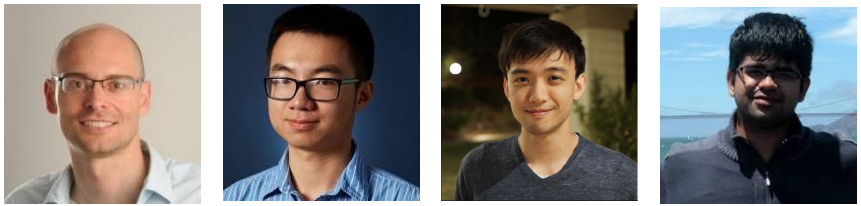
CS294-158 Deep Unsupervised Learning

Lecture 1b: Motivation



Pieter Abbeel, Peter Chen, Jonathan Ho, Aravind Srinivas

UC Berkeley

What is Unsupervised Learning?

- Capturing rich patterns in raw data with deep networks in a **label-free** way
  - Generative Models: recreate raw data distribution
  - Self-supervised Learning: “puzzle” tasks that require semantic understanding
- But why do we care?

"The brain has about 10^14 synapses and we only live for about 10^9 seconds. So we have a lot more parameters than data. This motivates the idea that we must do a lot of unsupervised learning since the perceptual input (including proprioception) is the only place we can get 10^5 dimensions of constraint per second."

- Geoffrey Hinton (in his 2014 AMA on Reddit)

Need tremendous amount of information to build machines that have common sense and generalize

- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- Unsupervised/Predictive Learning (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

[LeCun-20161205-NeurIPS-keynote]

Ideal Intelligence is all about compression (finding all patterns)

- Finding all patterns = short description of raw data (low Kolmogorov Complexity)
- Shortest code-length = optimal inference (Solomonoff Induction)
- Extensible to optimal action making agents (AIXI)

Aside from theoretical interests

- Deep Unsupervised Learning has many powerful applications
  - Generate novel data
  - Compression
  - Improve downstream tasks
  - Flexible building blocks

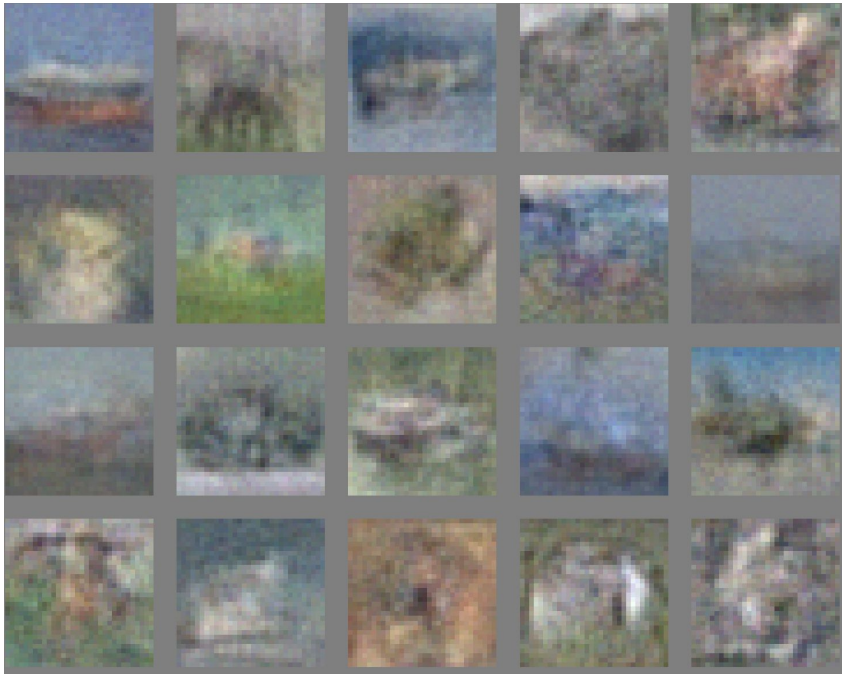
Generate Images

[Deep Belief Nets, Hinton, Osindero, Teh, 2006]

Generate Images

[VAE, Kingma and Welling, 2013]

Generate Images



[GAN, Goodfellow et al. 2014]

Generate Images

[DCGAN, Radford, Metz, Chintala 2015]

Generate Images

[DCGAN, Radford, Metz, Chintala 2015]

Generate Images

bicubic
(21.59dB/0.6423)

