A path to unsupervised learning

through Adversarial Networks

Soumith Chintala

Facebook AI Research

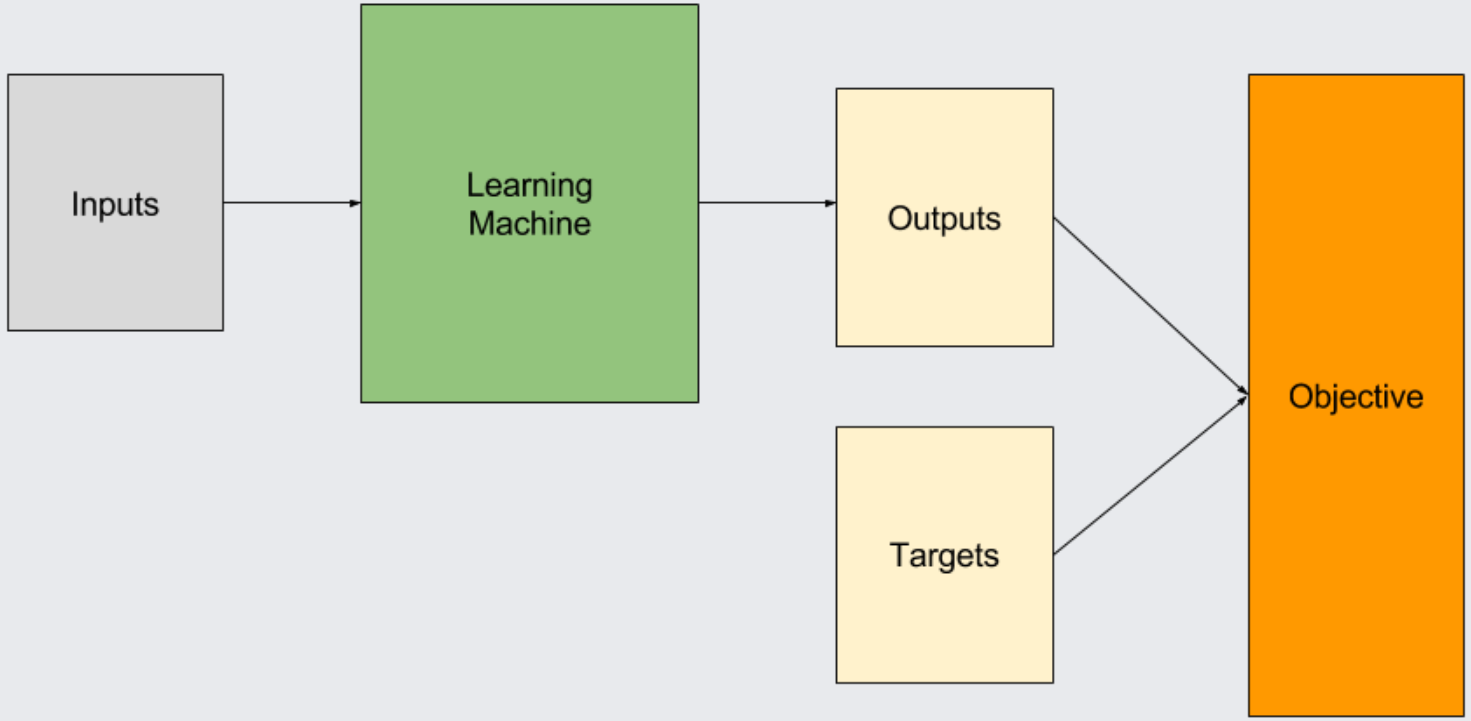
Overview

of the talk

- Unsupervised Learning
- Generative Adversarial Networks
- Advances
- Using the learnt representations
- What's next?

