Learning Representations with Deep InfoMax

Karan Grewal
University of Toronto & Princeton University

learning unsupervised representations

- supervised annotation is very costly

- uses:
  - extract useful information for downstream tasks

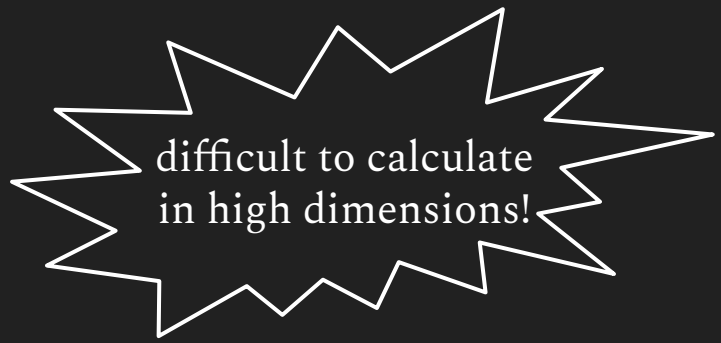
- good representations:
  - high correlation between data & representation
  - captures signal, ignores noise

mutual information

a measure of how informative one variable is of the other

$$I(\mathcal{X};\mathcal{Z}) = \mathbb{E}_{\mathbb{P}_{\mathcal{X}\mathcal{Z}}}\left[\log \frac{\mathbb{P}_{\mathcal{X}\mathcal{Z}}}{\mathbb{P}_{\mathcal{X}} \otimes \mathbb{P}_{\mathcal{Z}}}\right]$$

notice that if $\mathbb{P}_{\mathcal{X}\mathcal{Z}} = \mathbb{P}_{\mathcal{X}} \otimes \mathbb{P}_{\mathcal{Z}}$ then $I(\mathcal{X};\mathcal{Z}) = 0$



mutual information

1. contrastive loss lower bound (infoNCE)

$$I(\mathcal{X};\mathcal{Z}) \geq \log(N) \quad + \quad \underbrace{\mathbb{E}_S\left[\log \frac{h(x^*,z^*)}{\sum_j h(x_j,z_j)}\right]}_{-\mathcal{L}_{\text{contrast}}}$$

where $h(\cdot,\cdot)$ estimates the ratio $\frac{\mathbb{P}_{\mathcal{X}\mathcal{Z}}}{\mathbb{P}_{\mathcal{X}}\otimes\mathbb{P}_{\mathcal{Z}}}$,

$$S = \{(x^*, z^*)\} \quad \cup \quad \{(x_j, z_j)\}_{j=1}^{N}$$

from $\mathbb{P}_{\mathcal{X}\mathcal{Z}}$ — from $\mathbb{P}_{\mathcal{X}} \otimes \mathbb{P}_{\mathcal{Z}}$

van den Oord et al., 2018

mutual information

2. Donsker-Varadhan representation of KL divergence

$$
\begin{aligned}
I(\mathcal{X};\mathcal{Z}) &= \mathbb{E}_{\mathbb{P}_{\mathcal{X}\mathcal{Z}}}\left[\log \frac{\mathbb{P}_{\mathcal{X}\mathcal{Z}}}{\mathbb{P}_{\mathcal{X}} \otimes \mathbb{P}_{\mathcal{Z}}}\right] \\
&= D_{\mathrm{KL}}(\mathbb{P}_{\mathcal{X}\mathcal{Z}} || \mathbb{P}_{\mathcal{X}} \otimes \mathbb{P}_{\mathcal{Z}}) \\
&\geq \sup_{T_\omega : \Omega \to \mathbb{R}} \mathbb{E}_{\mathbb{P}_{\mathcal{X}\mathcal{Z}}}\left[T_\omega\right] - \log \mathbb{E}_{\mathbb{P}_{\mathcal{X}} \otimes \mathbb{P}_{\mathcal{Z}}}\left[e^{T_\omega}\right]
\end{aligned}
$$

