Cross-modal Generative Networks

Audio to Image Generation

Maciej Żelaszczyk

January 8, 2020

PhD Student in Computer Science
Division of Artificial Intelligence and Computational Methods
Faculty of Mathematics and Information Science

m.zelaszczyk@mini.pw.edu.pl

Warsaw University
of Technology

Research

Research hypotheses:

- It is possible to train a generative model using exclusively supervision from the alignment of datasets from different modalities.
- A model trained in this way could preserve a subset of features important for the output domain.

More concretely:

- It is possible to train a model to generate images from audio.
- The generated images are recognizable to a pretrained classifier.

Inspiration

Inspiration:

- Setups covering more than one data modality have not been covered for some domains.
- In principle, it is possible to use the temporal alignment of the data to ensure training without classic supervision.
- Going from one modality to another may be beneficial from a practical point of view (e.g. generating visual cues).
- Strands of neuroscience research show that information might be represented in a similar manner for different modalities. For instance, the semantic representations evoked by listening versus reading are almost identical [Deniz et al., 2019]. This is actually controversial.

Supervised vs. unsupervised

Supervised:

- Requires huge datasets.
- Annotating is costly.
- Extensive training.
- Driving a car off a cliff.
- Learns tasks, not skills.
- Some well-specified tasks have been largely solved.
- Limit to how much data we can obtain.
- Ignores physical world.

Supervised vs. unsupervised

How do children learn?

- A lot of evolutionary knowledge.
- Vision, hearing, touch etc. in place.
- Extensive observation.
- Build a model of the world.
- Model vs. physical world.
- Surprise, curiosity guide learning.
- Continuous refinement of model.
- Limited reinforcement learning.
- All initial learning is unsupervised.

Supervised vs. unsupervised

Unsupervised:

- In practice, very little labelled data available.
- Need to create model of world, confront it with reality.
- Attend to data.
- Manipulate world.
- Learn from little external reward.
- Learn from very few examples.
- Exploit physical structure of world to obtain links.
- Learn skills rather than tasks.

Desired architecture

What would we like our architecture to have?

- Unsupervised/weakly-supervised.
- Model of observed data.
- Potential to learn from observation only.
- Exploit structure of physical world.

Generative models

Models:

- Discriminative: $P(Y|X = x)$
- Generative. Joint probability distribution: $X \times Y, P(X, Y)$
- No hard demarcation line.

Standard generative models in deep learning:

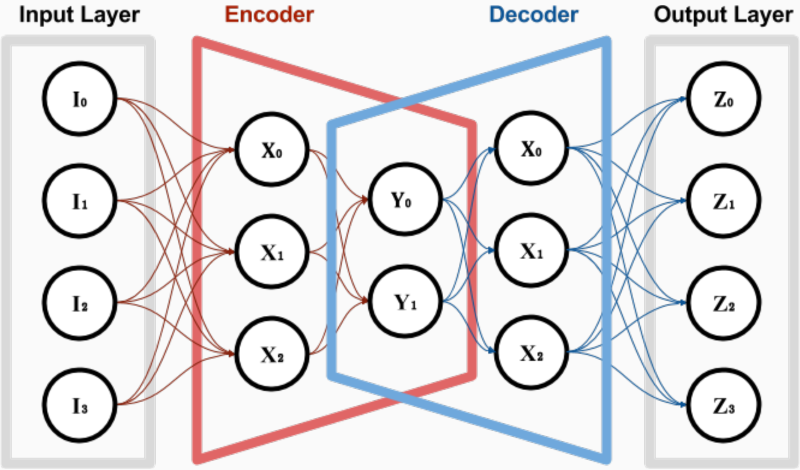
- Autoencoders.
- Variational autoencoders (VAEs).
- Generative adversarial networks (GANs).

Autoencoders

Main idea behind autoencoders:

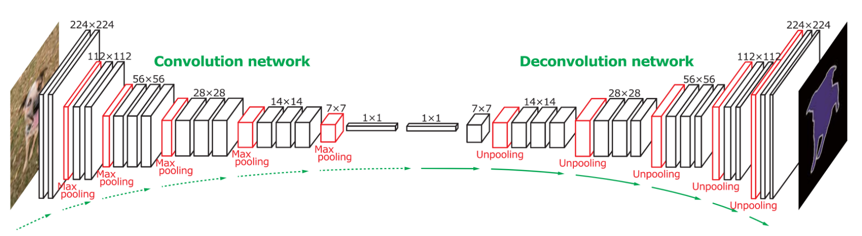
- One network to encode input.
- Second network to decode output.
- Bottleneck in between.
- Latent representation.