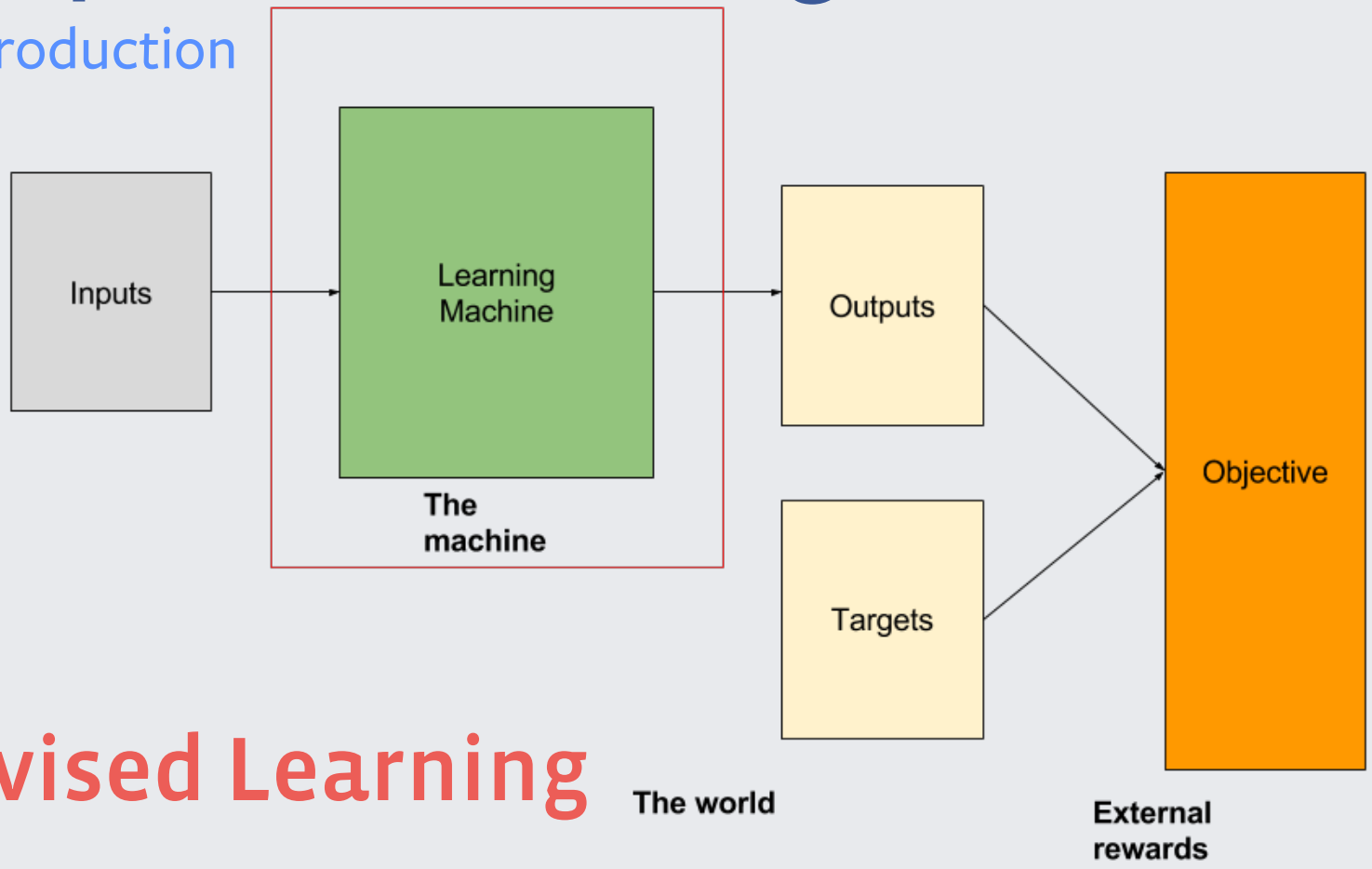
Unsupervised Learning

An introduction



Unsupervised Learning

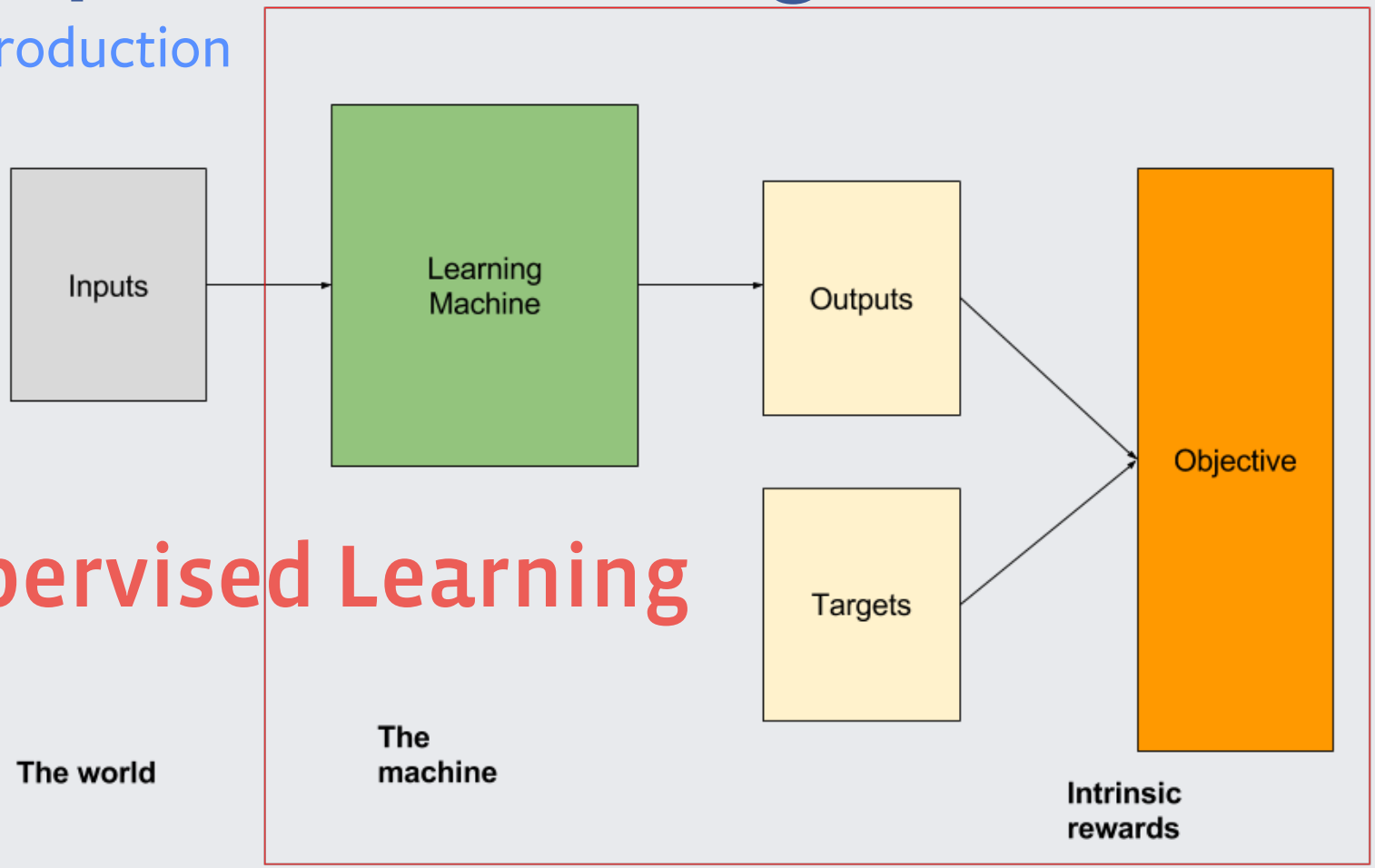
An introduction



Supervised Learning

Unsupervised Learning

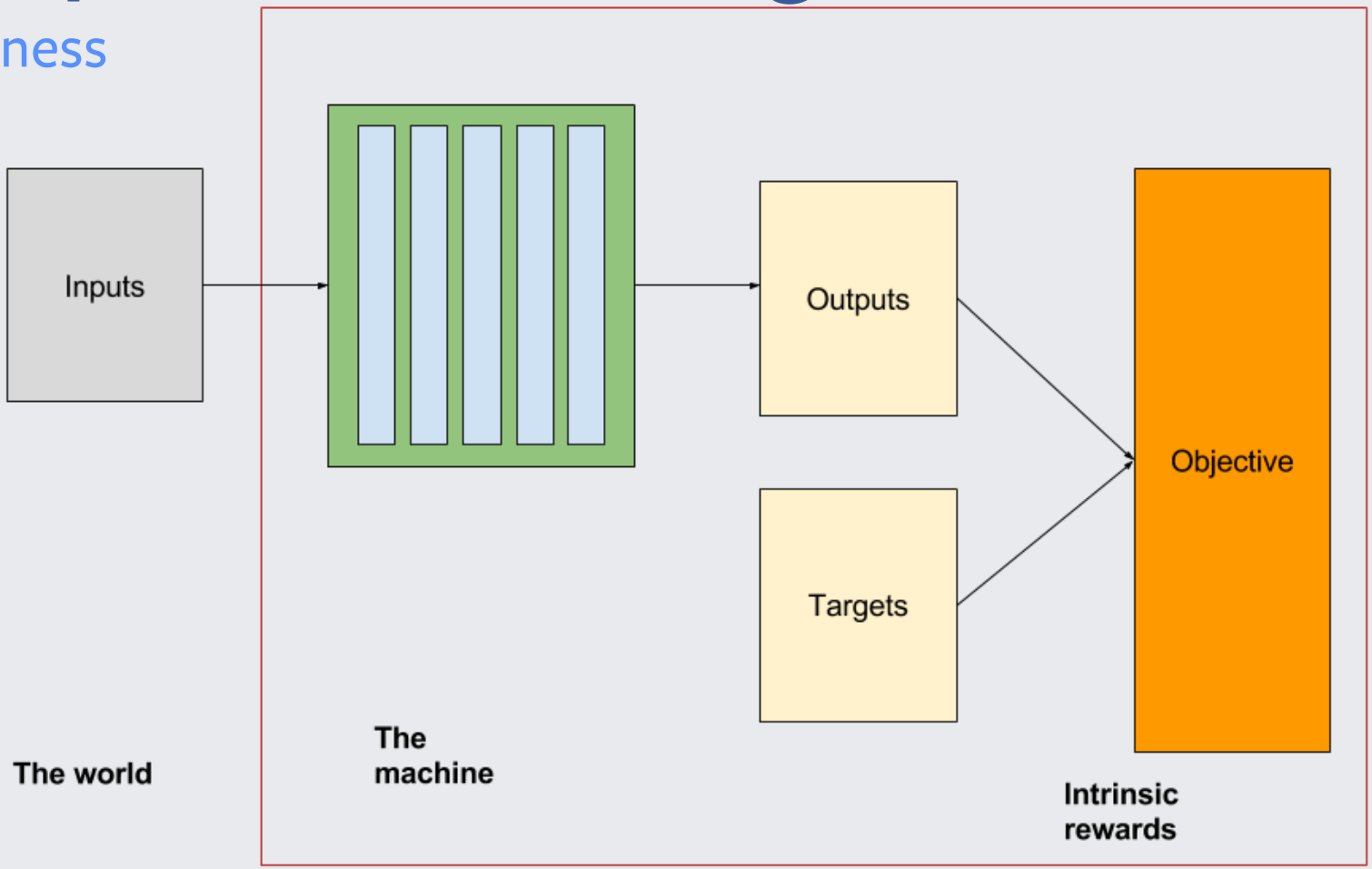
An introduction



Unsupervised Learning

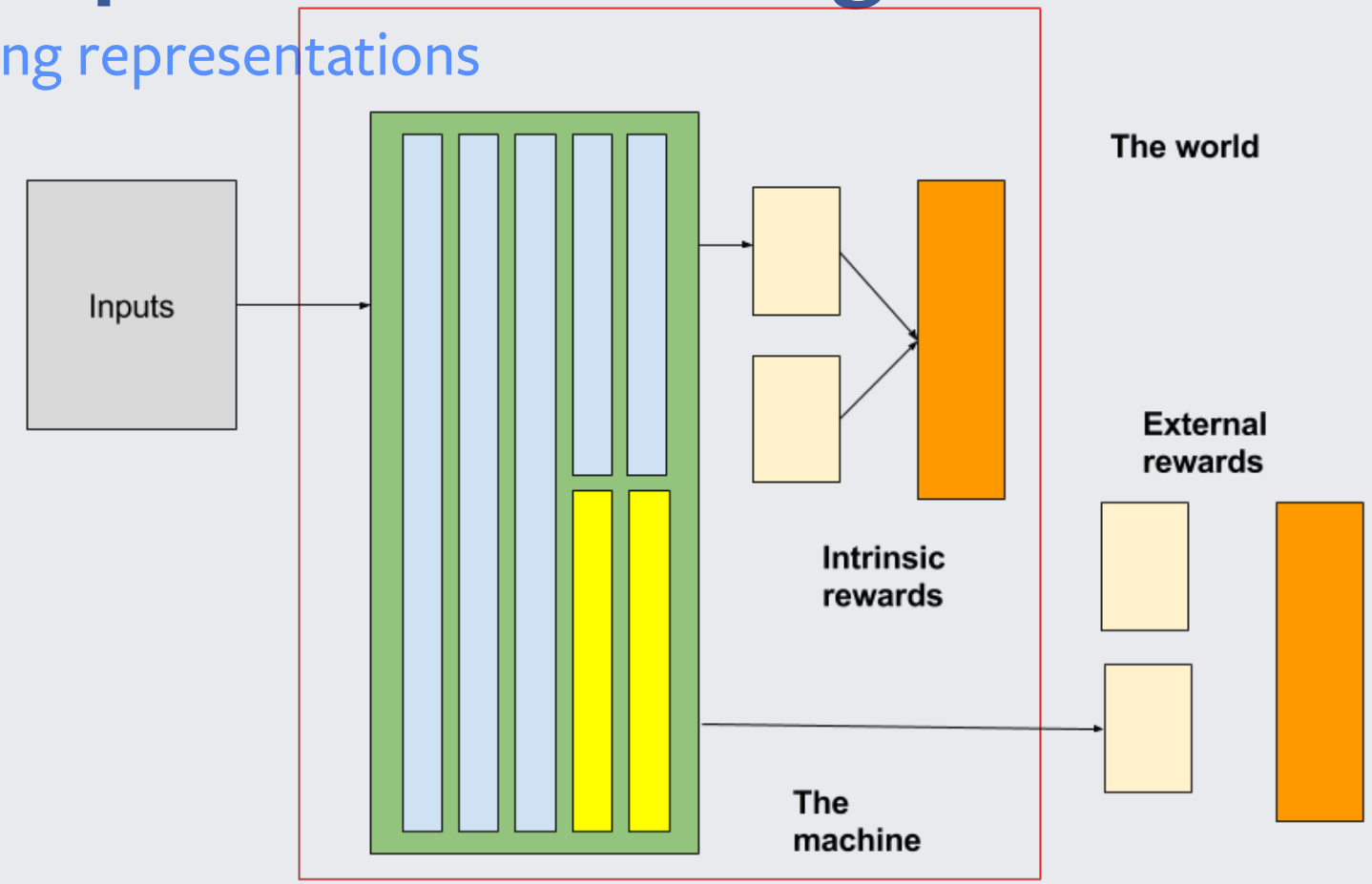
Unsupervised Learning

Usefulness



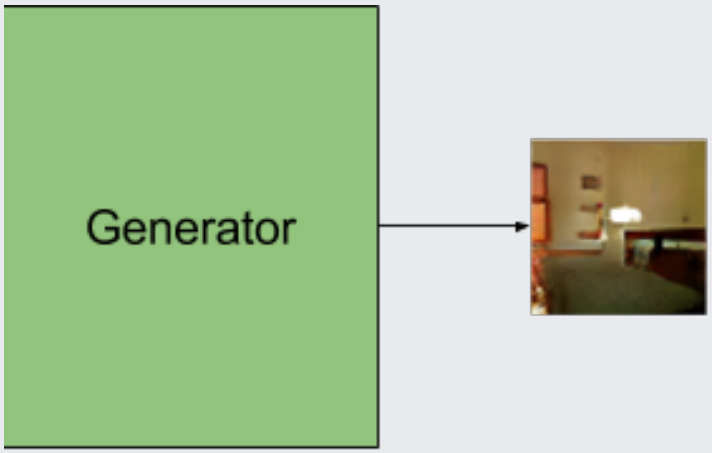
Unsupervised Learning

Reusing representations



Generative Models

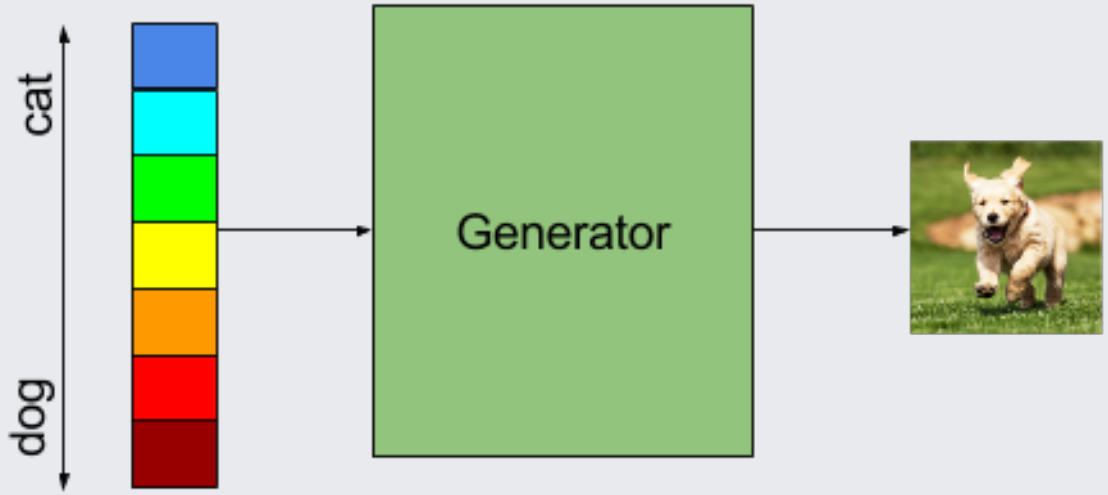
An introduction



A model that learns a distribution of images

Generative Models

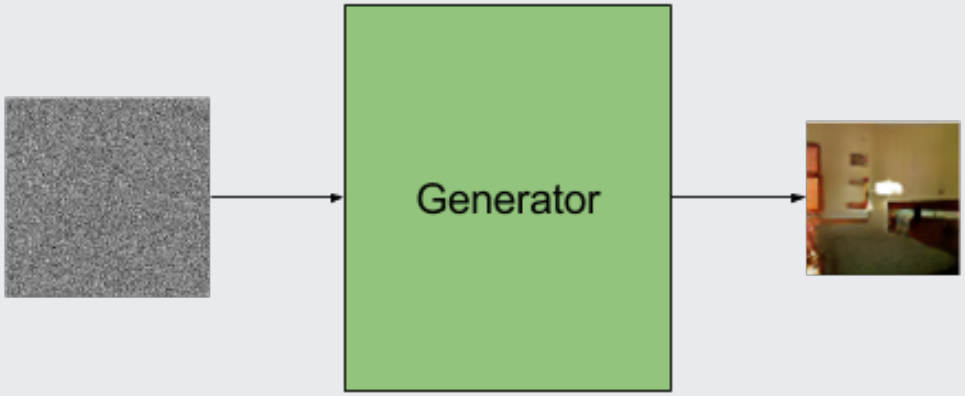
An introduction



X = P(z), z controls dogness or catness

Generative Models

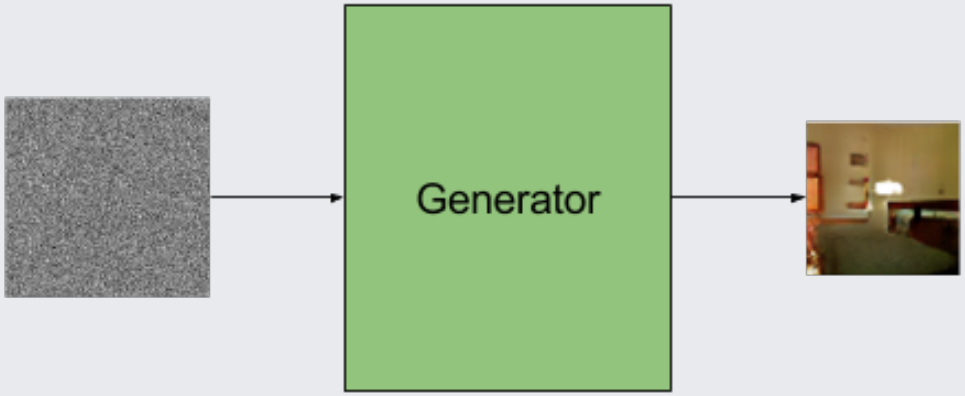
An introduction



X = P(z), z is a latent variable

Generative Models

An introduction



P(z) = neural network

Generative Adversarial Networks

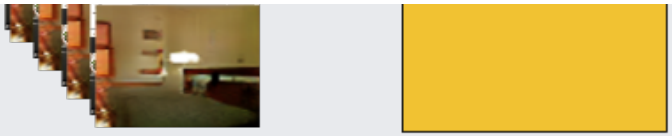
Generative Adversarial Networks

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio

(Submitted on 10 Jun 2014)