SRResNet
(23.53dB/0.7832)

SRGAN
(21.15dB/0.6868)

original

[Ledig, Theis, Huszar et al, 2017]

Generate Images

[CycleGAN: Zhu, Park, Isola & Efros, 2017]

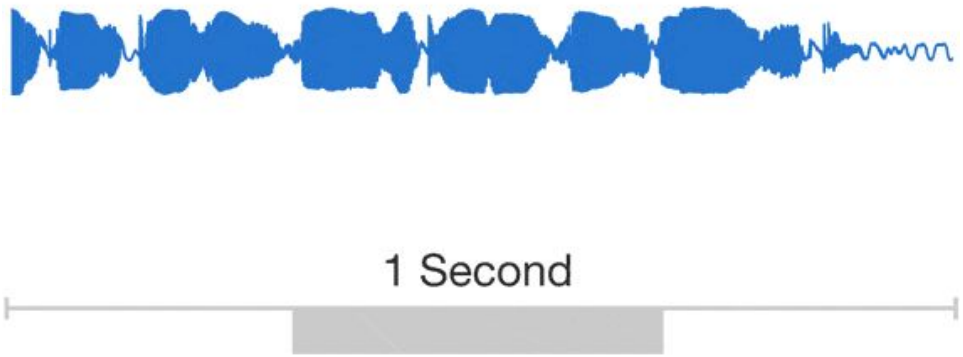
Generate Images

[BigGAN, Brock, Donahue, Simonyan, 2018]

Generate Images

[StyleGAN, Karras, Laine, Aila, 2018]

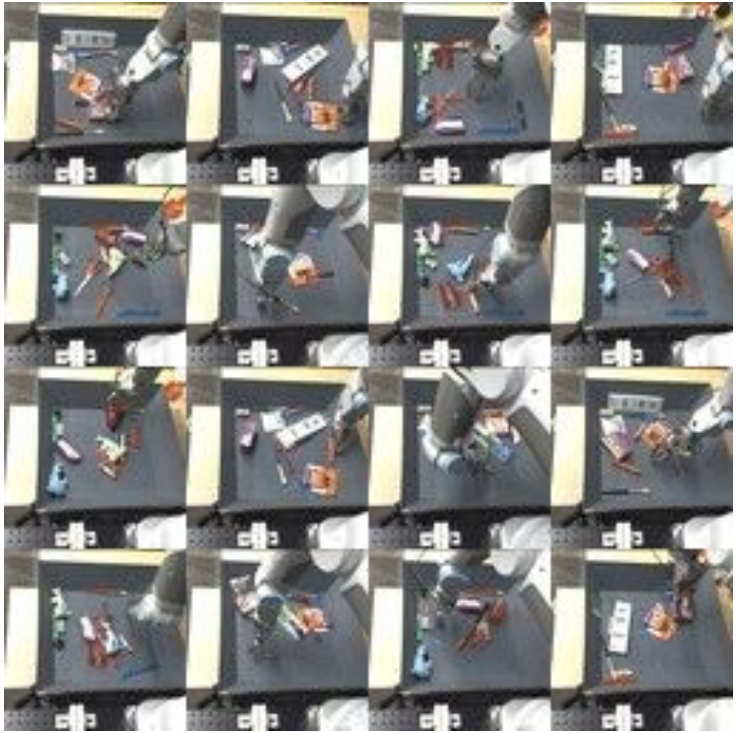
Generate Audio



Parametric

WaveNet

Generate Video



[Video Pixel Networks, Kalchbrenner, van den Oord, Simonyan, Danihelka, Vinyals, Graves, Kavukcuoglu, 2016]

Generate Text

PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

[Char-rnn, 2015]

Generate Text

```
\begin{proof}
We may assume that $\mathcal{I}$ is an abelian sheaf on
$\mathcal{C}$.
\item Given a morphism $\Delta : \mathcal{F} \to
\mathcal{I}$
is an injective and let $\mathfrak q$ be an abelian sheaf on
$X$.
Let $\mathcal{F}$ be a fibered complex. Let $\mathcal{F}$
be a category.
\begin{enumerate}
\item \hyperref[setain-construction-phantom]{Lemma}
\label{lemma-characterize-quasi-finite}
Let $\mathcal{F}$ be an abelian quasi-coherent sheaf on
$\mathcal{C}$.
Let $\mathcal{F}$ be a coherent $\mathcal{O}_X$-module.
Then
$\mathcal{F}$ is an abelian catenary over $\mathcal{C}$.
\item The following are equivalent
\begin{enumerate}
\item $\mathcal{F}$ is an $\mathcal{O}_X$-module.
\end{lemma}
```

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the compariconly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M^\bullet} = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example **??**. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma **??**. Namely, by Lemma **??** we see that R is geometrically regular over S.

