Generative and Multi-modal Networks

Generative adversarial networks and cross-modal retrieval

Maciej Żelaszczyk

December 19, 2018

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

Brief history of neural networks

- 1950-60s: initial models of perceptron.
- "Language is a summer research project."
- 1969: "Perceptrons: an introduction to computational geometry", Minsky and Papert.
- First AI winter. Symbolic AI.
- 1986: Backpropagation rediscovered, Rumelhart, Hinton and Williams.

Brief history of neural networks

- 1991-1994: Hard to train large nets, Hochreiter and Schmidhuber, Bengio.
- Second AI winter. SVMs.
- Ongoing work on RNNs [Hochreiter and Schmidhuber, 1997], CNNs, LeCun 1998, deep nets, Hinton 2006.
- 2012: AlexNet wins ImageNet Large Scale Visual Recognition Challenge, Krizhevsky, Sutskever, Hinton.
- Explosion of deep learning.
- Risk of AI winter (???)

State of deep learning

- Enormous success.
- Mostly relies on CNNs and LSTM variants.
- RL is poster boy.
- Various architecture extensions.
- Architectures geared toward dataset or task.
- Computationally expensive.
- In industry, strong reliance on simpler methods.
- Supervised learning.

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 lablled 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.

Importance of unsupervised learning

What if importance of various kinds of learning is like a cake?

- Pure reinforcement learning = cherry.
- Supervised learning = icing.
- Unsupervised/self-supervised/predictive learning = génoise.
- Perhaps we are still missing a sizeable pie crust? = meta-learning.

Source: LeCun, Y., *The Next Step Towards Artificial Intelligence*

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.
- Attention.
- Potential to be integrated within a meta-learning framework.

Desired architecture

What would we like our architecture to have?

- Unsupervised/weakly-supervised.
- Model of observed data.

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:

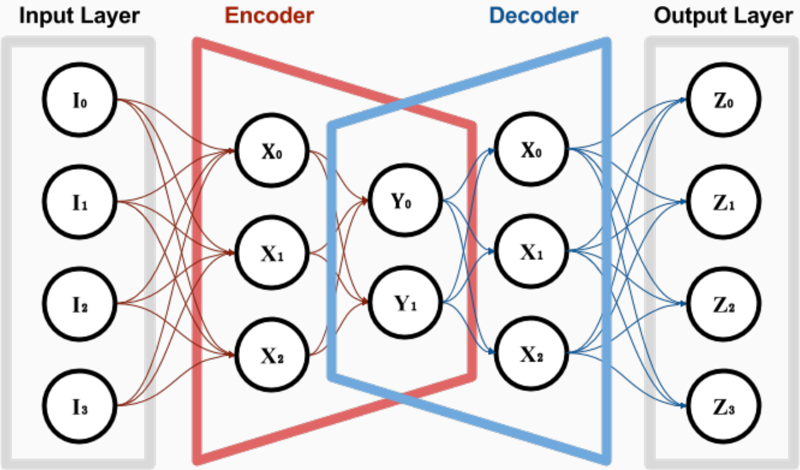
- Autoencoders.
- Variational autoencoders (VAEs).
- Generative adversarial networks (GANs).

Autoencoders

Main idea behind autoencoders:

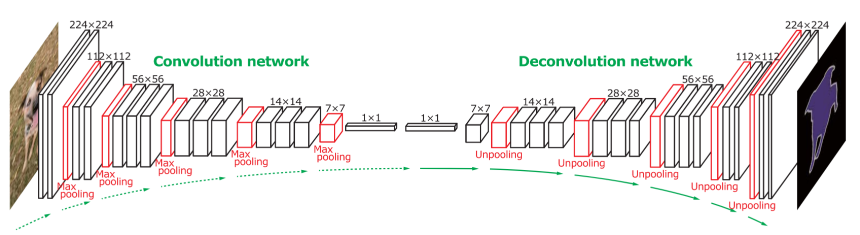
- One network to encode input.
- Second network to decode output.
- Bottleneck in between.
- Latent representation.