Autoencoders



Source: Zucconi, A., *An Introduction to Neural Networks and Autoencoders*

Autoencoders

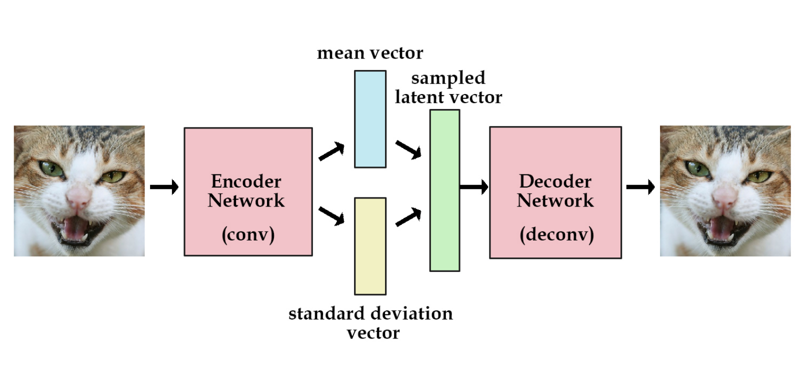


Source: [Noh et al., 2015]

Variational Autoencoders

Introduced in [Kingma and Welling, 2014]:

- Latent variable matches unit Gaussian.
- Loss = generation loss + KL divergence.



Source: Frans, K., *Variational Autoencoders Explained*

Generative Adversarial Nets

Approach model training from game-theoretic point of view [Goodfellow et al., 2014]:

- Two networks: Generator and Discriminator.
- Generator: from latent variable $\mathbf{z}$ generate into data space.
- Discriminator: distinguish between real and generated data.
- Generator tries to "fool" the Discriminator.
- Discriminator strives to "look through" the Discriminator.
- This can be represented by a minimax two-player game.

Generative Adversarial Nets

Training:

- Train D to maximize probability of assigning correct label to real data and samples from G.
- Train G to maximize probability of D assigning incorrect label to samples from G.
- D and G play:
- $\min_{G} \max_{D} V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})} [\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]$.
- $\log(1 - D(G(\mathbf{z})))$ may saturate early in training.
- Can train G to maximize $\log(D(G(\mathbf{z})))$ instead.

Source: [Goodfellow et al., 2014]

Desired architecture

What would we like our architecture to have?

- Unsupervised/weakly-supervised.
- Model of observed data.
- Potential to learn from observation only.
- Exploit structure of physical world.

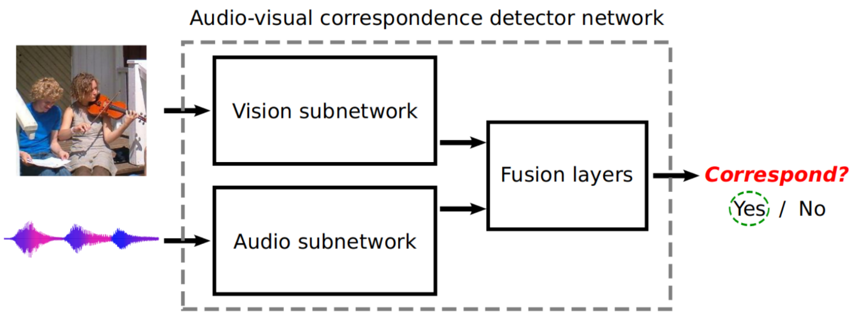
Multi-modal representation

Looking at data across modalities helps achieve some of our goals. For instance, let us consider visual data with corresponding audio:

- Extensive video datasets available.
- Sound aligned with video - exploit structure of the physical world.
- Data alignement obviates strong supervision.

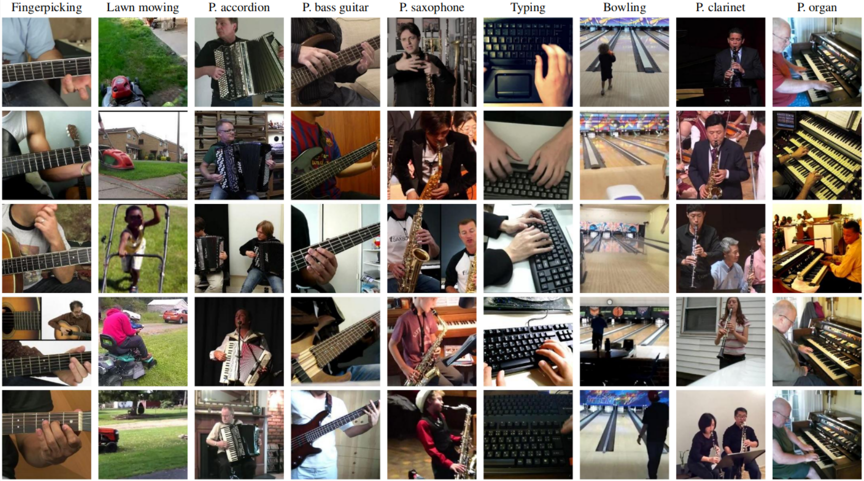
Audio-visual correspondence

What can be learnt by training audio and visual networks jointly to establish whether audio and visual information match?