Donsker & Varadhan, 1983

framework

image $x \in \mathcal{X}$

image $x' \in \mathcal{X}$

copy

$z \in \mathcal{Z}$

enc_ψ

sample from $\mathbb{P}_{xz}$

sample from $\mathbb{P}_x \otimes \mathbb{P}_z$

T_ω

$$\max_{\psi,\omega} \quad \mathbb{E}_{\mathbb{P}_{xz}}\left[T_\omega\right] - \log \mathbb{E}_{\mathbb{P}_x \times \tilde{\mathbb{P}}_x}\left[e^{T_\omega}\right]$$

algorithm

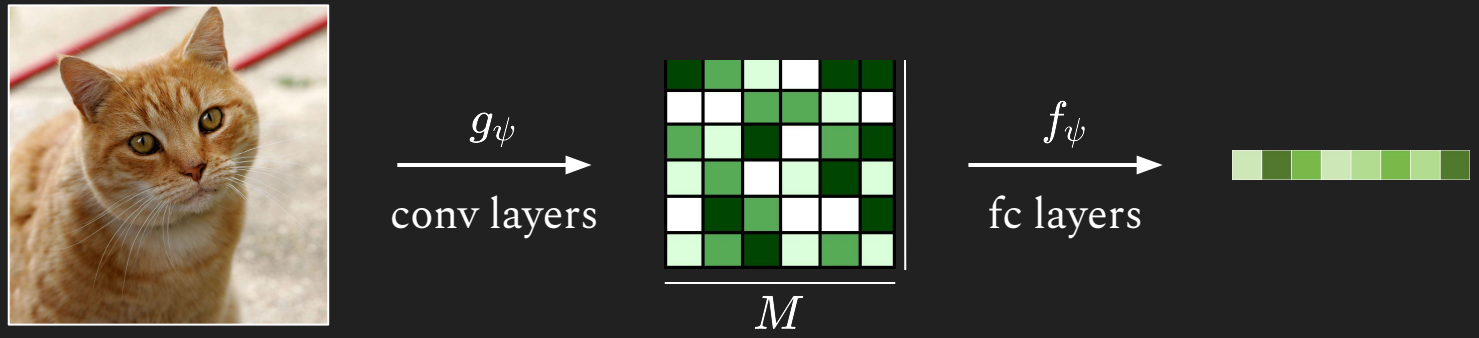
1. sample (+) examples $x_+^{(1)}, \ldots, x_+^{(n)} \sim \mathbb{P}_X$
2. compute representations $z^{(i)} = \mathsf{enc}_\psi(x_+^{(i)}) \quad \forall i$
3. let $\{(x_+^{(i)}, z^{(i)})\}_i$ be the (+) pairs
4. sample (-) examples $x_-^{(1)}, \ldots, x_-^{(n)} \sim \mathbb{P}_X$
5. let $\{(x_-^{(i)}, z^{(i)})\}_i$ be the (-) pairs
6. maximize $\frac{1}{n}\sum_{i=1}^n T_\omega(\underbrace{x_+^{(i)}, z^{(i)}}_{\text{(+) pairs}}) - \log \frac{1}{n}\sum_{i=1}^n e^{\underbrace{T_\omega(x_-^{(i)}, z^{(i)})}_{\text{(-) pairs}}}$

engineering footnotes

- encoder is a composition of convolutions followed by fully-connected layers

$$\mathsf{enc}_\psi = f_\psi \circ g_\psi$$

- first apply g_ψ to obtain an $M \times M$ feature map, then apply f_ψ to get representation