Autoencoders



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

Autoencoders

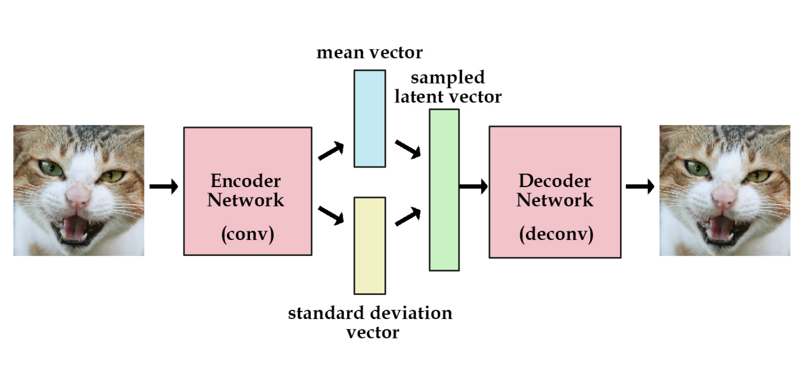


Source: [Noh et al., 2015]

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

More concretely:

- We aim to learn Generator's distribution p_g over data $\mathbf{x}$.
- Define prior $p_{\mathbf{z}}(\mathbf{z})$.
- Represent mapping to data space $G(\mathbf{z}; \theta_g)$.
- G is a neural network parametrized by θ_g.
- Define second neural network $D(\mathbf{x}; \theta_d)$ which outputs single scalar.
- $D(\mathbf{x})$ represents a probability that $\mathbf{x}$ came from the data rather than p_g.

Source: [Goodfellow et al., 2014]

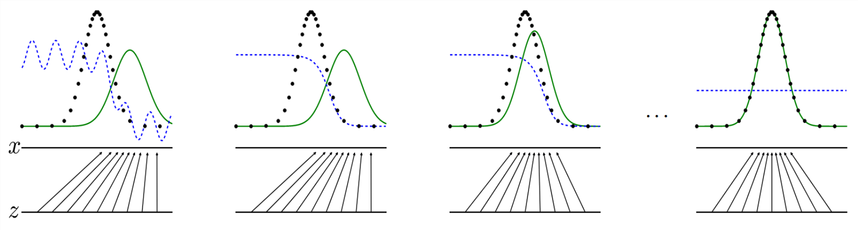
Generative Adversarial Nets

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]

Generative Adversarial Nets



Source: [Goodfellow et al., 2014]

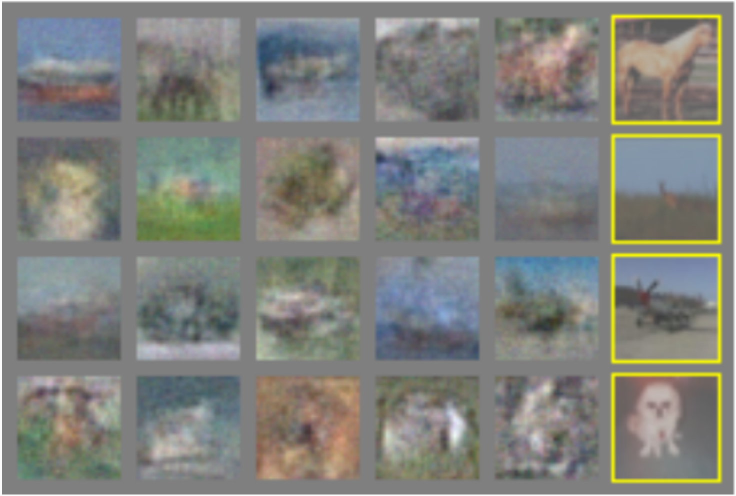
Generative Adversarial Nets

Source: [Goodfellow et al., 2014]