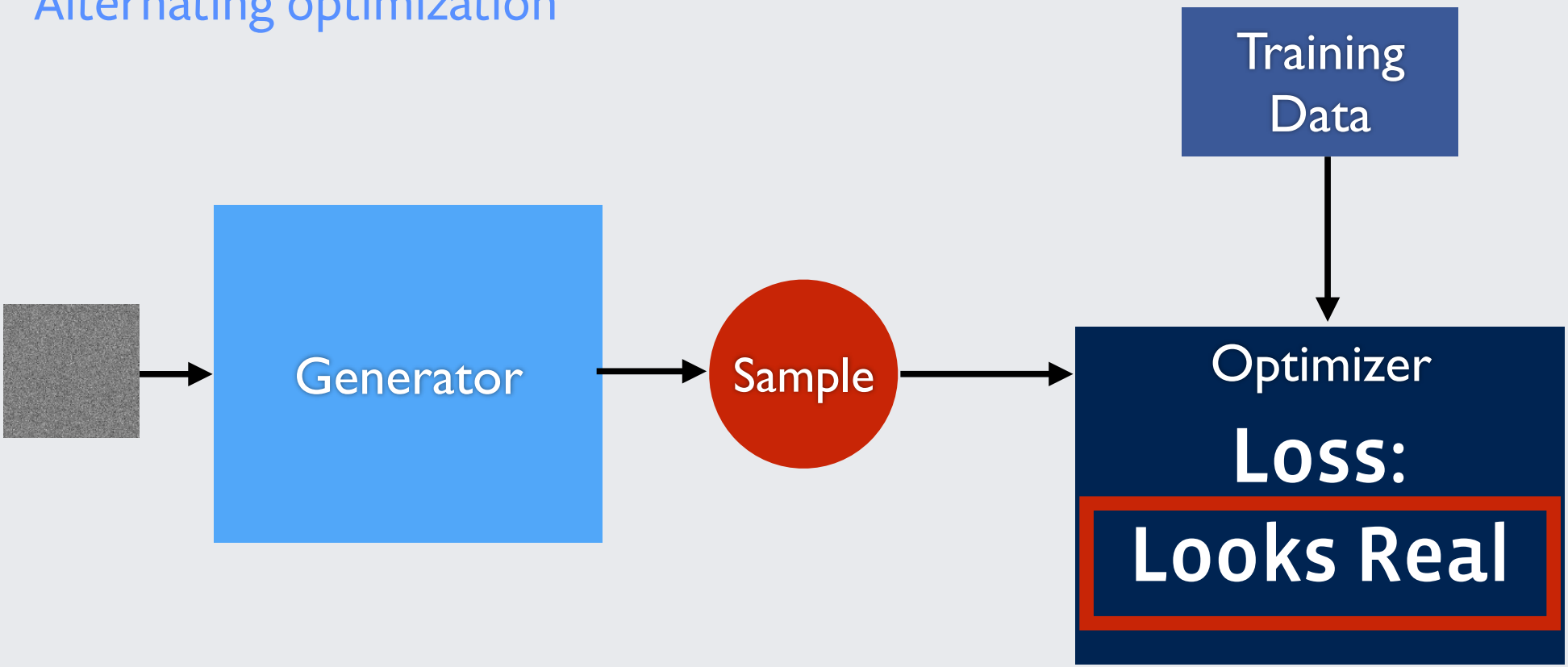
We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to 1/2 everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

Subjects: **Machine Learning (stat.ML)**; Learning (cs.LG)
Cite as: **arXiv:1406.2661 [stat.ML]**
(or **arXiv:1406.2661v1 [stat.ML]** for this version)