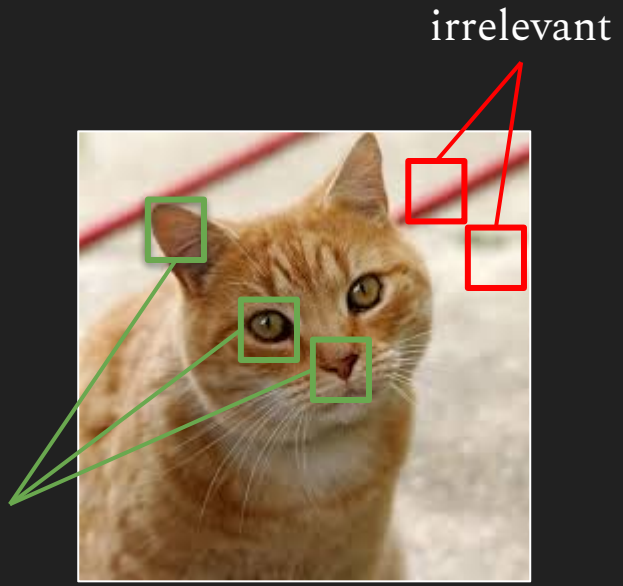


- maximize MI between feature map and representation

local InfoMax

- hypothesis: encoder is more likely to capture information shared across all patches

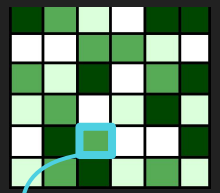
    - pro → global structure will be present
    - con → model has no incentive to focus on relevant information, pixel-level noise will be encoded



local InfoMax

g_ψ

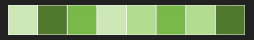
$M \times M$ patches



f_ψ

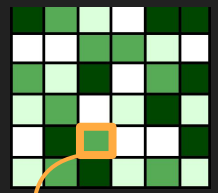
$z \in \mathcal{Z}$

copy



g_ψ

patches of other image

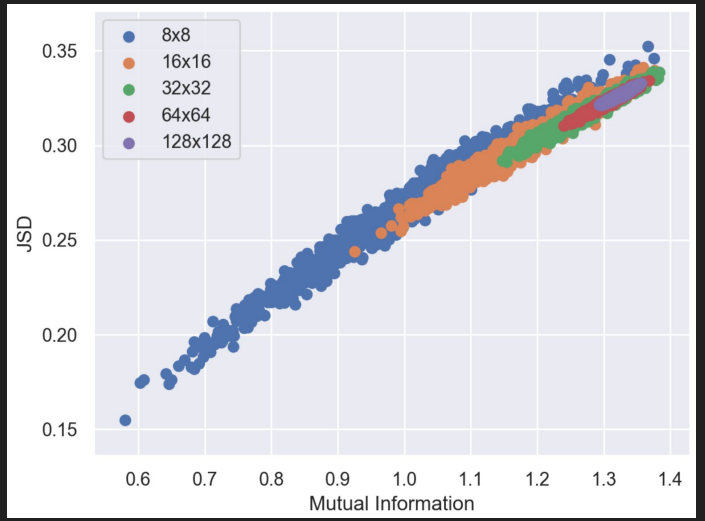


T_ω

$\mathbb{P}_{XZ}$ or $\mathbb{P}_X \otimes \mathbb{P}_Z$?

alternate MI objectives

objective	pros	cons
Donsker-Varadhan (DV)	tightest available bound on KL divergence	requires many (-) samples
Jensen-Shannon divergence (JSD)	stable; few (-) samples needed	not the tightest bound
noise contrastive estimation (NCE)	strongest results	requires many (-) samples



putting it all together

final objective for Deep InfoMax:

$$\max_{\omega,\psi} \alpha \hat{I}(\mathcal{X}, \mathcal{Z})$$

$$+ \max_{\omega,\psi} \frac{\beta}{M^2} \sum_{j=1}^{M^2} \hat{I}(\mathcal{X}_j, \mathcal{Z})$$

$$+ \min_{\psi} \max_{\theta} \gamma D_{\text{JSD}}(\mathbb{Q}_{\text{prior}} || \mathbb{Q}_{\psi})$$

e.g. estimate KL lower bound

global InfoMax

local InfoMax

prior matching

just like adversarial autoencoders

α, β, γ are tunable hyperparameters

results - linear classification

train SVM on learned representations

CIFAR-10

model	conv	fc	z
VAE	53.8	42.1	39.6
adversarial autoencoder	55.2	43.3	37.8
BiGAN	56.4	38.4	44.9
NAT	48.6	42.6	39.6
DIM	**57.6**	**45.6**	18.6
DIM - global only	46.8	28.8	29.1
DIM - local only	**63.3**	**54.1**	**49.6**