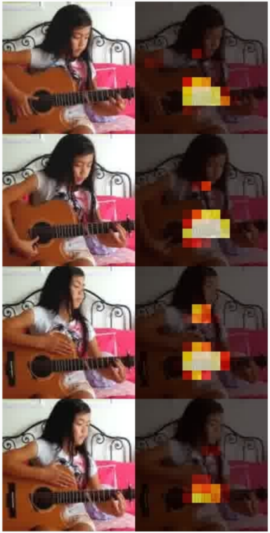
Source: [Arandjelovic and Zisserman, 2017]

Audio-visual correspondence



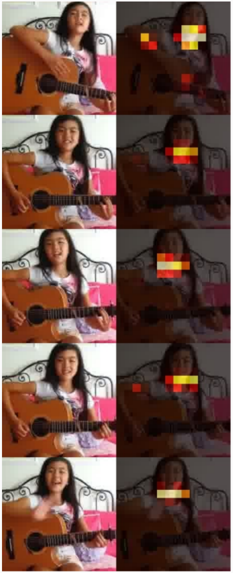
Source: [Arandjelovic and Zisserman, 2017]

Cross-modal retrieval



Source: [Arandjelovic and Zisserman, 2018]

Cross-modal retrieval



Source: [Arandjelovic and Zisserman, 2018]

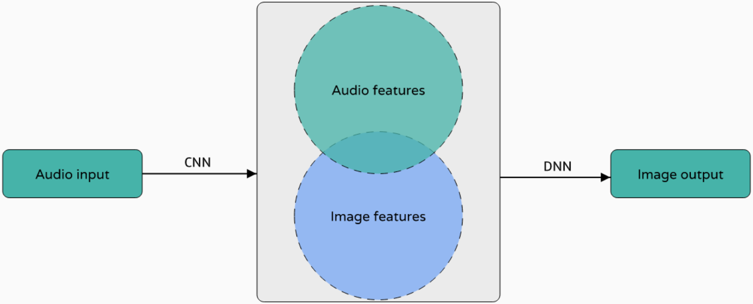
Idea

What if we can use aligned datasets to generate images from audio?

- Use an encoder to extract audio features.
- Condition on the audio features to generate images via a decoder.
- Multiple ways to train this.
- Possibility to use VAE and measure the reconstruction loss and KL divergence.
- We could also potentially train adversarially.

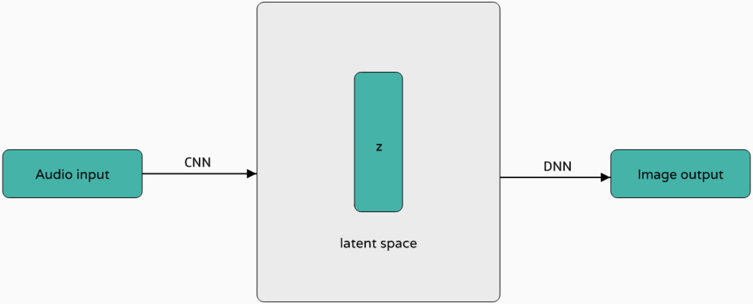
Idea

Encoder/decoder setup.



Idea

Encoder/decoder setup.



Datasets

We use a synthetic dataset combining datasets from the audio and image modalities:

- Audio: Free Spoken Digit Dataset (FSDD).
- Image: MNIST.

FSDD

Free Spoken Digit Dataset (FSDD):

- Recordings of spoken digits in *.wav* files at 8kHz. The recordings are trimmed so that they have near minimal silence at the beginnings and ends.
- $2,000$ recordings, 50 pronounciations of each digit for a speaker.
- 4 speakers.
- English pronounciations.

https://github.com/Jakobovski/free-spoken-digit-dataset

MNIST

Modified National Institute of Standards and Technology (MNIST) database:

- Images of digits, 28×28 pixels, antialiased.
- Train set: $60,000$ images.
- Test set: $10,000$ images.

http://yann.lecun.com/exdb/mnist/

MNIST/FSDD Mix

We apply the following procedure to combine MNIST with FSDD:

