages adhering to examples seen in the dataset, the KL divergence (the second term) ensures that the encoder remains consistent with the standard normal distribution assumed for \( p(z) \). Such an assumption simplifies sampling from a trained model.

**B. Generative adversarial networks (GANs)**

GANs, first introduced in [21], apply the concept of a minimax game to generative models. In this setting, we consider two neural networks, the generator \( G(z) \) and the discriminator \( D(x) \). We define a latent space prior \( p_\theta(z) \). \( G \) maps from the latent space to the data space, while \( D \) maps from the data space to a scalar output, which represents the probability that a sample comes from the data space rather than that it has been produced by \( G \). \( G \) and \( D \) are trained in turns, essentially playing a two-player minimax game:

\[
\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{x \sim p_{\text{data}}} \left[ \log(D(x)) \right] + \mathbb{E}_{z \sim p_{\phi}(z)} \left[ \log(1 - D(G(z))) \right] \tag{2}
\]

This minimax game has a global optimum for \( p_\theta = p_{\text{data}} \), i.e. for \( G \) generating images indistinguishable from the ones in the data space and \( D \) continually producing \( \frac{1}{2} \) as its assessment of the probability that the samples come from the data distribution. This global optimum can theoretically be achieved via backpropagation in sequential weight updates for both networks. In reality, the training process for GANs is frequently unstable and convergence is by no means automatic.

**IV. Audio-to-Image Architectures**

We consider two generative setups geared towards extracting audio representations from sound and using a subset of these features relevant for the visual domain, the audio-visual features, for image generation (Figure 1). Our first model is a modification of the classic VAE architecture. Our second architecture is a combination of a VAE and a GAN network. We use VAE as the GAN generator and train the whole system adversarially. Both architectures are presented in Figure 2.

**A. Audio-to-Image VAE**

We modify the VAE encoder to process audio data and map it to the parameters of the latent space. It now describes a distribution \( q_\phi(f|a) \), where \( a \) represents the audio input and \( f \) denotes the \textit{audio-visual features} - features from the audio space, which would be relevant for image generation. These features are used as parameters \( \mu \) and \( \sigma \) in the process of generating random vectors \( f^{(i)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon \), in line with the classic VAE architecture. The random vectors are then passed through the decoder \( p_\theta(i|f) \), which maps to the image space. The loss minimized in training is:

\[
\mathcal{L}(a^{(i)}, i^{(i)}, \theta, \phi) = -E_{f \sim q_\phi(f|a^{(i)})} \left[ \log p_\theta(i^{(i)}|f) \right] + KL(q_\phi(f|a^{(i)}) || p(f)) \tag{3}
\]

The whole architecture is trained with backpropagation just like a vanilla VAE. However, the Audio-to-Image VAE (AIVAE) requires two aligned data sources, as each training example is a tuple \( (a^{(i)}, i^{(i)}) \).

**B. Audio-to-Image VAE-GAN**

AIVAE can be further extended in a setting where it is treated as the generator in an adversarial framework, \( G(a) = \text{AIVAE}(a) \). Now, \( G : a \to i \), where \( a \) and \( i \) are the audio and image spaces, respectively. The discriminator \( D(i) \) tries to discern between real images and the ones manufactured by \( G \). It is worth noting that \( D \) does not have explicit access to the pairing between the image and audio and so is not
likely to enforce correct classes from \( G \). The information about the pairing \((a^{(i)}, i^{(i)})\) is incorporated via a reconstruction loss term in the objective function:

\[
\min_G \max_D V(D, G) = \mathbb{E}_{i \sim p_{\text{data}}(i)} \{ \log (D(i)) \} \\
+ \mathbb{E}_{a \sim p_{\text{data}}(a)} \mathbb{E}_{f \sim q_{\phi}(f|a)} \{ \log (1 - D(G(f))) \} \\
- \alpha \log p_{\theta}(i|f) + \text{KL}(q_{\phi}(f|a) || p(f))
\] (4)

This objective function incorporates the opposing goals of \( G \) and \( D \), while taking into account the information about the expected reconstruction and desired latent (feature) distribution. The hyperparameter \( \alpha \) is designed to regulate the importance of reconstruction vs. adversarial goals. Similarly to the AIVAE setup, the Audio-to-Image VAE-GAN (AIVAEGAN) loss relies on aligned data without explicit supervision.