conv → last conv layer, fc → 2nd last fc layer, z → representation

results - nonlinear classification (1/2)

train shallow neural network on learned representations

	CIFAR-10				CIFAR-100		
model	conv	fc	z		conv	fc	z
fully supervised	75.4				42.3		
VAE	60.7	60.5	54.6		37.2	34.1	24.2
β-VAE	62.4	57.9	55.4		32.3	26.9	29.0
adversarial autoencoder	59.4	57.2	52.8		36.2	33.4	23.3
BiGAN	62.6	62.7	52.5		37.6	33.3	21.5
DIM - global	52.2	52.8	43.2		27.7	24.4	20.0
DIM - local (DV)	**72.7**	**70.6**	**64.7**		**48.5**	**44.4**	**39.3**
DIM - local (JSD)	**73.3**	**73.6**	**67.0**		**48.1**	**45.9**	**39.6**
DIM - local (NCE)	**75.2**	**75.6**	**69.1**		**49.7**	**47.7**	**41.6**

conv → last conv layer, fc → 2nd last fc layer, z → representation

results - nonlinear classification (2/2)

train shallow neural network on learned representations

	Tiny ImageNet				STL-10		
model	conv	fc	Y		conv	fc	z
fully supervised	75.4				42.3		
VAE	18.6	16.9	11.9		58.3	56.7	46.5
β-VAE	19.3	16.8	12.4		57.2	55.1	46.9
adversarial autoencoder	18.0	17.3	11.5		59.5	54.5	43.9
BiGAN	24.4	20.2	13.1		**71.5**	67.2	58.5
DIM - global	11.3	6.3	5.0		42.0	30.8	28.1
DIM - local (DV)	**30.4**	**29.5**	**28.2**		**69.2**	**63.8**	**61.9**
DIM - local (JSD)	**33.5**	**36.9**	**31.7**		**72.9**	**70.9**	**65.9**
DIM - local (NCE)	**34.2**	**38.1**	**33.3**		**72.6**	**70.0**	**67.1**

conv → last conv layer, fc → 2nd last fc layer, z → representation

results - MI neural estimate

train a neural network to estimate MI between input and representations

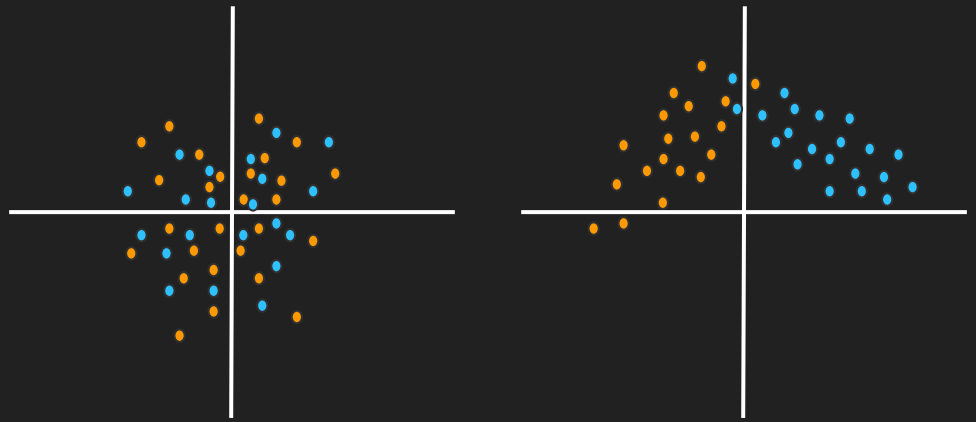
CIFAR-10

model	MINE estimate
VAE	93.0
adversarial autoencoder	87.5
BiGAN	37.7
NAT	6.0
DIM	**101.7**
DIM - global only	49.6
DIM - local only	45.1