- We represent FSDD recordings as 48×48 MEL-scaled spectrograms.
- We perform a 90/10 random train/test split on FSDD.
- For each image from the MNIST train set, we select a subset of the FSDD train set with the same labels as the image and we randomly choose (with replacement) a spectrogram from this subset.
- Similarly, for each image from the MNIST test set, we select a subset of the FSDD test set with the same labels as the image and we randomly choose (with replacement) a spectrogram from this subset.

http://yann.lecun.com/exdb/mnist/

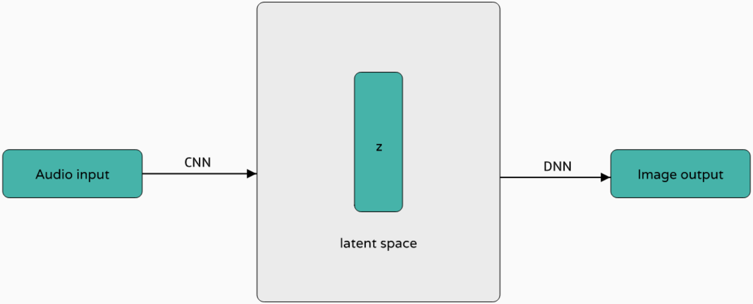
MNIST/FSDD Mix

We apply the following procedure to combine MNIST with FSDD:

- The train set consists of $60,000$ audio-image pairs aligned along labels. The images are unique while the spectrograms are not.
- The test set consists of $10,000$ audio-image pairs aligned along labels. The images are unique while the spectrograms are not.
- Prior to evaluation, this is the only point when labels are used.

VAE

Encoder/decoder setup.



VAE

Audio Encoder	Image Decoder
Input 48x48	Input 64
Conv 4x4, 64, str=2, ReLU	FC 512, ReLU
Conv 4x4, 128, str=2, ReLU	FC 1024, ReLU
FC 1024, ReLU	FC 7x7x128
FC 512, ReLU	Upconv 4x4, 64, str=2, ReLU
μ, σ: FC 64, ReLU	Upconv 4x4, 1, str=2, Sigmoid
Output 2x64	Output 28x28

VAE Training

Hyperparameters:

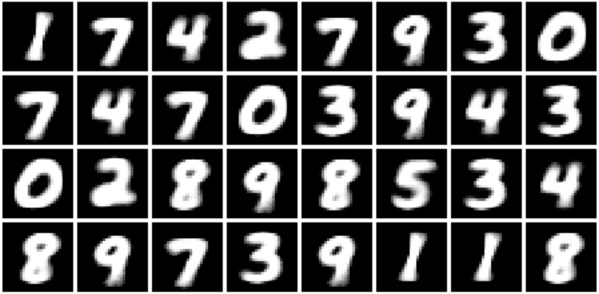
- Epochs $= 100$.
- Batch size $= 128$.
- Learning rate $= 0.001$.

Total number of parameters $= 20,950,529$.

Training time ≈ 2 hours.

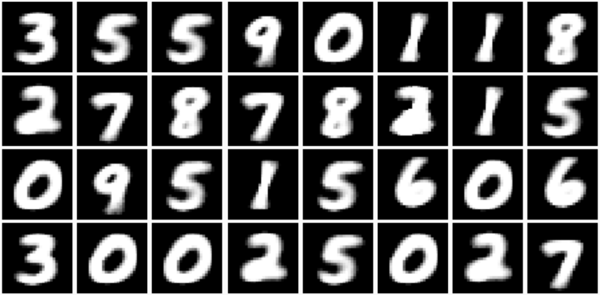
VAE Results

Images generated in last epoch.



VAE Results

Images for test set audio.



VAE Results

Qualitative results:

- Generated images are similar to those from data.
- There seems to be an *archetype* of each digit.

What about a qualitative assessment of these results?

LeNet5

LeNet5
Input 32x32
Conv 1x1, 6, str=5, ReLU
Max Pool 2x2, str=2
Conv 6x6, 16, str=5, ReLU
Max Pool 2x2, str=2
Conv 16x16, 120, str=5, ReLU
Flatten 120
FC 84, ReLU
FC 10, Softmax
Output 10

Trained for 100 epochs, batch size = 128, learning rate = 0.001.
Accuracy on test set = 98.7%.

VAE Results

Quantitative evaluation:

- Present the images generated from test set audio to the pre-trained LeNet5 architecture.
- Measure agreement between the labels predicted by LeNet5 and actual labels.
- LeNet5 is able to correctly classify 93.7% of instances.

It seems that Audio-Image VAE is able to retain a lot of information helpful in image classification.