V. EXPERIMENTS

We test AIVAE and AIVAEGAN on two synthetic datasets in order to assess their capability to generate images from audio data.

A. Datasets

We construct two synthetic datasets for our experiments. Both of them rely on the MNIST dataset as a source of digit images to be paired with audio data.

1) MNIST-FSDD: The first dataset combines \( 28 \times 28 \)-pixel MNIST digits with audio data from the Free Spoken Digit Dataset (FSDD). \(^1\) FSDD consists of 2000 WAV files, which are the recorded pronunciations of digits by 4 speakers, which amounts to 50 pronunciations per digit per speaker. We represent the recordings as MEL-scaled spectrograms resized to \( 48 \times 48 \) pixels. We then perform a random train/test split, where \( 90\% \) of the data is assigned to the train set and the remaining \( 10\% \) of the data is in the test set.

After initial preprocessing, the FSDD data is paired with the MNIST digits in the following procedure. We sample digits from MNIST without replacement and for each of them, we sample with replacement from FSDD spectrograms with matching labels. Each MNIST digit is used once, but each FSDD spectrogram is paired multiple times with various MNIST digits from the corresponding class. This many-to-one MNIST \( \rightarrow \) FSDD mapping enforces the alignment of datasets along labels. This is the only time the label information is used prior to evaluation. The outlined steps are repeated separately for the train and test sets. The synthesized dataset consists of 60,000 audio-image pairs for the train set and 10,000 pairs for the test set.

2) MNIST-SCD: Our second dataset is based on MNIST for image data and on the Speech Commands Dataset (SCD) for audio data \(^2\). SCD has 65,000 one-second long recordings of 30 words. We limit our analysis to the words representing digits. This leaves us with 23,666 utterances distributed among 10 classes. For simplicity, we will henceforth refer to this subset as \( SCD \), having in mind that we consider the digit-related subset of the original dataset. Similarly to the previous case, we convert the WAV files to MEL-scaled spectrograms, resize them to \( 48 \times 48 \) pixels and randomly split the dataset so that \( 90\% \) of the data is used as a train set and \( 10\% \) is held out for evaluation. We align MNIST and the truncated SCD in a sequential approach. For each class label we randomly sample a spectrogram without replacement and match it to a MNIST image with the same class via random sampling without replacement. This results in a one-to-one MNIST \( \rightarrow \) SCD mapping. Again, these steps are performed separately for the train and test sets. The resulting train set is comprised of 21,160 audio-image pairs, while the produced test set consists of 2,360 audio-image pairs.

\(^1\) The FSDD dataset is available at https://github.com/Jakobovski/free-spoken-digit-dataset. This work is based on commit 9b8dcea.

\(^2\) The SCD dataset can be downloaded from https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html.
TABLE I: Audio-to-image variational autoencoder architecture (AIVAE).

| AUDIO ENCODER | IMAGE DECODER |
|---------------|---------------|
| INPUT 48x48   | INPUT 64      |
| Conv 4x4, 64, str=2 | FC 512         |
| ReLU          | ReLU          |
| Conv 4x4, 128, str=2 | FC 1024        |
| ReLU          | FC 1024       |
| FC 512        | ReLU          |
| ReLU          | FC 7x7x128    |
| Conv 4x4, 64, str=2 | Upconv 4x4, 1, str=2 |
| ReLU          | ReLU          |
| Output 2x64   | Output 2x64   |

TABLE II: Audio-to-image variational autoencoder - generative adversarial network architecture (AIVAEGAN).

| AUDIO ENCODER | IMAGE DECODER | DISCRIMINATOR |
|---------------|---------------|---------------|
| INPUT 48x48   | INPUT 64      | Input 28x28   |
| Conv 4, 128, 2 | Upconv 3, 512, 2 | Conv 4, 128, 2 |
| Conv 4, 256, 2 | Upconv 3, 256, 2 | Conv 4, 256, 2 |
| Conv 4, 512, 2 | Upconv 2, 128, 2 | Conv 4, 512, 2 |
| μ, σ: Conv 4, 64, 2 | Upconv 2, 1, 2 | Conv 1, 1, 1, Sigmoid |
| Output 2x64   | Output 28x28  | Output 1      |
| LeakyReLU(0.2) | ReLU          | LeakyReLU(0.2) |
| BatchNorm     | BatchNorm     | BatchNorm     |

B. Results

We test the AIVAE (Table I) and AIVAEGAN (Table II) architectures on the MNIST-FSDD and MNIST-SCD datasets with the latent space dimension set to 64 in each experiment. This dimensionality allows us to obtain satisfactory quality of generated images without presenting too much of a computational burden. All models are trained for 100 epochs with batches of 128 examples and are optimized with Adam [36]. The initial learning rate is set to $10^{-3}$ for AIVAE and to $2 \times 10^{-4}$ for both the generator and the discriminator of AIVAEGAN.