Compression - Lossless

Method	CIFAR-10 (bits / dim)
PNG (1996)	5.92
WebP (2010)	4.61
NICE (Dinh et al., 2014)	4.48
Conv DRAW (Gregor et al., 2016)	3.5
Real NVP (Dinh et al., 2016)	3.49
VAE with IAF (Kingma et al., 2016)	3.11
PixelRNN (Oord et al., 2016b)	3.00
Gated PixelCNN (van den Oord et al., 2016b)	3.03
Image Transformer (Anonymous, 2018)	2.98
PixelCNN++ (Salimans et al., 2017)	2.92
Block Sparse PixelCNN++ (OpenAI, 2017)	2.90
PixelSNAIL (2018)	**2.85**

100K LOC, years of engineering

Days of training on GPU

No pattern hardcoded

Compression - Lossy

JPEG

JPEG2000

WaveOne

[Rippel & Bourdev, 2017]

Downstream Task - Sentiment Detection

This is one of Crichton's best books. The characters of Karen Ross, Peter Elliot, Munro, and Amy are beautifully developed and their interactions are exciting, complex, and fast-paced throughout this impressive novel. And about 99.8 percent of that got lost in the film. Seriously, the screenplay AND the directing were horrendous and clearly done by people who could not fathom what was good about the novel. I can't fault the actors because frankly, they never had a chance to make this turkey live up to Crichton's original work. I know good novels, especially those with a science fiction edge, are hard to bring to the screen in a way that lives up to the original. But this may be the absolute worst disparity in quality between novel and screen adaptation ever. The book is really, really good. The movie is just dreadful.

[Radford et al., 2017]

Downstream Task - NLP benchmark

Rank	Name	Model	URL	Score
1	bigbird he	Microsoft D365 AI & MSR AI		81.9
2	Jacob Devlin	BERT: 24-layers, 1024-hidden, 16-heads		80.4
3	Jason Phang	GPT on STILTs		76.9
4	Alec Radford	Singletask Pretrain Transformer		72.8
5	Samuel Bowman	BiLSTM+ELMo+Attn		70.5
6	GLUE Baselines	BiLSTM+ELMo+Attn		68.9
		GenSen		66.6
		Single Task BiLSTM+ELMo		66.2
		BiLSTM+Attn		65.7
		BiLSTM+ELMo		64.9
		Single Task BiLSTM+ELMo+Attn		64.8
		InferSent		64.7
		BiLSTM+CoVe+Attn		64.3
		BiLSTM		63.5

- SOTA unsupervised pre-training + transformer
- Unsupervised pre-training + BiLSTM + Attn
- No pre-training + BiLSTM + Attn

[https://gluebenchmark.com/leaderboard]

Downstream Task - Vision Task

Method	Ref	Class.	Det.	Segm.
Supervised [20]	[43]	79.9	56.8	48.0
Random	[33]	53.3	43.4	19.8
Context [9]	[19]	55.3	46.6	-
Context [9]*	[19]	65.3	51.1	-
Jigsaw [30]	[30]	67.6	**53.2**	37.6
ego-motion [1]	[1]	52.9	41.8	-
ego-motion [1]*	[1]	54.2	43.9	-
Adversarial [10]*	[10]	58.6	46.2	34.9
ContextEncoder [33]	[33]	56.5	44.5	29.7
Sound [31]	[44]	54.4	44.0	-
Sound [31]*	[44]	61.3	-	-
Video [41]	[19]	62.8	47.4	-
Video [41]*	[19]	63.1	47.2	-
Colorization [43]*	[43]	65.9	46.9	35.6
Split-Brain [44]*	[44]	67.1	46.7	36.0
ColorProxy [22]	[22]	65.9	-	**38.0**
WatchingObjectsMove [32]	[32]	61.0	52.2	-
Counting		**67.7**	51.4	36.6

- Various unsupervised pre-training improved on randomly initialized representation;
- There is still a sizable gap to fully supervised version