VAE Results

Some considerations:

- The generated images strongly adhere to specific archetypes.
- There is very little variability between the samples.
- While this could be beneficial in some settings, it could be a hurdle in others.

Is it possible to generate more diverse images?

VAE-GAN

Audio Encoder	Image Decoder	Discriminator
Input 48x48 Conv 4, 128, 2 Conv 4, 256, 2 Conv 4, 512, 2 μ, σ: Conv 4, 64, 2 Output 2x64	Input 64 Upconv 3, 512, 2 Upconv 3, 256, 2 Upconv 2, 128, 2 Upconv 2, 1, 2 Output 28x28	Input 28x28 Conv 4, 128, 2 Conv 4, 256, 2 Conv 4, 512, 2 Conv 1, 1, 1, Sigm Output 1
LeakyReLU(0.2)	ReLU	LeakyReLU(0.2)
BatchNorm	BatchNorm	BatchNorm

There are no fully-conected layers anywhere in this setup. This is crucial.

VAE-GAN

It turns out that using a simple GAN formulation results in images that look like ones from the training set but do not correspond to audio labels. We remedy this by incorporating the reconstruction loss and KL divergence within the adversarial framework:

- $\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})} [\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] + \alpha \mathrm{RL}(G(\mathbf{z}), \mathbf{x}) + \beta D_{\mathrm{KL}}(p_{\mathbf{z}(\mathbf{z})} \mid\mid \mathcal{N}(0, 1))$
- In simpler terms: we add a reconstruction loss and KL divergence term to the Generator loss.
- We set $\beta = 0.5$.

VAE-GAN Training

Hyperparameters:

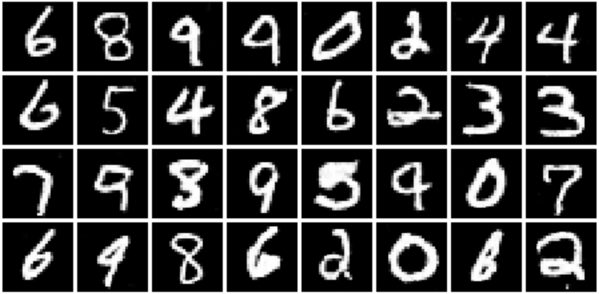
- Epochs $= 100$.
- Batch size $= 128$.
- Learning rate $= 0.0002$.

Total number of parameters $= 7,909,382$.

Training time ≈ 2 hours.

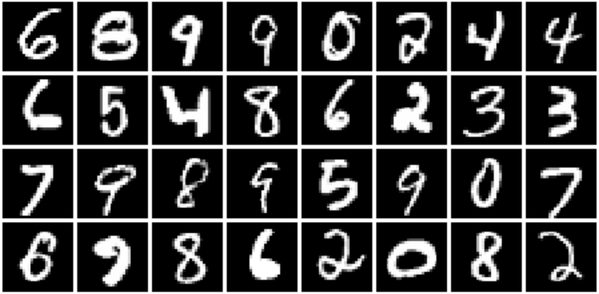
VAE-GAN Results

Images generated in last epoch ($\alpha = 0.2$).



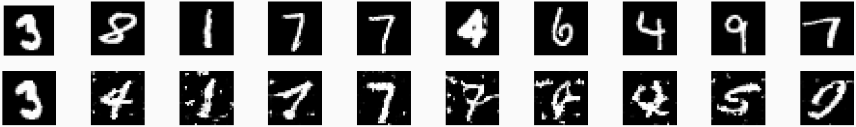
VAE-GAN Results

Real images from training set.



VAE-GAN Results

Actual vs. reconstructed images for the test set.



LeNet5 is able to correctly classify 79% of instances.

VAE-GAN Results

Results:

- A hybrid VAE-GAN approach is able to generate more diverse images from audio.
- The quality of the images still needs improvement.
- In spite of the substandard quality, generated images preserve features important for classification.
- There is a tradeoff between image quality and diversity. Relatively high values of α result in images similar to the ones observed for the VAE setup.

Arandjelovic, R. and Zisserman, A. (2017).
Look, listen and learn.
ICCV.

Arandjelovic, R. and Zisserman, A. (2018).
Objects that sound.
ECCV.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. (2019).
The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality.
Journal of Neuroscience, 39(39):7722–7736.

Goodfellow, I. J., Pouget-Abadie, J., et al. (2014).
Generative adversarial networks.
NIPS.

- Kingma, D. P. and Welling, M. (2014).
  Auto-encoding variational bayes.
  ICLR.
- Noh, H., Hong, S., and Han, B. (2015).
  Learning deconvolution network for semantic segmentation.
  ICCV.