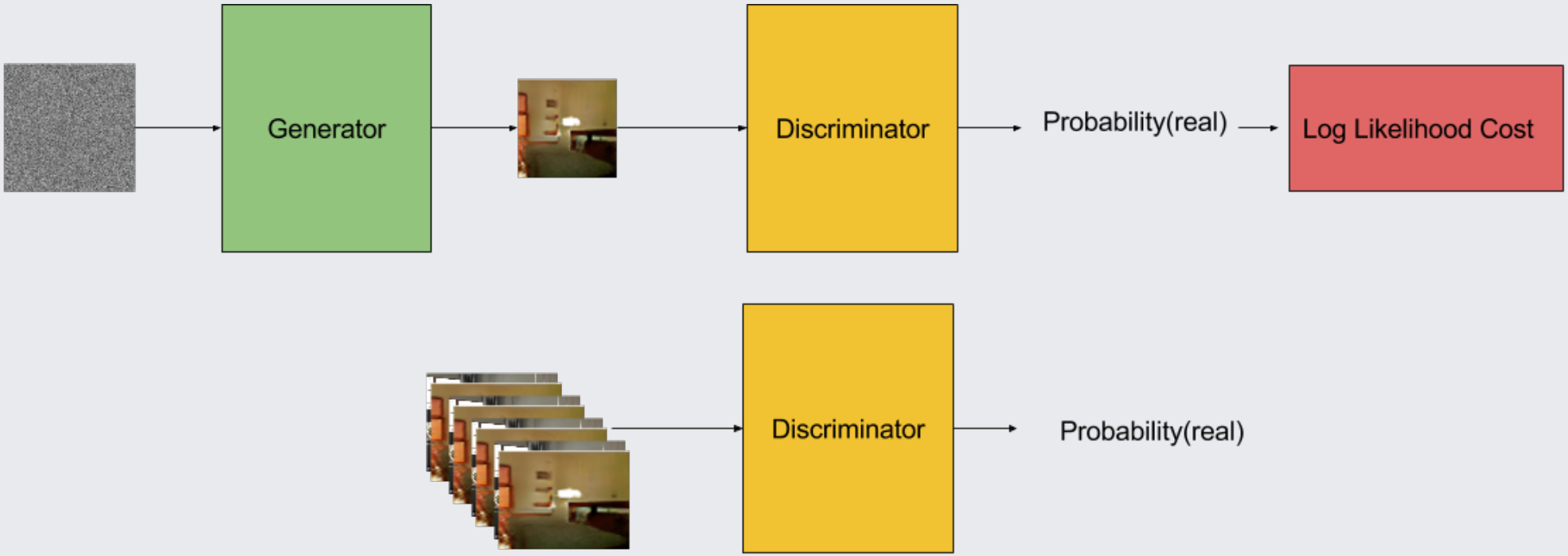


Generative Adversarial Networks

Alternating optimization

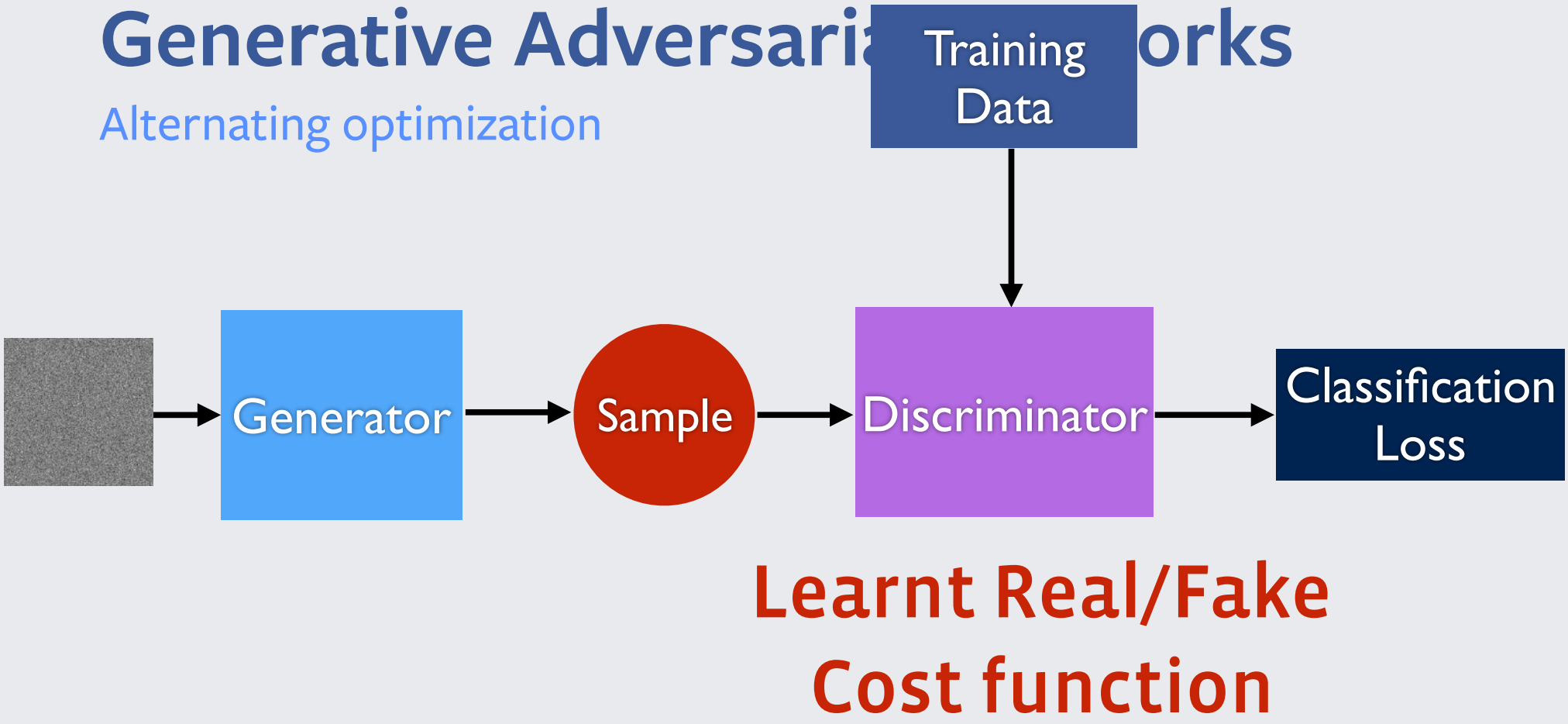


Generative Adversarial Networks



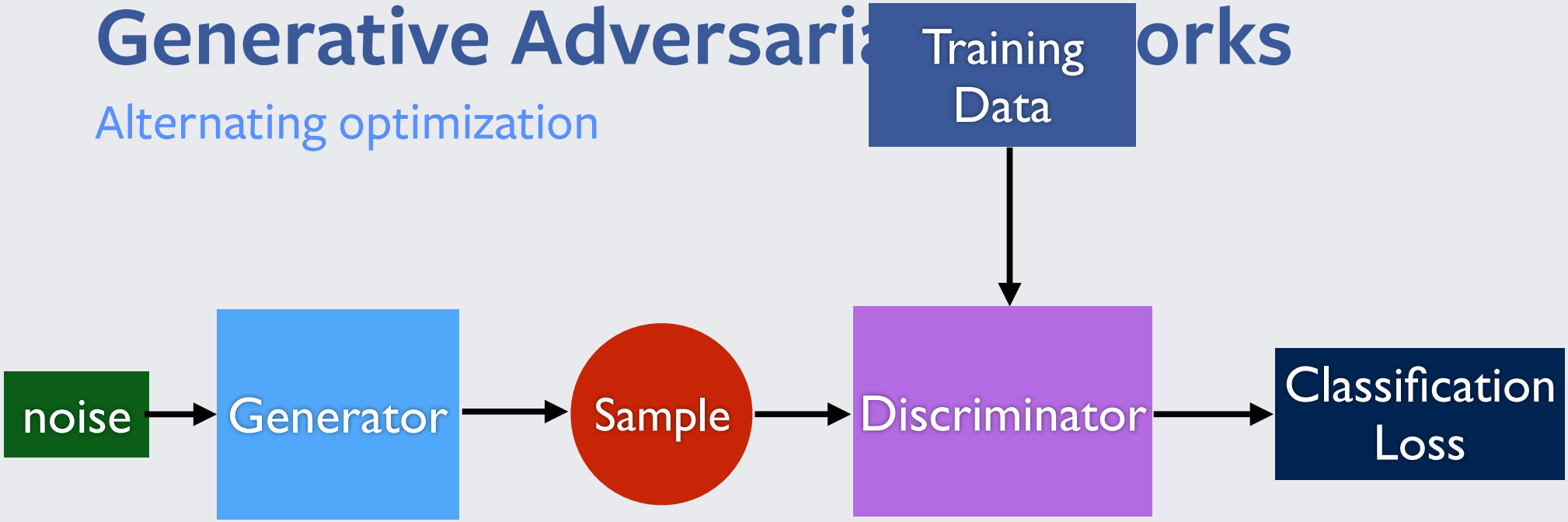
Generative Adversaria orks

Alternating optimization



Generative Adversaria orks

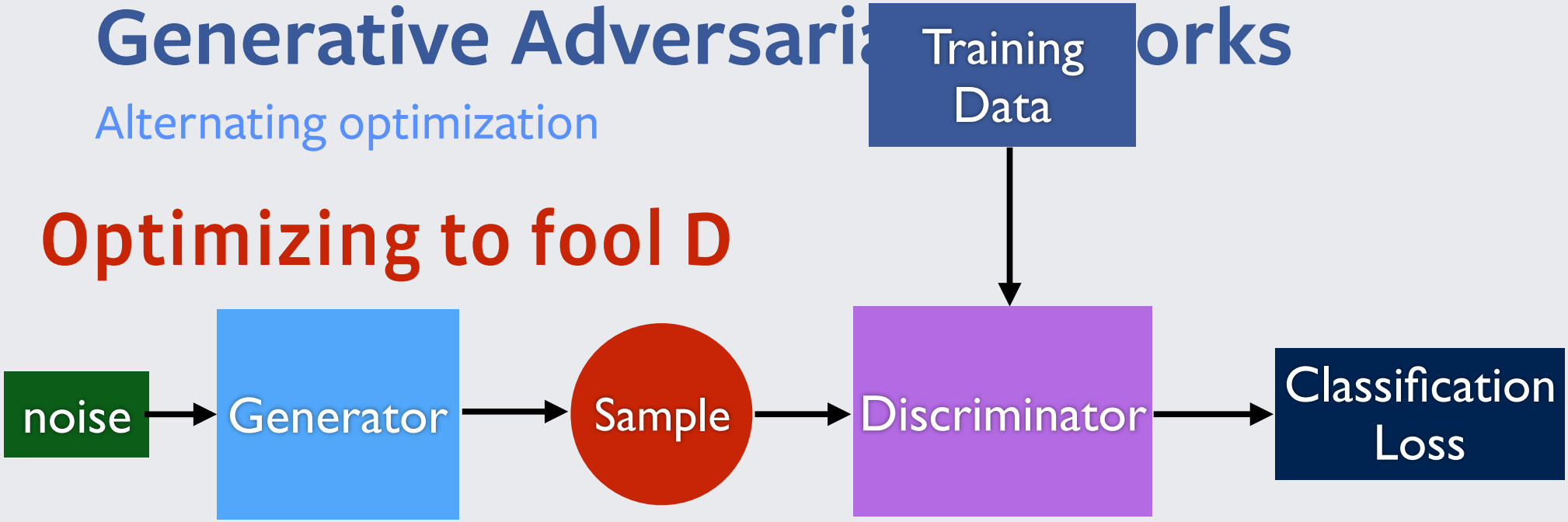
Alternating optimization



Trained via Gradient Descent

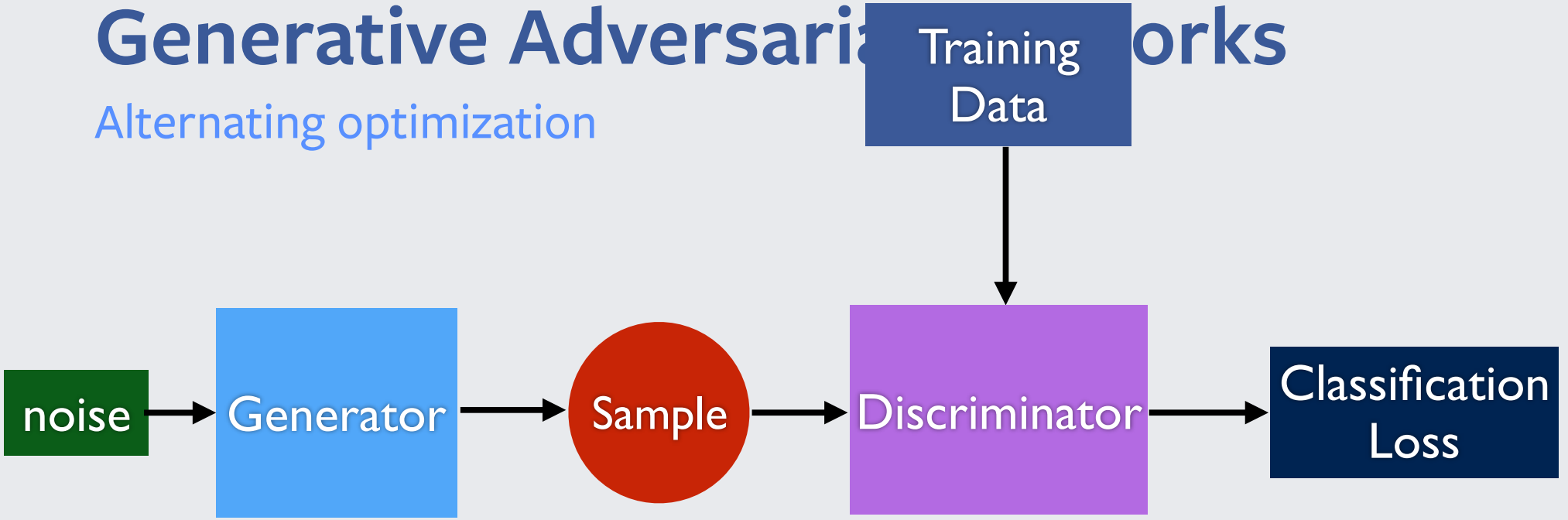
Generative Adversaria orks

Alternating optimization



Generative Adversaria orks

Alternating optimization



Optimizing to not get fooled by G

Generative Adversarial Networks

Optimizes Jensen-Shannon Divergence

Theorem 1. *The global minimum of the virtual training criterion $C(G)$ is achieved if and only if $p_g = p_{data}$. At that point, $C(G)$ achieves the value $-\log 4$.*

Proof. For $p_g = p_{\text{data}}$, $D_G^*(\boldsymbol{x}) = \frac{1}{2}$, (consider Eq. 2). Hence, by inspecting Eq. 4 at $D_G^*(\boldsymbol{x}) = \frac{1}{2}$, we find $C(G) = \log \frac{1}{2} + \log \frac{1}{2} = -\log 4$. To see that this is the best possible value of $C(G)$, reached only for $p_g = p_{\text{data}}$, observe that

$$\mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} [-\log 2] + \mathbb{E}_{\boldsymbol{x} \sim p_g} [-\log 2] = -\log 4$$

and that by subtracting this expression from $C(G) = V(D_G^*, G)$, we obtain:

$$C(G) = -\log(4) + KL\left(p_{\text{data}} \middle\| \frac{p_{\text{data}} + p_g}{2}\right) + KL\left(p_g \middle\| \frac{p_{\text{data}} + p_g}{2}\right) \tag{5}$$

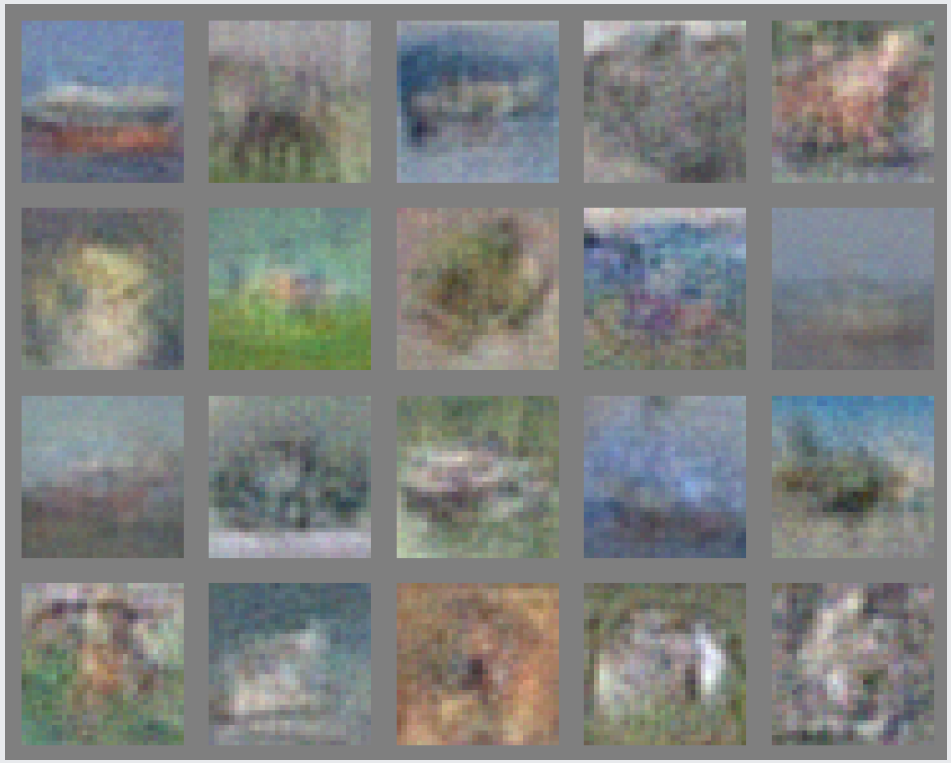
where KL is the Kullback–Leibler divergence. We recognize in the previous expression the Jensen–Shannon divergence between the model's distribution and the data generating process:

$$C(G) = -\log(4) + 2 \cdot JSD\left(p_{\text{data}} \| p_g\right) \tag{6}$$

Since the Jensen–Shannon divergence between two distributions is always non-negative and zero only when they are equal, we have shown that $C^* = -\log(4)$ is the global minimum of $C(G)$ and that the only solution is $p_g = p_{\text{data}}$, i.e., the generative model perfectly replicating the data generating process. □

Generative Adversarial Networks

Samples



Class-conditional GANs

arxiv.org/abs/1506.05751

Cornell University Library

arXiv.org > cs > arXiv:1506.05751

Search

Computer Science > Computer Vision and Pattern Recognition

Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks

Emily Denton, Soumith Chintala, Arthur Szlam, Rob Fergus

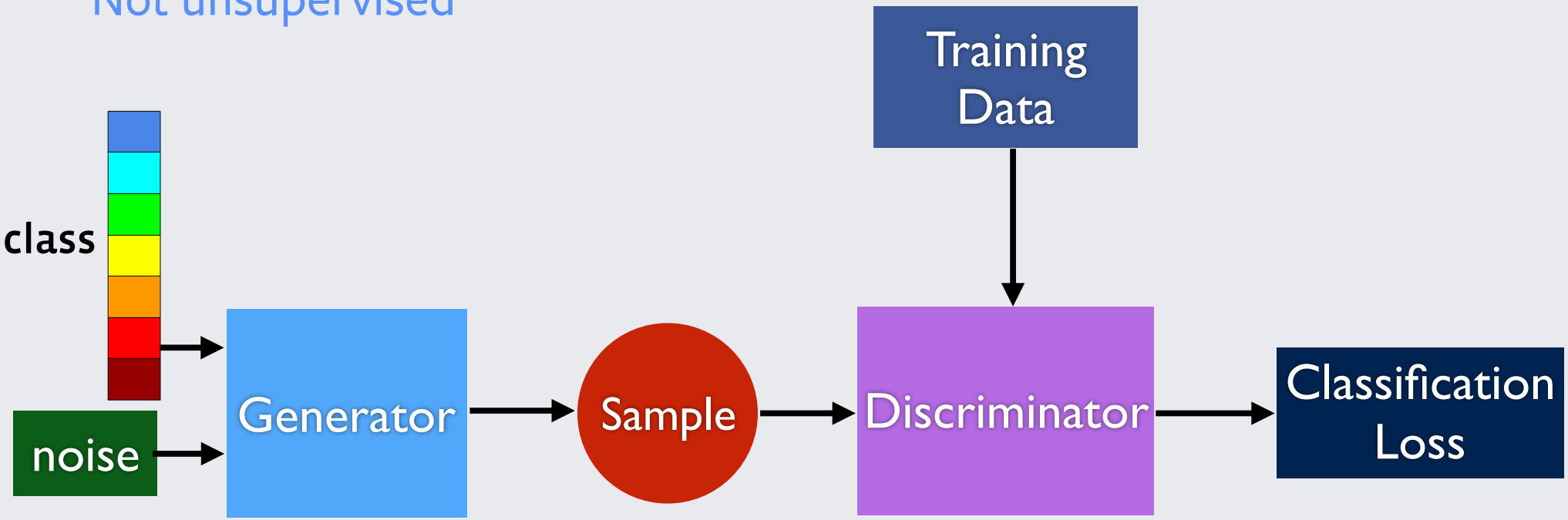
(Submitted on 18 Jun 2015)