Belghazi et al., 2018

results - neural dependency measure

shuffle representations along batch axis and train discriminator to tell real from fake

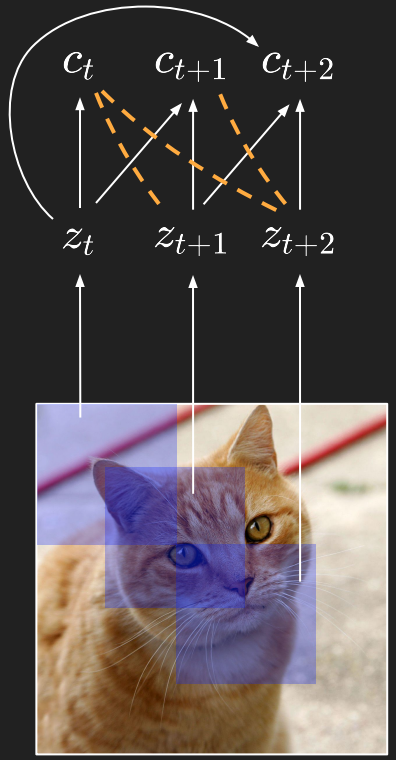


CIFAR-10

model	discriminator loss
VAE	1.6
adversarial autoencoder	0.1
BiGAN	**24.5**
NAT	0.1
DIM	**22.9**
DIM - global only	**10.0**
DIM - local only	**9.2**

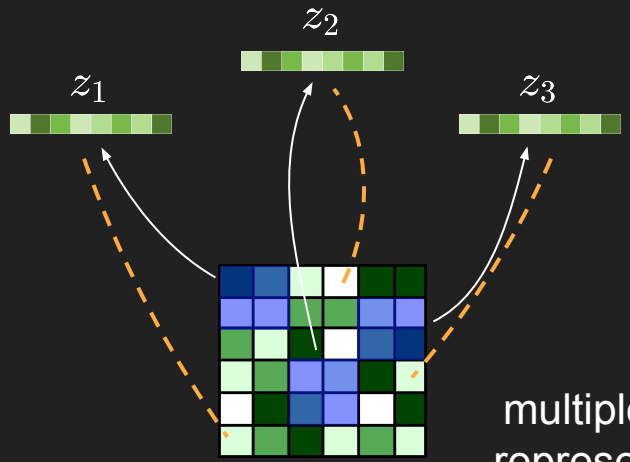
Brakel & Bengio, 2017

results - occlusion



Contrastive Predictive Coding

maximize MI



multiple “local” representations

Deep InfoMax + occlusion

results - occlusion

non-linear classification with occlusion

model	CIFAR-10	STL-10
CPC	77.5	77.8
DIM - original	**81.0**	77.0
DIM - multiple representations	77.5	**78.2**

extensions

(a) maximize MI between different views of the same object/scene

- minimize contrastive loss between representations obtained from 2 separate views / encoders

Tian et al., 2019; Bachman et al., 2019

(b) maximize MI between audio representations from the same speaker

- similar to (a), but the two “views” correspond to audio waveforms taken from the same speaker

Ravanelli & Bengio, 2019

conclusions

pros

- Deep InfoMax doesn’t require a decoder
- MI-based objectives can be extended to other tasks
- local information can be encoded - advantageous for downstream tasks

cons

- still requires 2 discriminators
- may be hard to scale - encoding local information can be harmful

collaborators

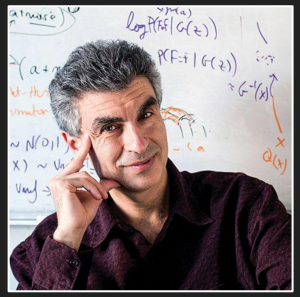
Devon Hjelm
Microsoft Research, Montréal & Mila

Alex Fedorov
Georgia Tech

Samuel Lavoie-Marchildon
Mila

Phil Bachman
Microsoft Research, Montréal

Adam Trischler
Microsoft Research, Montréal



Yoshua Bengio
Mila