Generative Adversarial Nets

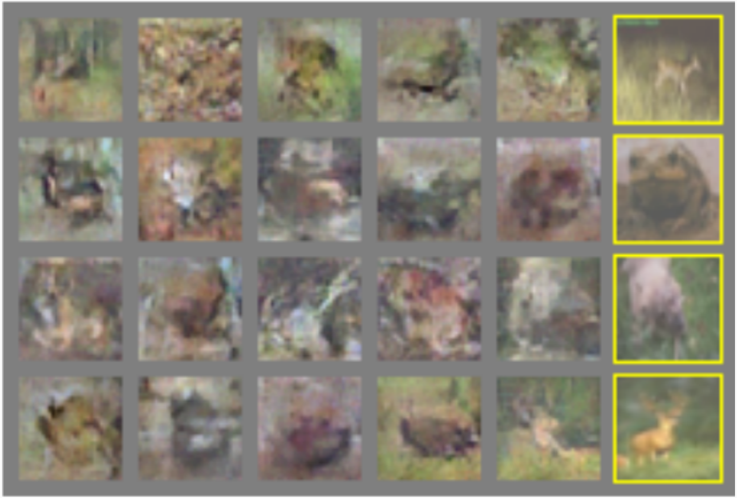
Source: [Goodfellow et al., 2014]

Generative Adversarial Nets



Source: [Goodfellow et al., 2014]

Generative Adversarial Nets



Source: [Goodfellow et al., 2014]

Generative Adversarial Nets

Source: [Radford et al., 2016]

Generative Adversarial Nets

Source: [Radford et al., 2016]

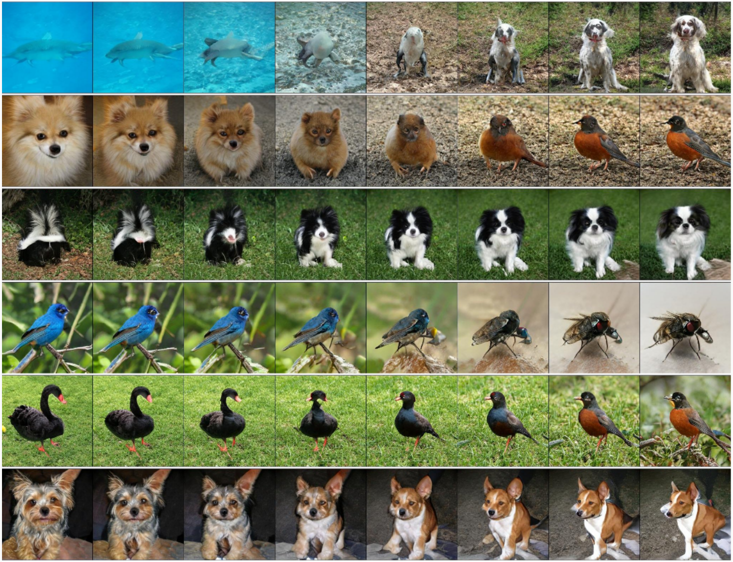
Generative Adversarial Nets

Source: [Brock et al., 2018]

Generative Adversarial Nets

Source: [Brock et al., 2018]

Generative Adversarial Nets



Source: [Brock et al., 2018]

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.

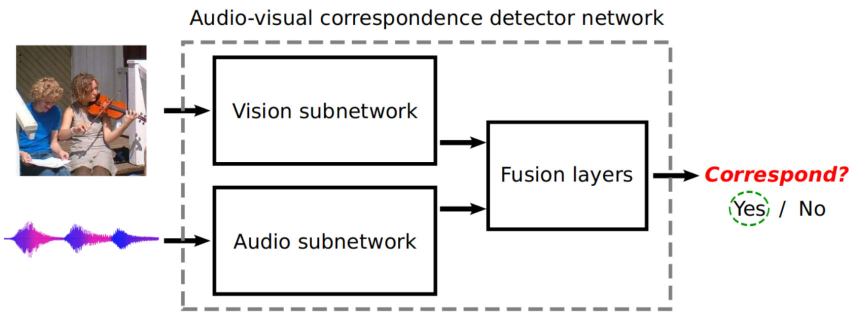
Multi-modal representation

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.