1) Qualitative evaluation: We first investigate the reconstruction capabilities of AIVAE for MNIST-FSDD and MNIST-SCD. Figure 3 presents real images from the last epoch of training together with their reconstructions. The classes of the generated images are consistent with their real counterparts for both datasets. The images generated for MNIST-SCD display more variability and less blur, while the ones generated for MNIST-FSDD display a pattern in which each class is represented by visually indistinguishable images - the archetypes for each class. The archetypical nature of these images stems from the fact that the MNIST-FSDD dataset was constructed via a many-to-one mapping and each audio data point was associated with multiple digits from the MNIST database. Consequently, for each digit the model has learned its average reconstruction.

The reconstructions are further evaluated on the test set with the results presented in Figure 4. These outcomes support the initial idea that the model trained on the many-to-one dataset recovers digit archetypes. The model trained on the one-to-one dataset is less consistent, as evidenced by partially unintelligible reconstructions, however, it does retain reconstruction capabilities and shows more variability in the generated data.

For the AIVAEGAN model, the train-set reconstructions on MNIST-FSDD and MNIST-SCD are presented in Figures 5 and 6, respectively. For MNIST-FSDD, the quality of reconstructions at the end of training is relatively high, with lower levels of $\alpha$ promoting more variation in the generated images. Conversely, for $\alpha = 1$ and $\alpha = 2$, stricter adherence to train data is observed. The results for MNIST-SCD follow the same pattern, however, the one-to-one mapping in this dataset ensures that more variability is preserved without necessarily sacrificing the correct labels.

These outcomes are verified on the test sets (Figure 7) and are consistent with our earlier remarks. Lower $\alpha$ levels result in more variation and less adherence to proper classes. Higher values of $\alpha$ result in more archetypical data, however, even for larger $\alpha$ values a one-to-one training set mapping preserves diversity in the generated data.

2) Quantitative evaluation: While eyeball tests confirm the possibility of image generation from audio features for...
algorithms tuned to different levels of reconstruction loss importance (\(\alpha\) AIV AE and \(\alpha\) AIV AEGAN, MNIST-FSDD).

aligned datasets, such a form of inspection is inherently subjective. A quantitative evaluation of the ability of AIV AE and AIV AEGAN to preserve audio features important for image generation would be a more objective measure. To this end, the following procedure is proposed. First, we generate images from audio for the test sets of both MNIST-FSDD and MNIST-SCD. Then, these images are fed to a pretrained MNIST classifier and the accuracy of prediction is recorded. Such a process would allow to assess the extent to which the images generated by our models retain information necessary for image classification.

We choose a LeNet5 classifier [37] - a classic CNN with convolutional/pooling blocks followed by fully-connected layers. This architecture is chosen because of its relative simplicity, speed of training and accuracy. We train it on the MNIST train set for 100 epochs with Adam, with an initial learning rate of \(10^{-3}\). The network achieves a 98.7% accuracy on the MNIST test set. This is the benchmark we would compare our models against. Intuitively, the accuracy of the pretrained classifier on the data generated by our models should be lower as the audio features used in the data generation process potentially encompass only a subset of the relevant visual features (the audio-visual features).

It turns out that for AIV AE the accuracy remains close to 95% for the many-to-one data setting. The one-to-one data alignment results in lower accuracy, albeit still above 85% (see Table III). AIV AEGAN results display a pattern where for many-to-one data the accuracy of the LeNet5 classifier remains considerably higher than for the one-to-one dataset. At the same time, a higher weight associated with the reconstruction loss improves visual feature retention.