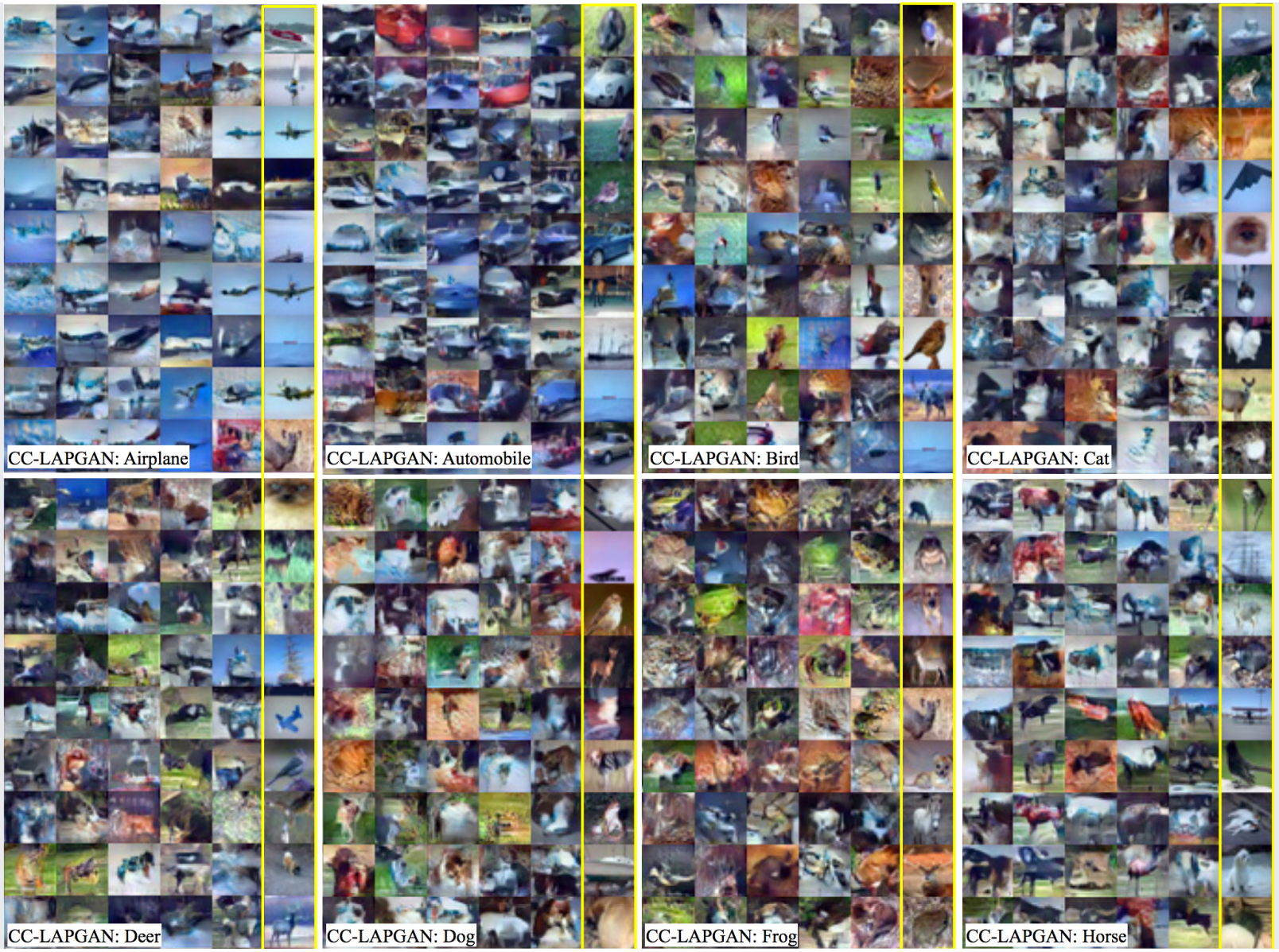
In this paper we introduce a generative parametric model capable of producing high quality samples of natural images. Our approach uses a cascade of convolutional networks within a Laplacian pyramid framework to generate images in a coarse-to-fine fashion. At each level of the pyramid, a separate generative convnet model is trained using the Generative Adversarial Nets (GAN) approach (Goodfellow et al.). Samples drawn from our model are of significantly higher quality than alternate approaches. In a quantitative assessment by human evaluators, our CIFAR10 samples were mistaken for real images around 40% of the time, compared to 10% for samples drawn from a GAN baseline model. We also show samples from models trained on the higher resolution images of the LSUN scene dataset.

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**
Cite as: **arXiv:1506.05751 [cs.CV]**
(or **arXiv:1506.05751v1 [cs.CV]** for this version)

Class-conditional GANs

Not unsupervised




CC-LAPGAN: Airplane
CC-LAPGAN: Automobile
CC-LAPGAN: Bird
CC-LAPGAN: Cat
CC-LAPGAN: Deer
CC-LAPGAN: Dog
CC-LAPGAN: Frog
CC-LAPGAN: Horse

Video Prediction GANs

arxiv.org/abs/1511.05440

Cornell University Library

arXiv.org > cs > arXiv:1511.05440

Computer Science > Learning

Deep multi-scale video prediction beyond mean square error

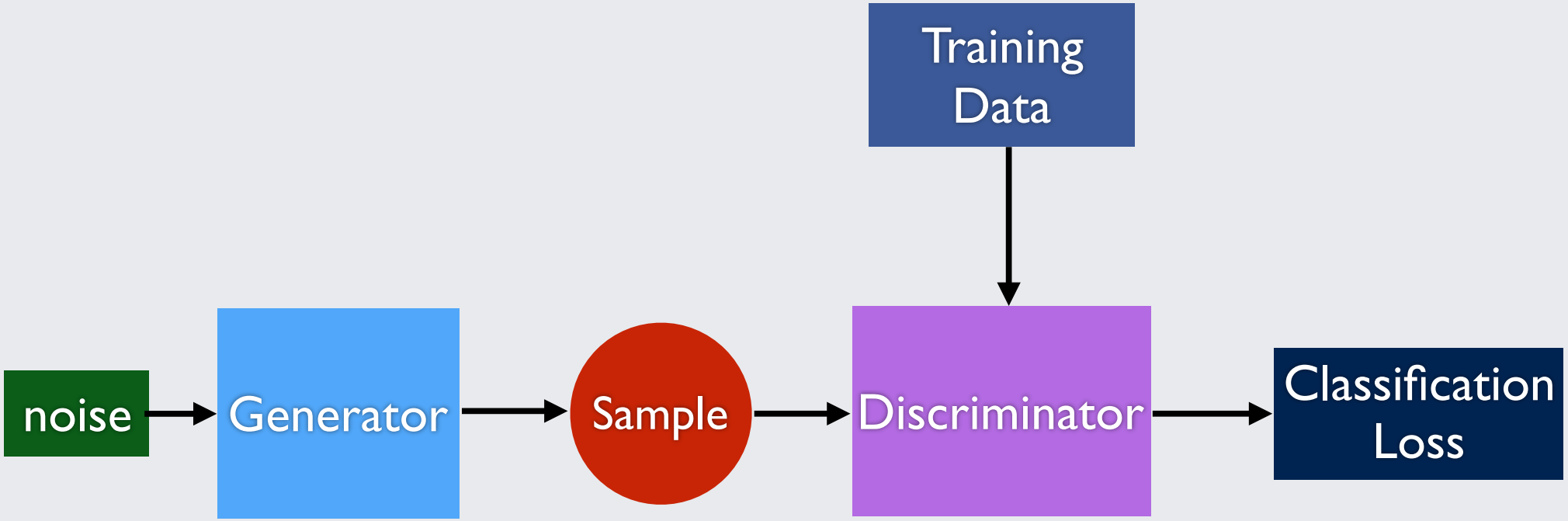
Michael Mathieu, Camille Couprie, Yann LeCun

(Submitted on 17 Nov 2015 (v1), last revised 26 Feb 2016 (this version, v6))

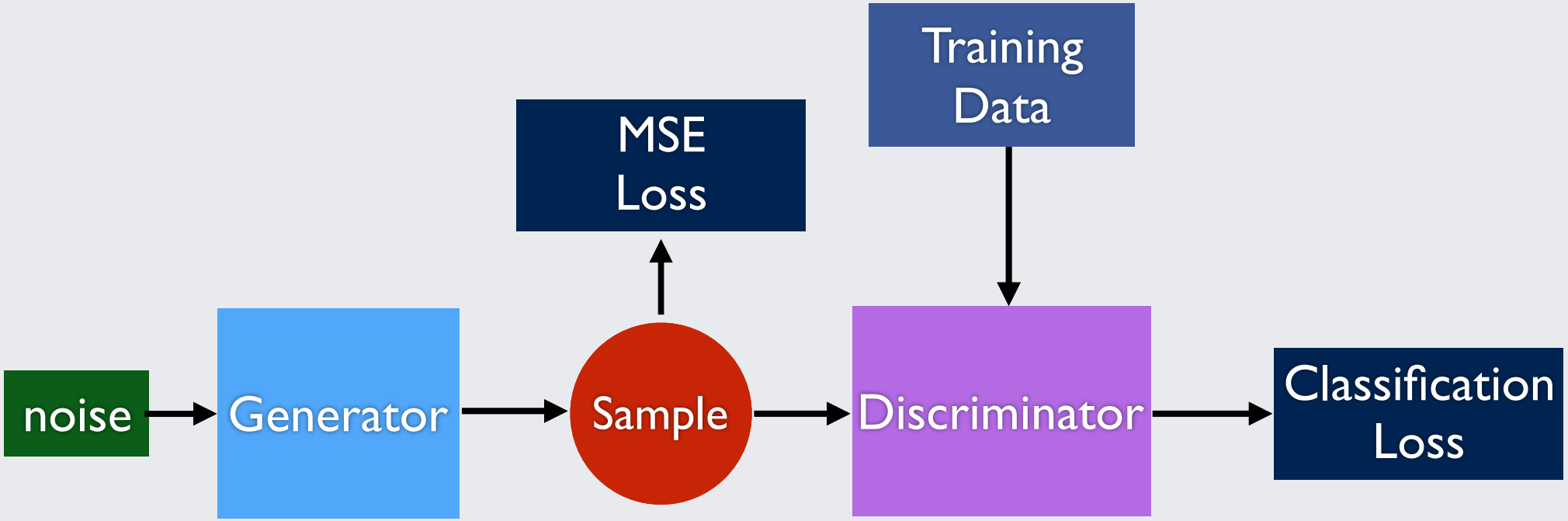
Learning to predict future images from a video sequence involves the construction of an internal representation that models the image evolution accurately, and therefore, to some degree, its content and dynamics. This is why pixel-space video prediction may be viewed as a promising avenue for unsupervised feature learning. In addition, while optical flow has been a very studied problem in computer vision for a long time, future frame prediction is rarely approached. Still, many vision applications could benefit from the knowledge of the next frames of videos, that does not require the complexity of tracking every pixel trajectories. In this work, we train a convolutional network to generate future frames given an input sequence. To deal with the inherently blurry predictions obtained from the standard Mean Squared Error (MSE) loss function, we propose three different and complementary feature learning strategies: a multi-scale architecture, an adversarial training method, and an image gradient difference loss function. We compare our predictions to different published results based on recurrent neural networks on the UCF101 dataset

Subjects: **Learning (cs.LG)**; Computer Vision and Pattern Recognition (cs.CV); Machine Learning (stat.ML)
Cite as: **arXiv:1511.05440 [cs.LG]**
(or **arXiv:1511.05440v6 [cs.LG]** for this version)