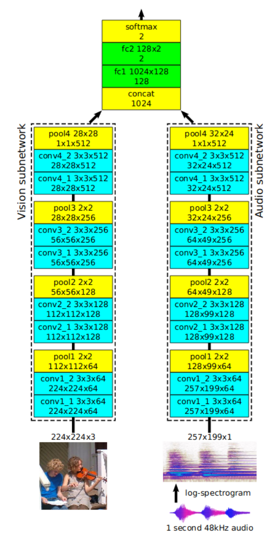
Audio-visual correspondence

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



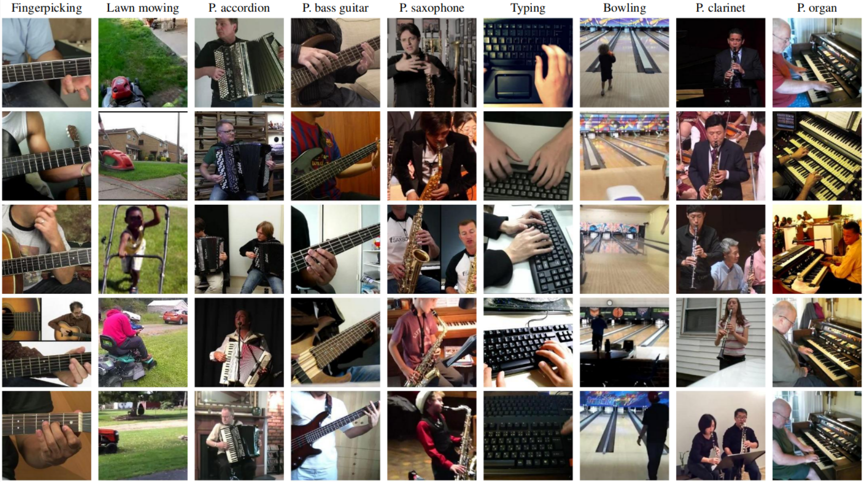
Source: [Arandjelovic and Zisserman, 2017]

Audio-visual correspondence



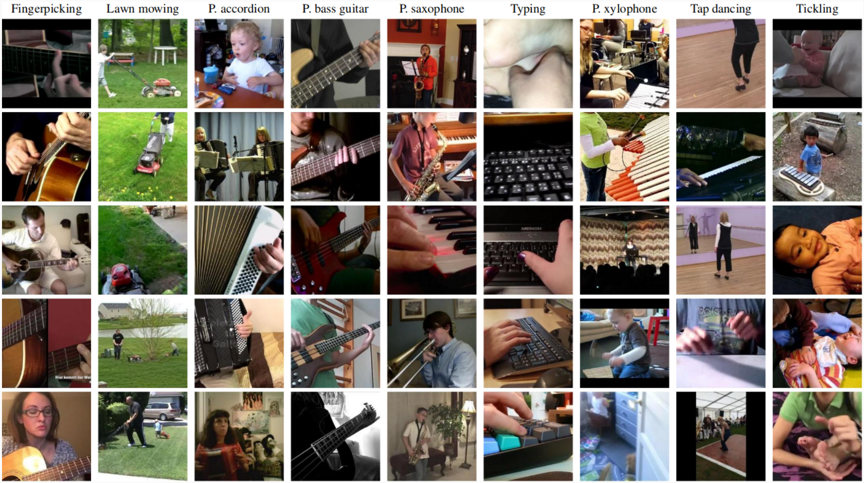
Source: [Arandjelovic and Zisserman, 2017]

Audio-visual correspondence



Source: [Arandjelovic and Zisserman, 2017]

Audio-visual correspondence



Source: [Arandjelovic and Zisserman, 2017]

Audio-visual correspondence

Method	Flickr-SoundNet	Kinetics-Sounds
Supervised direct	–	65%
Supervised pretraining	–	74%
L^3-Net	78%	74%

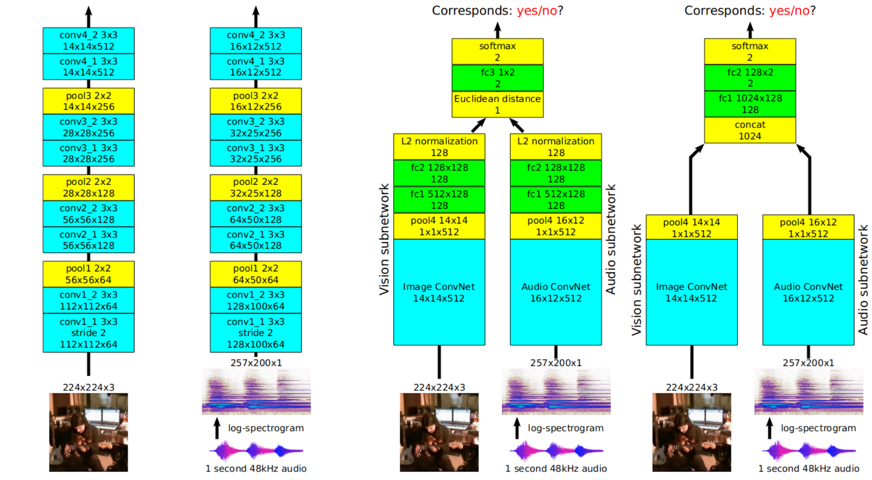
Source: [Arandjelovic and Zisserman, 2017]

Cross-modal retrieval

So far, AVC only shows whether audio and visual data correspond. The data are not aligned in any systematic way.