A comparison of AIV AE and AIV AEGAN reveals that the gap in performance between the two models is influenced by the level of \(\alpha\) set for AIV AEGAN. Higher levels of \(\alpha\) result in a smaller difference in performance between the two models, particularly for \(\alpha = 2\) and the MNIST-FSDD dataset. This is perhaps not surprising, as the reconstruction loss explicitly enforces stronger adherence to images from the data space.
lower $\alpha$ promotes more diversity in the generated digits, but this comes at a cost of losing visual features distinguishable for a pretrained classifier. It does seem that the effect of the reconstruction loss on the error rate is more pronounced for the many-to-one mapping. Going from $\alpha = 0.2$ to $\alpha = 2$, the error rate drops 44% in the one-to-one setting, while the corresponding drop for the many-to-one setting is 70%. The lower sensitivity of the error rate to changes in $\alpha$ for the one-to-one setting, combined with the less stable adversarial learning process, may also account for a lower top accuracy on MNIST-SCD when comparing AIVAEGAN against AIVAE. Preliminary experiments, omitted in this work for brevity, suggest that setting $\alpha > 2$ does not change the conclusions.

We conduct an additional evaluation of the quality of features of AIVAE trained on MNIST-FSDD and MNIST-SCD. Namely, we randomly set $k$ elements of the latent vector to 0 and record the classification accuracy of the LeNet5 classifier on both test sets. We repeat this for $k \in \{0, 1, \ldots, 64\}$. The results are presented in Figure 8. With $k$ increasing, the classification accuracy naturally diminishes, confirming the usefulness of the audio-visual features for image generation. The interesting result is that the gap in accuracy between the models trained on the many-to-one dataset and the one-to-one dataset narrows down and actually disappears for larger numbers of the elements of the latent vector restricted to 0. This suggests that for the many-to-one alignment, the whole set of features is crucial for performance, while the one-to-one alignment results in a model that is more robust to a removal of a small subset of features.

While the obtained overall results concern the scenario of generating images of digits from sound when paired datasets are available, extending this approach to larger and potentially unaligned datasets would be an interesting task. We intend to pursue these lines of research in future work.

VI. Conclusions

In this work we show that it is possible to generate images based on audio features extracted from sounds. We train cross-modal VAEs and their adversarial extensions on two synthetic datasets and obtain results which suggest that the audio-visual features - the subset of relevant visual features corresponding to audio features - are sufficient to generate images similar to the ones from the data space. We also show that the degree of adherence of the generated images to the ones from the data space can be modeled by choosing an appropriate weight for the reconstruction loss. For audio which repeats itself for a given image, our models tend to produce image archetypes. In a setting where each data point is uniquely paired with another one from the other modality, archetypes are weaker and the data retains more diversity without necessarily losing much of the relevant visual features.

Our study provides results in a hitherto underrepresented area of cross-modal learning related to audio-visual data. We show that audio-to-image generation is possible both in a VAE architecture and by adversarially extending it. This provides a new approach toward architectures which could be employed in a cross-modal setting. Additionally, we provide an analysis of the impact the kind of dataset alignment (many-to-one vs. one-to-one) may have on the generated data. It is not hard to imagine scenarios where the generation of archetypes might be of interest (e.g. when we are interested in only the general notion and not a specific object) vs. scenarios where diversity could be preferable to strict adherence (e.g. when we are interested in obtaining many examples of one specific notion).

On a more general note, our work provides a method for generating data in an unsupervised manner, which could contribute toward creating a more realistic model of the world in artificial intelligence systems. Generating images from sound could potentially be used in a variety of situations, e.g. to help people with impaired hearing, to visualize sounds, to create art, etc.

Presented results point to several potential lines of research which we plan to consider in our future work. First of all, while in this study we have discounted the possibility of forming one representation per modality rather than one universal shared representation, it might be beneficial to analyze settings where two or more representations are formed with a shared subset as this may allow to retain more relevant features from each
analyzed modality. Second, the obtained results suggest that the supervisory signal contained in the alignment of datasets may be enough to at least somewhat successfully generate data from one modality based on the aligned data from another modality. This, in turn, suggests the possibility of leveraging a number of unlabeled datasets to obtain shared cross-modal representations. Online streaming services which present their users with videos seem like a natural choice since videos contain naturally aligned audio and visual data. It would be interesting to see extensions handling temporal data such as audio-video streams. Finally, it would be interesting to see extensions handling temporal data such as audio-video streams. Finally, it would be interesting to test the extent to which cross-modal representations can be formed from unaligned datasets, and the extent to which such representations could be used to bind two data points from different modalities.