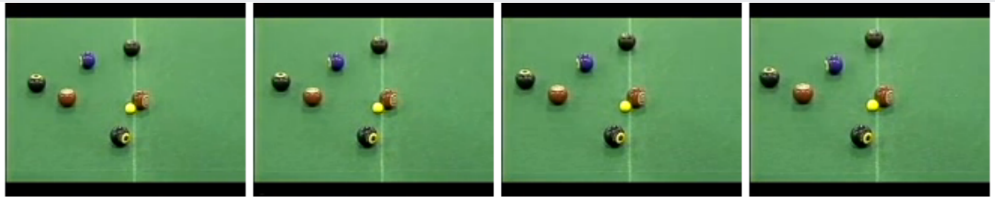
Video Prediction GANs



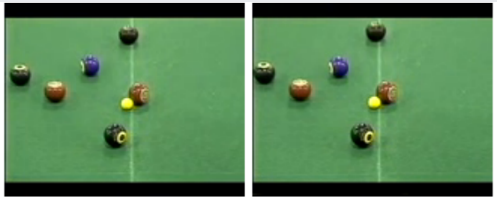
Video Prediction GANs



Video Prediction GANs

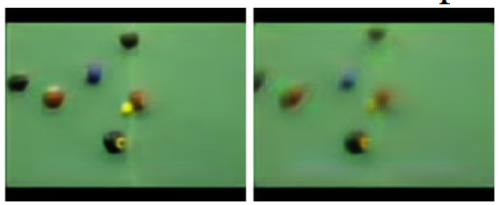


Input frames

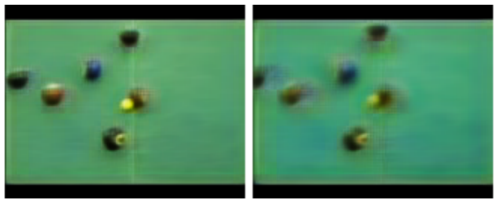


Ground truth

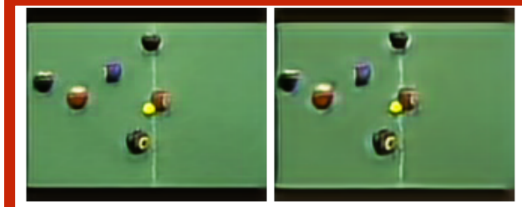
ℓ_2 result



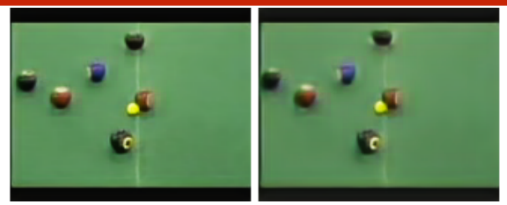
ℓ_1 result



GDL ℓ_1 result



Adversarial result



Adversarial+GDL result

DCGANs

arxiv.org/abs/1511.06434

Cornell University Library

arXiv.org > cs > arXiv:1511.06434

Search

Computer Science > Learning

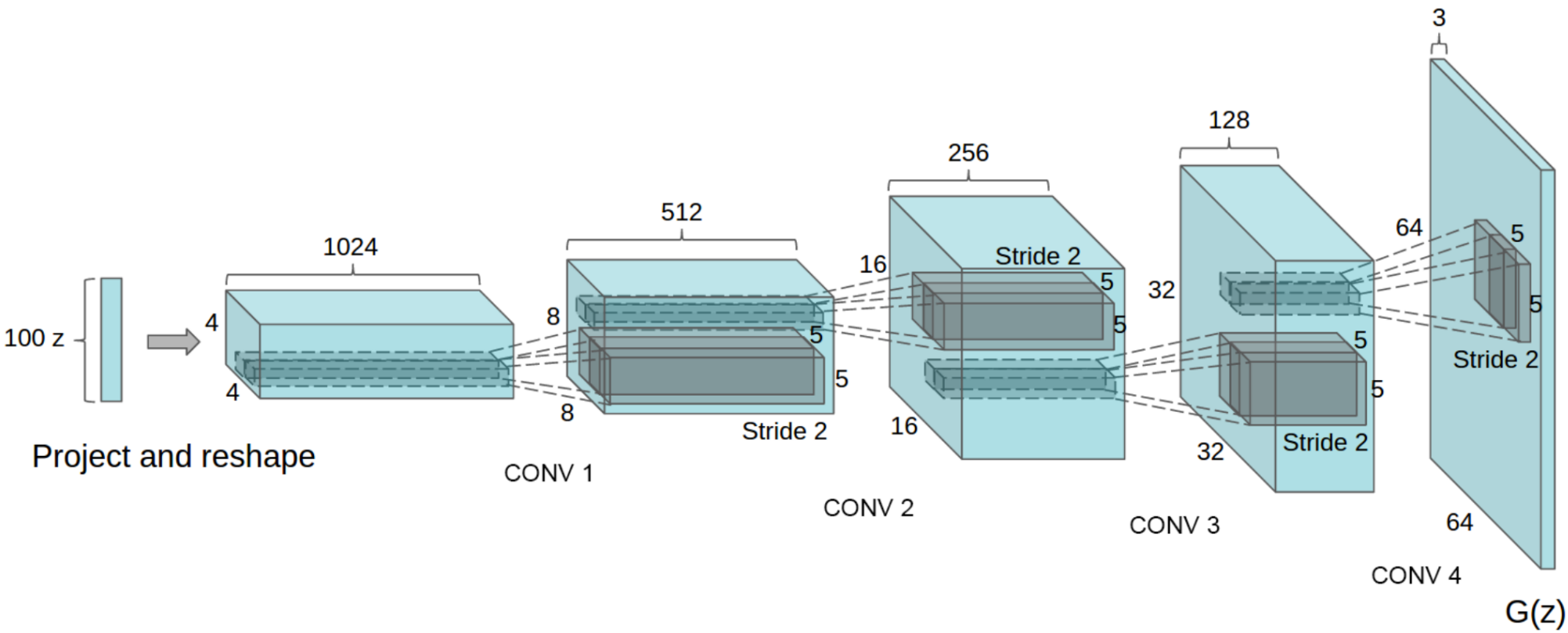
Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

Alec Radford, Luke Metz, Soumith Chintala

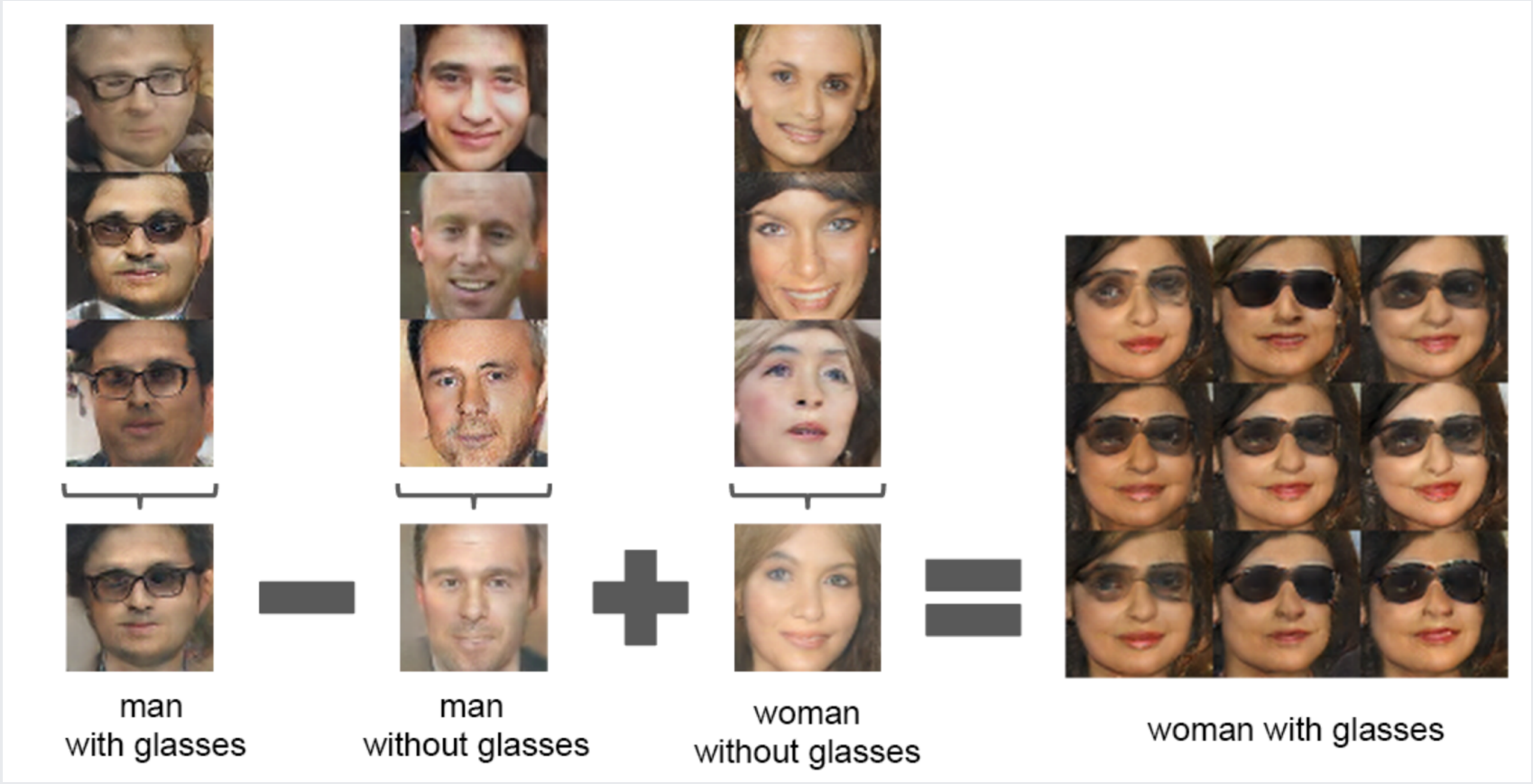
(Submitted on 19 Nov 2015 (v1), last revised 7 Jan 2016 (this version, v2))

In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. In this work we hope to help bridge the gap between the success of CNNs for supervised learning and unsupervised learning. We introduce a class of CNNs called deep convolutional generative adversarial networks (DCGANs), that have certain architectural constraints, and demonstrate that they are a strong candidate for unsupervised learning. Training on various image datasets, we show convincing evidence that our deep convolutional adversarial pair learns a hierarchy of representations from object parts to scenes in both the generator and discriminator. Additionally, we use the learned features for novel tasks - demonstrating their applicability as general image representations.

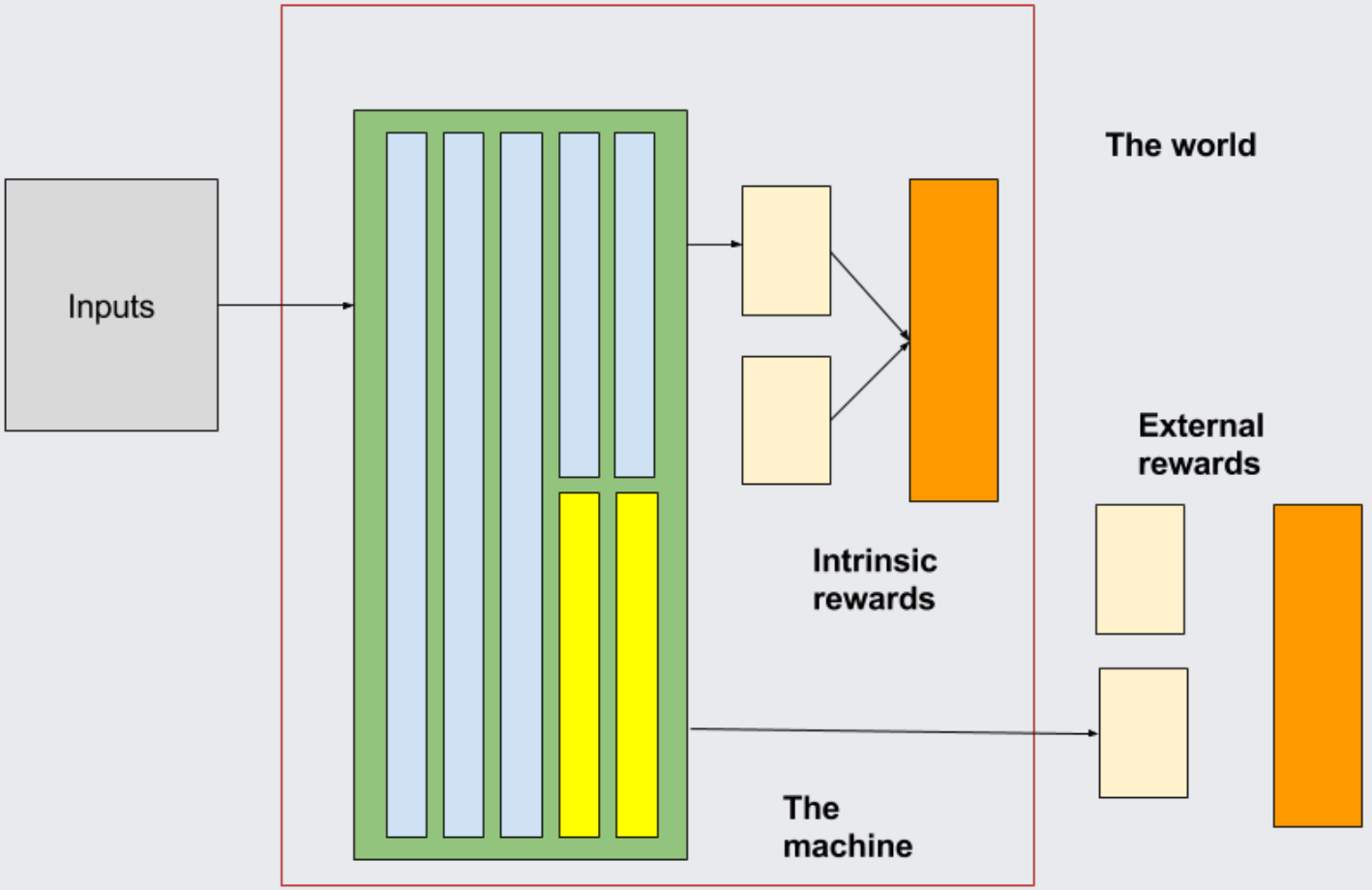
Comments: Under review as a conference paper at ICLR 2016
Subjects: **Learning (cs.LG)**; Computer Vision and Pattern Recognition (cs.CV)
Cite as: **arXiv:1511.06434 [cs.LG]**
(or **arXiv:1511.06434v2 [cs.LG]** for this version)


100 z
Project and reshape
4
4
1024
8
8
512
CONV 1
Stride 2
16
16
256
CONV 2
Stride 2
32
32
128
CONV 3
Stride 2
64
64
3
CONV 4
Stride 2
5
5
5
5
5
5
5
5
G(z)