- We would want to align audio and visual features.
- This would allow to retrieve data from one modality based on the other.
- Answer the question: "What object in the frame is making the sound?"

Cross-modal retrieval



(a) Vision ConvNet (b) Audio ConvNet (c) AVE-Net (d) L^3-Net [4]

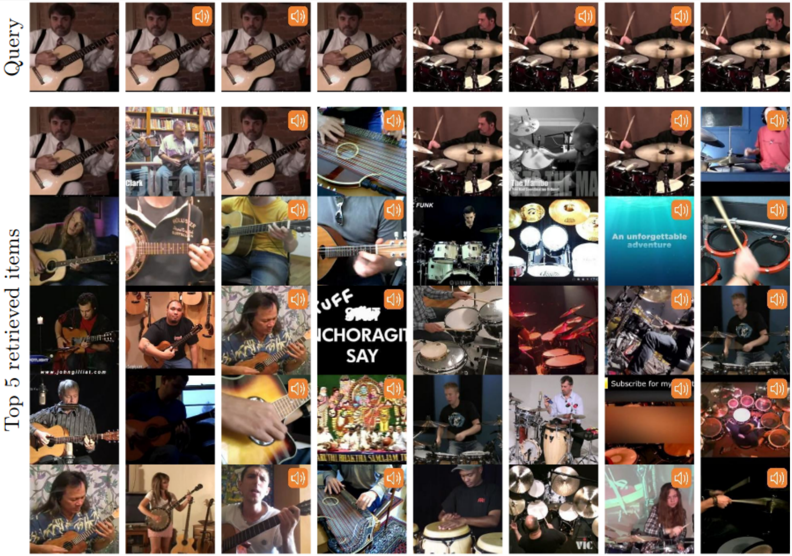
Source: [Arandjelovic and Zisserman, 2018]

Cross-modal retrieval

Method	im-im	im-aud	aud-im	aud-aud
Random chance	.407	.407	.407	.407
L^3-Net [4]	.567	.418	.385	.653
L^3-Net with CCA	.578	.531	.560	.649
VGG16-ImageNet [29]	.600	–	–	–
VGG16-ImageNet + L^3-Audio CCA	.493	.458	.464	.618
AVE-Net	**.604**	**.561**	**.587**	**.665**

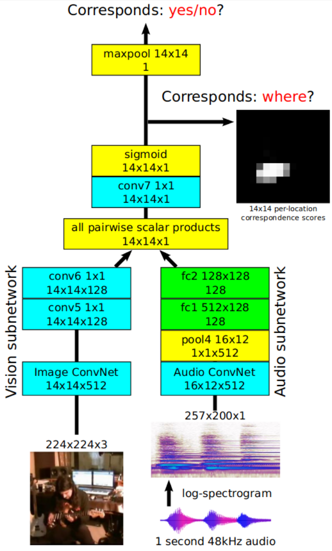
Source: [Arandjelovic and Zisserman, 2018]