Acknowledgements

The project was funded by the POB Research Centre Cybersecurity and Data Science of the Warsaw University of Technology within the Excellence Initiative Program - Research University (ID-UB).

REFERENCES

[1] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart, Distributed Representations. MIT Press, 1986, p. 77–109.
[2] J. Schmidhuber, “Learning factorial codes by predictability minimization,” Neural Computation, vol. 4, no. 6, pp. 863–879, Nov 1992.
[3] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, Aug 2013.
[4] J. Weston, S. Bengio, and N. Usunier, “Large scale image annotation: learning to rank with joint word-image embeddings;” Machine Learning, vol. 81, no. 1, pp. 21–35, Oct 2010.
[5] R. Arandjelovic and A. Zisserman, “Objects that sound,” in Proceedings of the European Conference on Computer Vision (ECCV), September 2018.
[6] X. Wang, Q. T., J. Zhu, A. Hanjalic, and O. Scharenborg, “Generating visually aligned sound from videos,” in Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019, pp. 6306–6315.
[7] C. Li, S. Gao, C. Deng, D. Xie, and W. Liu, “Cross-modal learning from one modality based on the aligned data from another modality;” in Proceedings of The 33rd International Conference on Machine Learning (ICML), May 2016.
[8] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan;” in International Conference on Machine Learning, August 2017.
[9] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks;” in International Conference on Learning Representations, 2018.
[10] M. A. Kramer, “Nonlinear principal component analysis using autoassociative neural networks;” AChE Journal, vol. 37, no. 2, pp. 233–243, Apr 1991.
[11] G. E. Hinton and R. S. Zemel, “Autoencoders, minimum description length and helmholtz free energy;” in Proceedings of the 6th International Conference on Neural Information Processing Systems. Morgan Kaufmann Publishers Inc., 1993, p. 3–10.
[12] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders;” in Proceedings of the 25th International Conference on Machine Learning, ser. ICML ’08. Association for Computing Machinery, 2008, p. 1096–1103.
[13] D. P. Kingma and M. Welling, “Auto-encoding variational bayes;” in Proceedings of the 2nd International Conference on Learning Representations (ICLR), 2014.
[14] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning;” in Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017, pp. 6306–6315.
[15] A. Razavi, A. van den Oord, and O. Vinyals, “Generating diverse high-fidelity images with vq-vae-2;” in Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019, pp. 14837–14847.
[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets;” in Advances in Neural Information Processing Systems 27. Curran Associates, Inc., 2014, pp. 2672–2680.
[17] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks;” in International Conference on Learning Representations (ICLR), May 2016.
[18] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio;” arXiv:1609.03499, 2016.
[19] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text-to-image synthesis;” in Proceedings of The 33rd International Conference on Machine Learning, 2016.
[20] 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, Mountain View, CA, USA, October 23 - 27, 2017, pp. 349–357.
[21] W. Hao, Z. Zhang, and H. Guan, “Cmegan: A uniform framework for cross-modal visual-audio-visual generation;” in AAAI Conference on Artificial Intelligence, 2018.
[22] W.-N. Hsu and J. Glass, “Disentangling by partitioning: A representation learning framework for multimodal sensory data;” arXiv:1805.11264, 2018.
[23] Y. Shi, S. N. B. Paige, and P. Torr, “Variational mixture-of-experts autoencoders for multi-modal deep generative models;” in Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019, pp. 15692–15703.
[24] M. Müller-Eberstern and N. van Noord, “Translating visual art into music;” in International Conference on Computer Vision (ICCV), 2019.
[25] D. Das, S. Biswas, S. Sinha, and B. Bhowmick, “Speech-driven facial animation using cascaded gans for learning of motion and texture;” in Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol. 12375. Springer, 2020.
[26] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan, “Generating visually aligned sound from videos;” in IEEE Transactions on Image Processing, vol. 29, 2020.
[27] S. Lee, S. Yoon, J. Kim, and S. Kim, “Sound-guided semantic image manipulation;” in NeurIPS Workshop, CtrlGen: Controllable Generative Modeling in Language and Vision, 2021.
[28] X. Wang, Q. T., J. Zhu, A. Hanjalic, and O. Scharenborg, “Generating images from spoken descriptions;” IEEE - ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 850–865, 2021.
[29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization;” ICLR, 2015.
[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition;” in Proceedings of the IEEE, 1998, pp. 2278–2324.