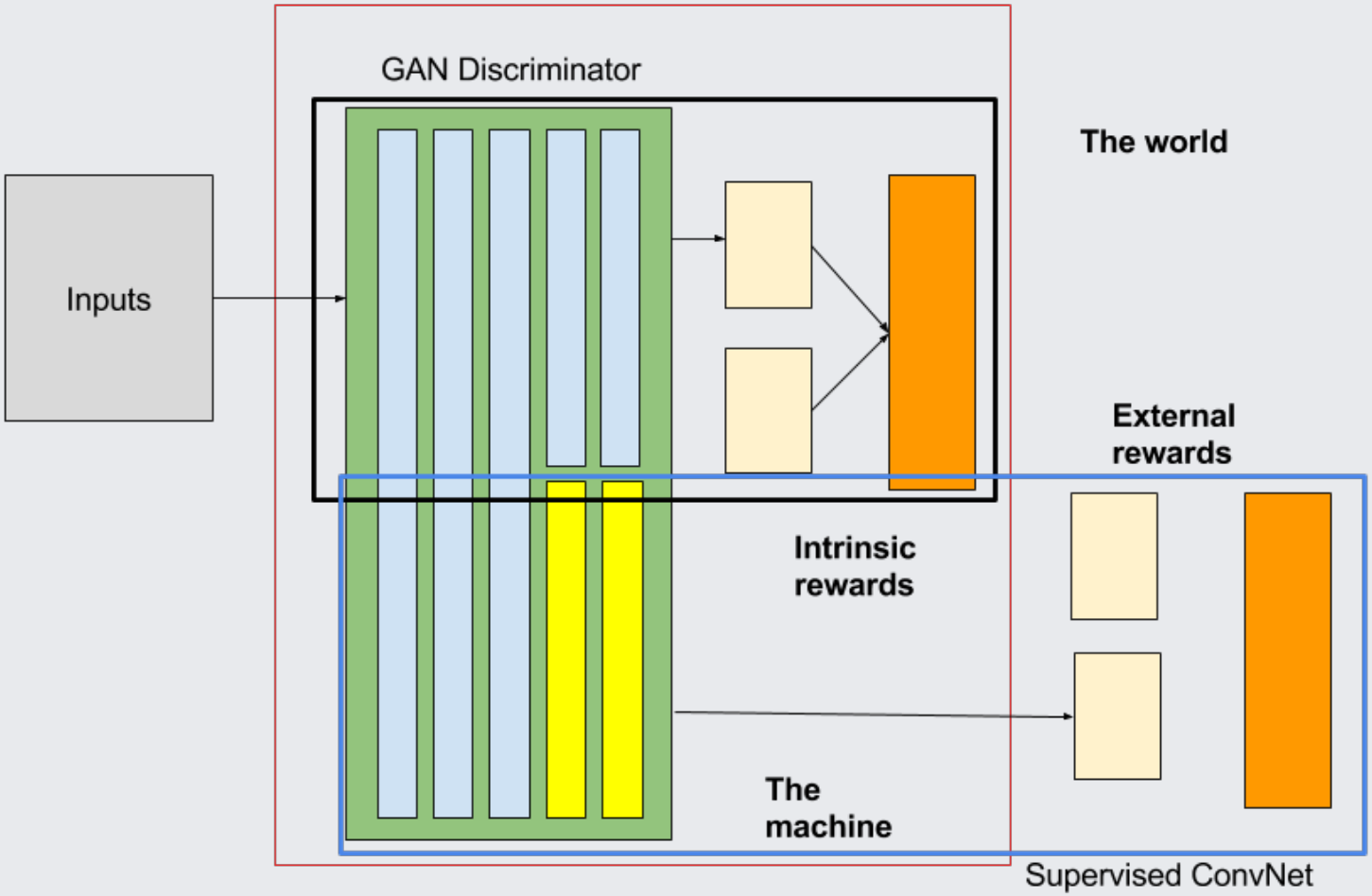
Latent space arithmetic



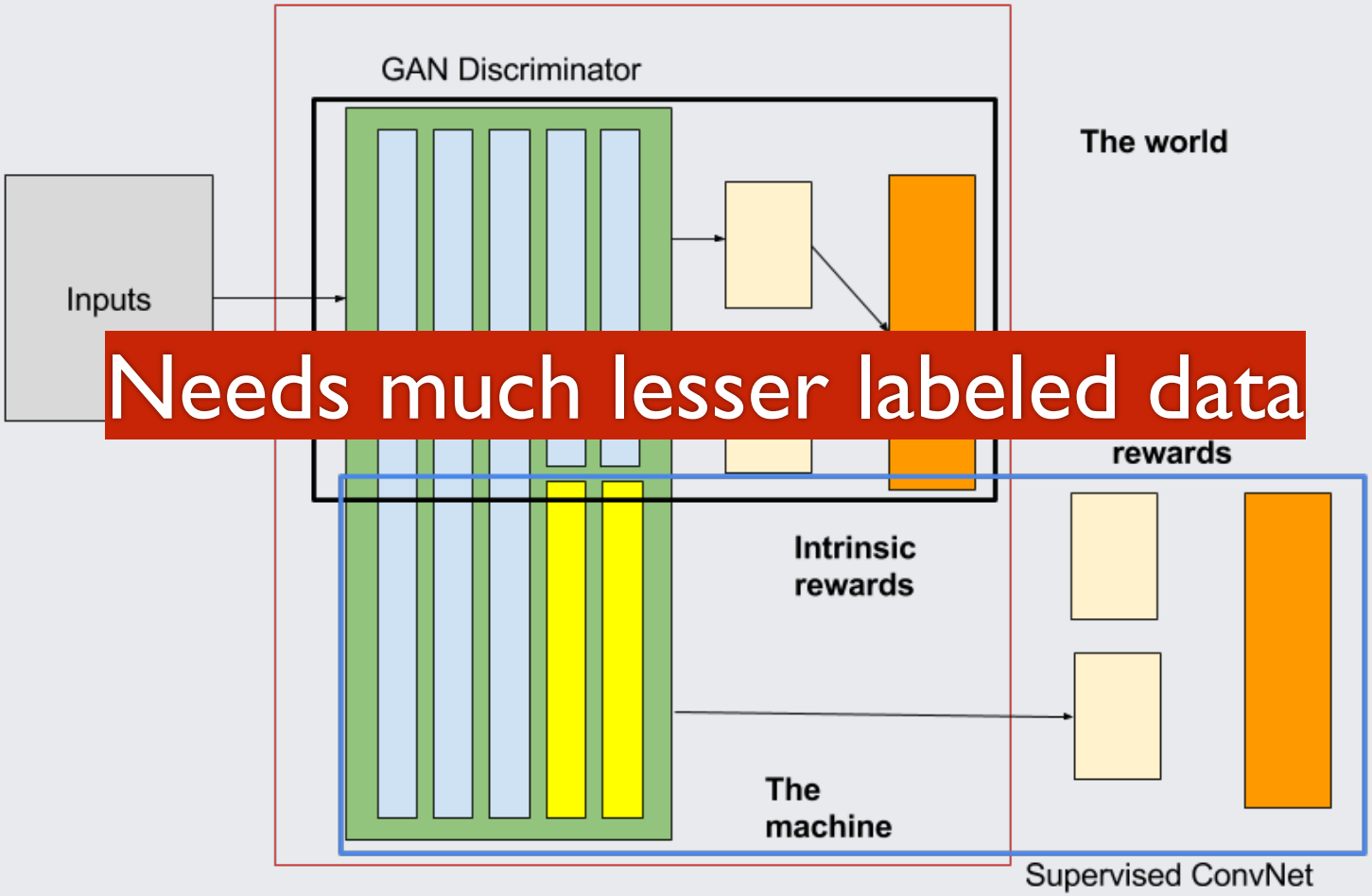
Using the GAN feature representation



Using the GAN feature representation



Using the GAN feature representation



Using the GAN feature representation

Table 2: SVHN classification with 1000 labels

Model	error rate
KNN	77.93%
TSVM	66.55%
M1+KNN	65.63%
M1+TSVM	54.33%
M1+M2	36.02%
SWWAE without dropout	27.83%
SWWAE with dropout	23.56%
DCGAN (ours) + L2-SVM	22.48%
Supervised CNN with the same architecture	28.87% (validation)

In-painting GANs

Context Encoders: Feature Learning by Inpainting

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros

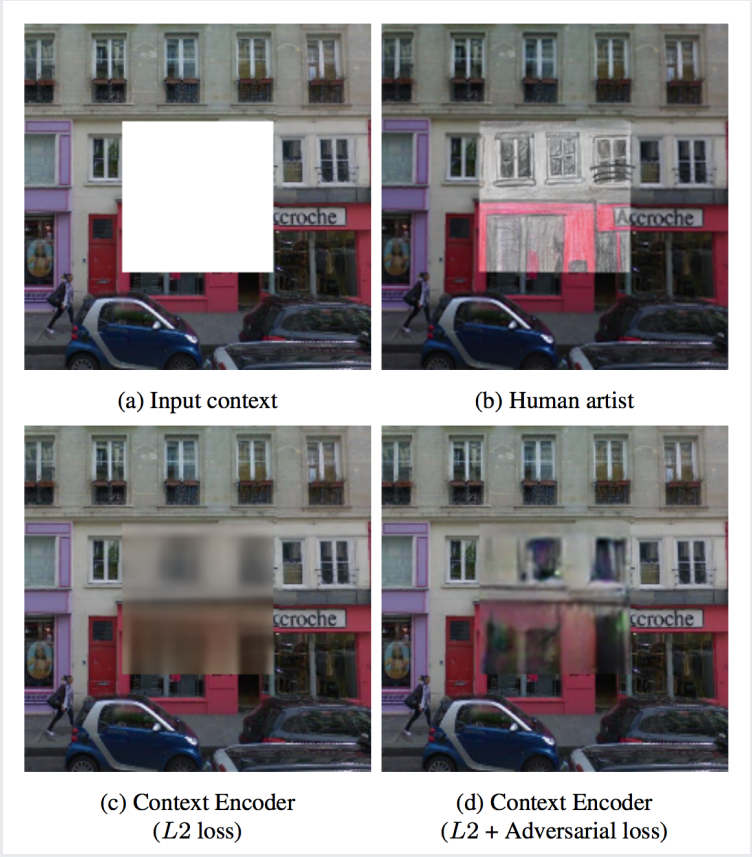
(Submitted on 25 Apr 2016)

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose Context Encoders -- a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.

Comments: CVPR 2016
Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI); Graphics (cs.GR); Learning (cs.LG)
Cite as: **arXiv:1604.07379 [cs.CV]**
(or **arXiv:1604.07379v1 [cs.CV]** for this version)