Cross-modal retrieval



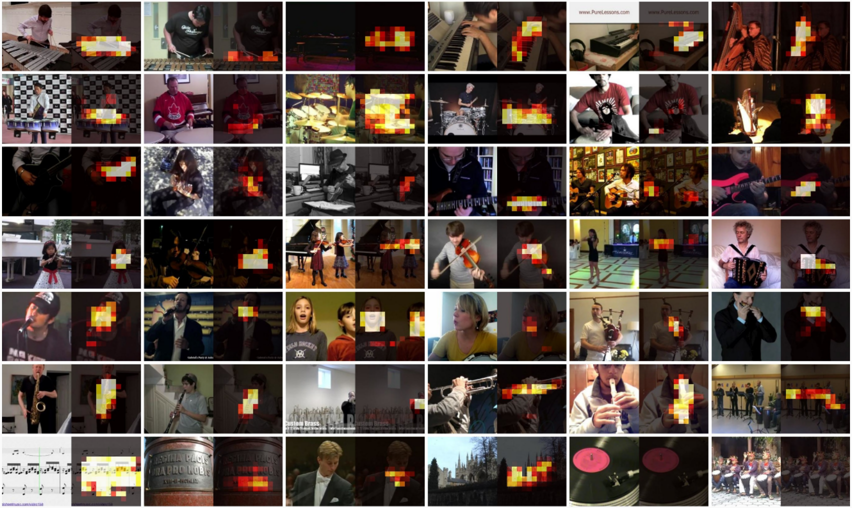
Source: [Arandjelovic and Zisserman, 2018]

Cross-modal retrieval



Source: [Arandjelovic and Zisserman, 2018]

Cross-modal retrieval



Source: [Arandjelovic and Zisserman, 2018]

Cross-modal retrieval

Source: [Arandjelovic and Zisserman, 2018]

Cross-modal retrieval

Source: [Arandjelovic and Zisserman, 2018]

Cross-modal retrieval

Source: [Arandjelovic and Zisserman, 2018]

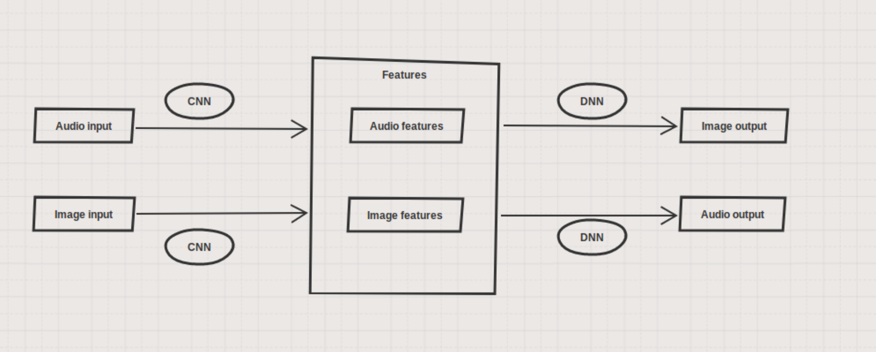
Idea

What if we can use AVC to generate visual/audio data?

- Use AVC setup on audio/visual pairs.
- Separate audio/visual encoders.
- Mix data from them into one representation.
- Use this representation to separate data via audio/visual decoders.
- Multiple ways to train this.
- Ideally, we would like to train adversarially.

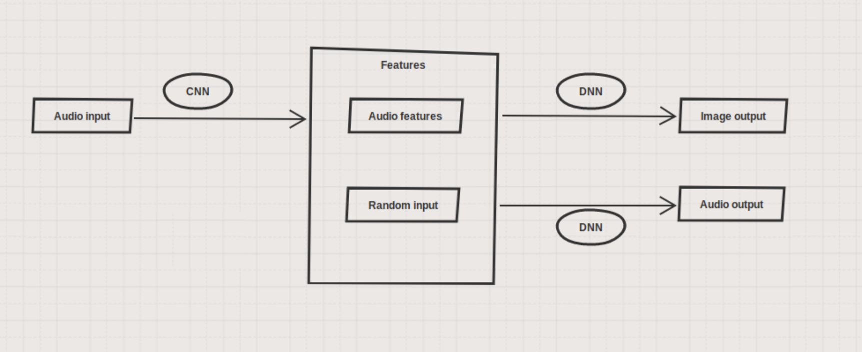
Idea

Audio-visual network.



Idea

Use audio-visual network to generate data.



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

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

Brock, A., Donahue, J., and Simonyan, K. (2018).
Large scale gan training for high fidelity natural image synthesis.
arXiv.

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

Hochreiter, S. and Schmidhuber, J. (1997).
Long short-term memory.

Neural Computation, 9(8):1735–1780.

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.

Radford, A., Metz, L., and Chintala, S. (2016).
Unsupervised representation learning with deep convolutional generative adversarial networks.
ICLR.