In-painting GANs

In-painting GANs



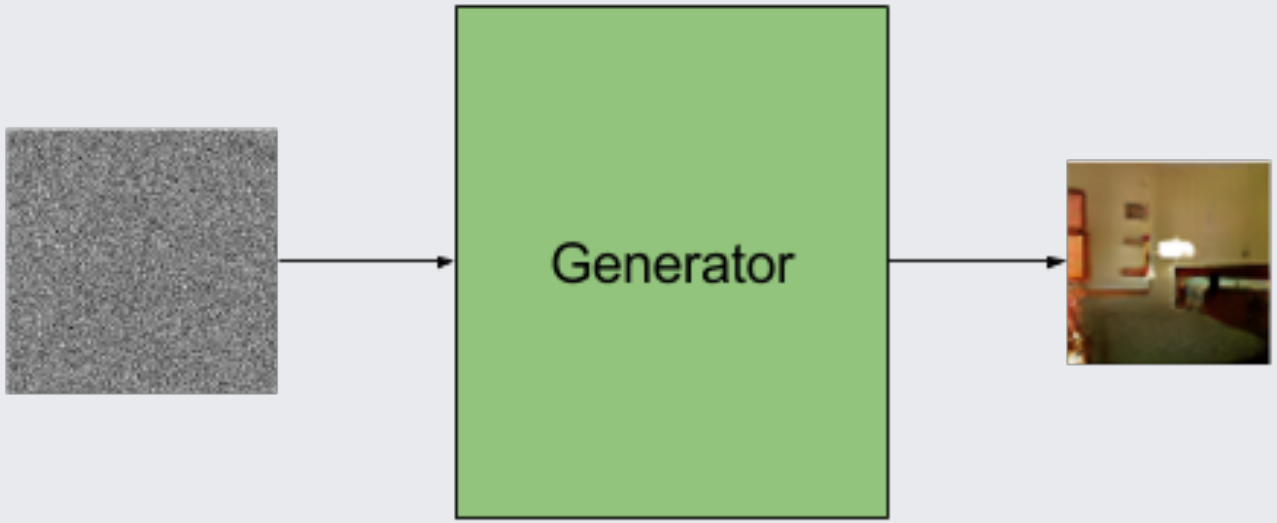
(a) Input context

(b) Human artist

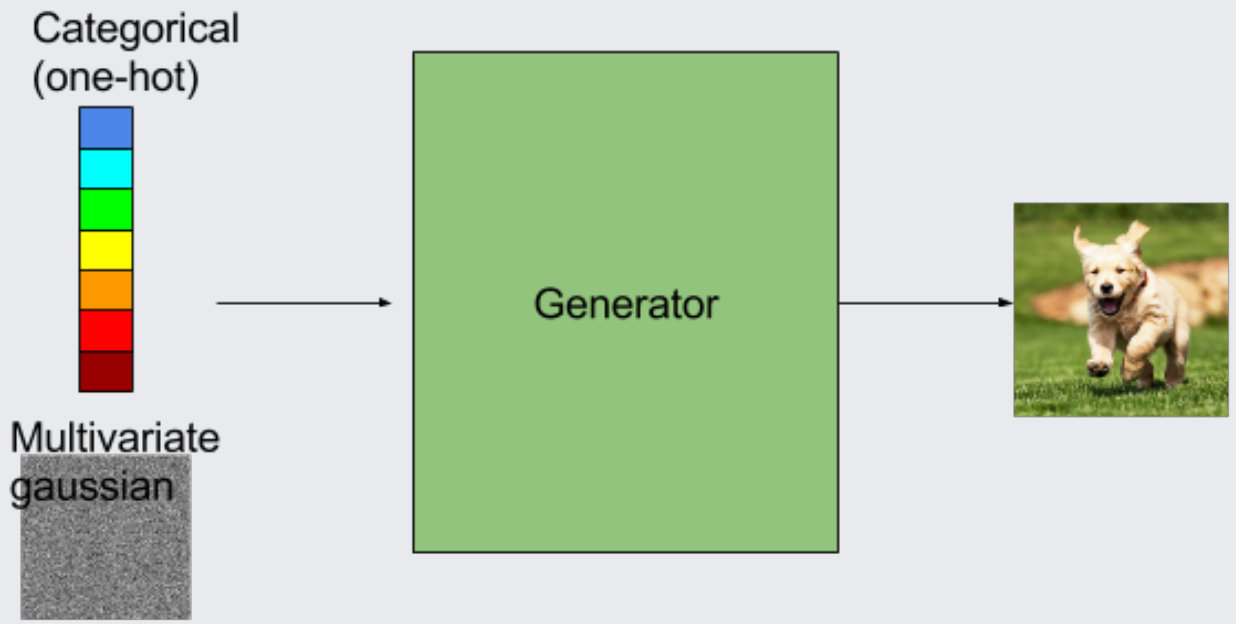
(c) Context Encoder ($L2$ loss)

(d) Context Encoder ($L2$ + Adversarial loss)

Disentangling representations



Disentangling representations



Disentangling representations

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

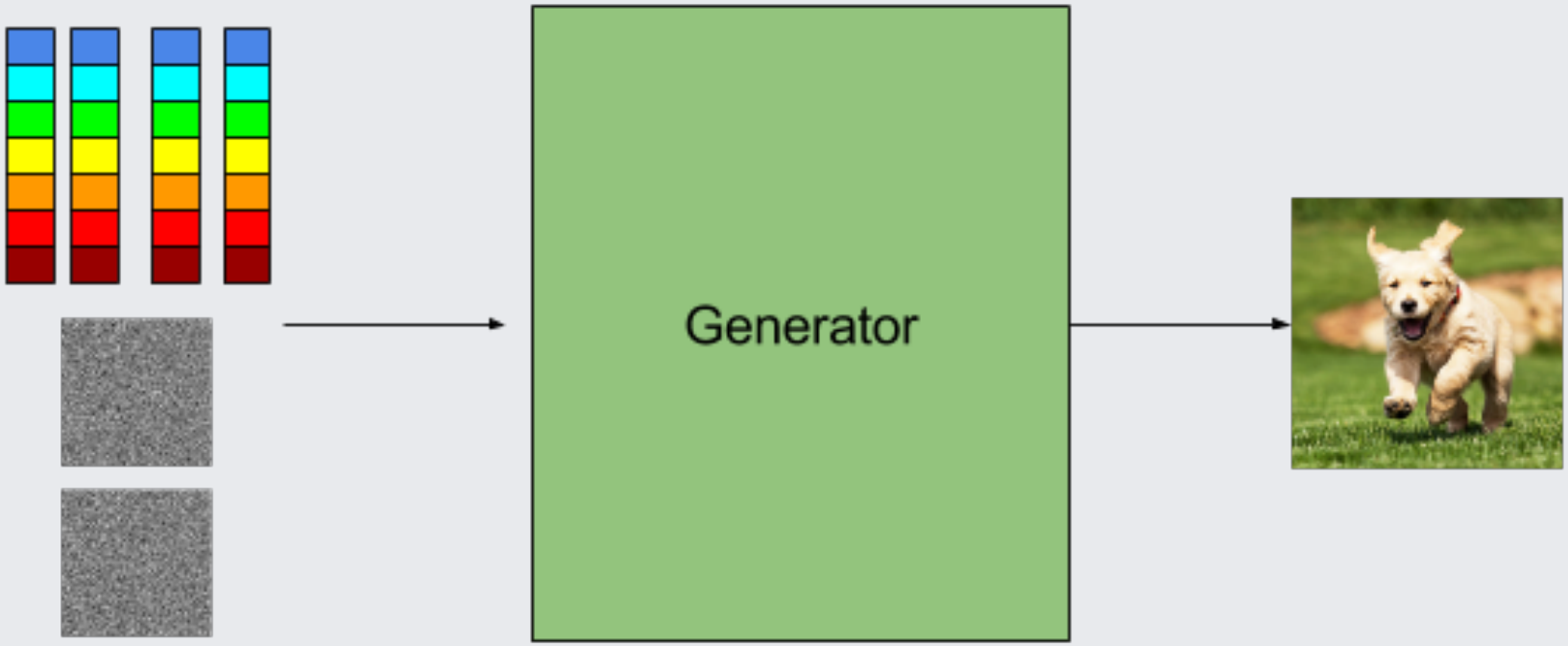
Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, Pieter Abbeel

(Submitted on 12 Jun 2016)

This paper describes InfoGAN, an information-theoretic extension to the Generative Adversarial Network that is able to learn disentangled representations in a completely unsupervised manner. InfoGAN is a generative adversarial network that also maximizes the mutual information between a small subset of the latent variables and the observation. We derive a lower bound to the mutual information objective that can be optimized efficiently, and show that our training procedure can be interpreted as a variation of the Wake-Sleep algorithm. Specifically, InfoGAN successfully disentangles writing styles from digit shapes on the MNIST dataset, pose from lighting of 3D rendered images, and background digits from the central digit on the SVHN dataset. It also discovers visual concepts that include hair styles, presence/absence of eyeglasses, and emotions on the CelebA face dataset. Experiments show that InfoGAN learns interpretable representations that are competitive with representations learned by existing fully supervised methods.

Subjects: **Learning (cs.LG)**; Machine Learning (stat.ML)
Cite as: **arXiv:1606.03657 [cs.LG]**
(or **arXiv:1606.03657v1 [cs.LG]** for this version)

Disentangling representations



Disentangling representations

(a) Varying c_1 on InfoGAN (Digit type)

(b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)

(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

Disentangling representations

(a) Rotation (b) Width

Stability and Representation Reuse

Improved Techniques for Training GANs

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen

(Submitted on 10 Jun 2016)

We present a variety of new architectural features and training procedures that we apply to the generative adversarial networks (GANs) framework. We focus on two applications of GANs: semi-supervised learning, and the generation of images that humans find visually realistic. Unlike most work on generative models, our primary goal is not to train a model that assigns high likelihood to test data, nor do we require the model to be able to learn well without using any labels. Using our new techniques, we achieve state-of-the-art results in semi-supervised classification on MNIST, CIFAR-10 and SVHN. The generated images are of high quality as confirmed by a visual Turing test: our model generates MNIST samples that humans cannot distinguish from real data, and CIFAR-10 samples that yield a human error rate of 21.3%. We also present ImageNet samples with unprecedented resolution and show that our methods enable the model to learn recognizable features of ImageNet classes.

Subjects: **Learning (cs.LG)**; Computer Vision and Pattern Recognition (cs.CV); Neural and Evolutionary Computing (cs.NE)
Cite as: **arXiv:1606.03498 [cs.LG]**
(or **arXiv:1606.03498v1 [cs.LG]** for this version)

Stability and Representation Reuse

- Feature matching
- Minibatch discrimination
- Label smoothing
- What's next?

Stability and Representation Reuse

Model	Percentage of incorrectly predicted test examples for a given number of labeled samples		
	500	1000	2000
DGN [21]		36.02±0.10	
Virtual Adversarial [22]		24.63	
Auxiliary Deep Generative Model [23]		22.86	
Skip Deep Generative Model [23]		16.61±0.24	
Our model	18.44 ± 4.8	8.11 ± 1.3	6.16 ± 0.58
Ensemble of 10 of our models		5.88 ± 1.0	

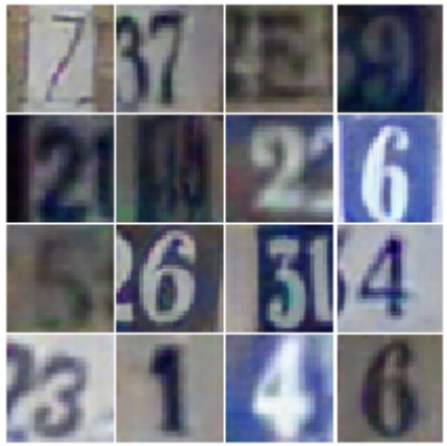


Figure 5: *(Left)* Error rate on SVHN. *(Right)* Samples from the generator for SVHN.

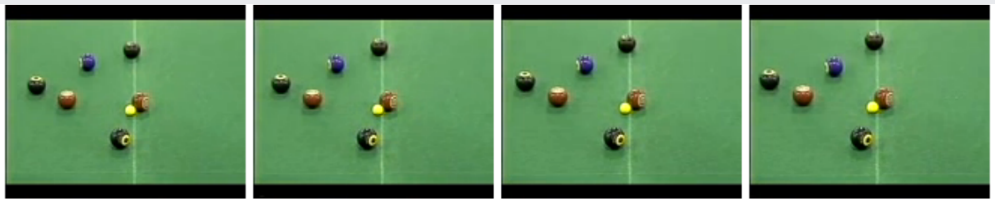
Stability and Representation Reuse

Model	Test error rate for a given number of labeled samples			
	1000	2000	4000	8000
Ladder network [24]			20.40 ± 0.47	
CatGAN [14]			19.58 ± 0.46	
Our model	21.83 ± 2.01	19.61 ± 2.09	18.63 ± 2.32	17.72 ± 1.82
Ensemble of 10 of our models	19.22 ± 0.54	17.25 ± 0.66	15.59 ± 0.47	14.87 ± 0.89

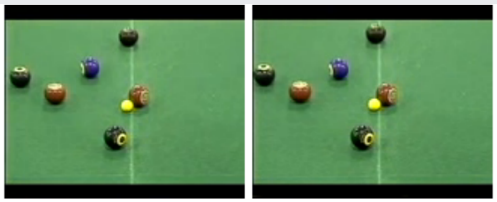
Table 2: Test error on semi-supervised CIFAR-10. Results are averaged over 10 splits of data.

What's next?

- Planning and forward modeling

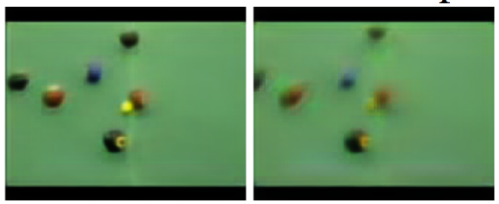


Input frames

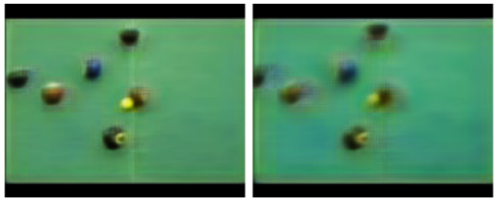


Ground truth

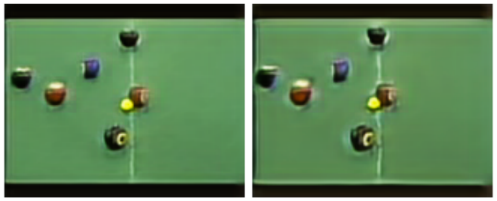
ℓ_2 result



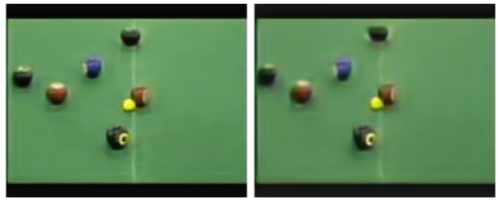
ℓ_1 result



GDL ℓ_1 result



Adversarial result



Adversarial+GDL result

Questions

- When will adversarial networks take over the world?
  - Soon.