Contrastive learning, multi-view redundancy, and linear models

Daniel Hsu
Columbia University

Joint work with:
Akshay Krishnamurthy (*Microsoft Research*)
Christopher Tosh (*Columbia University*)

Representation learning

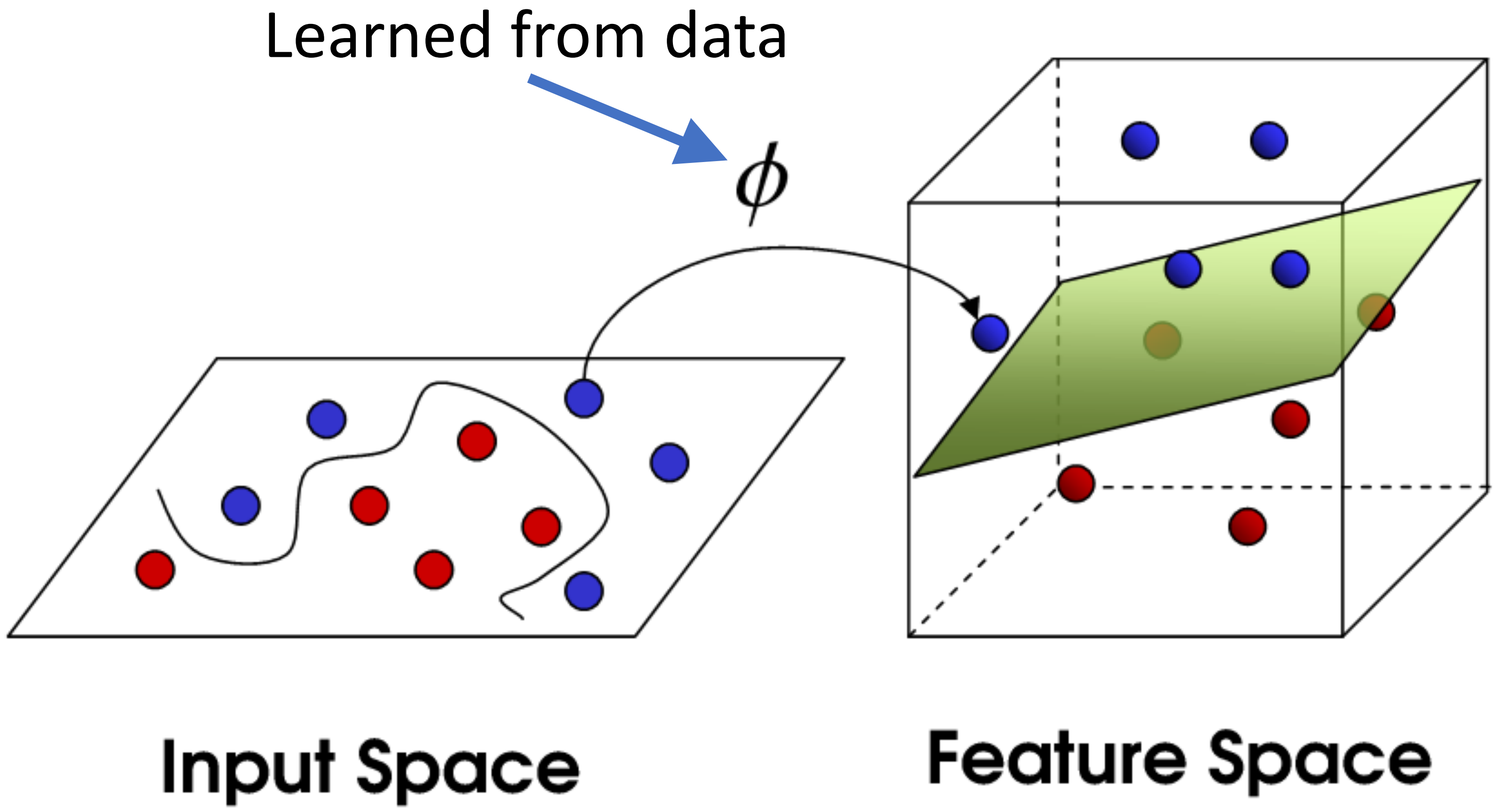
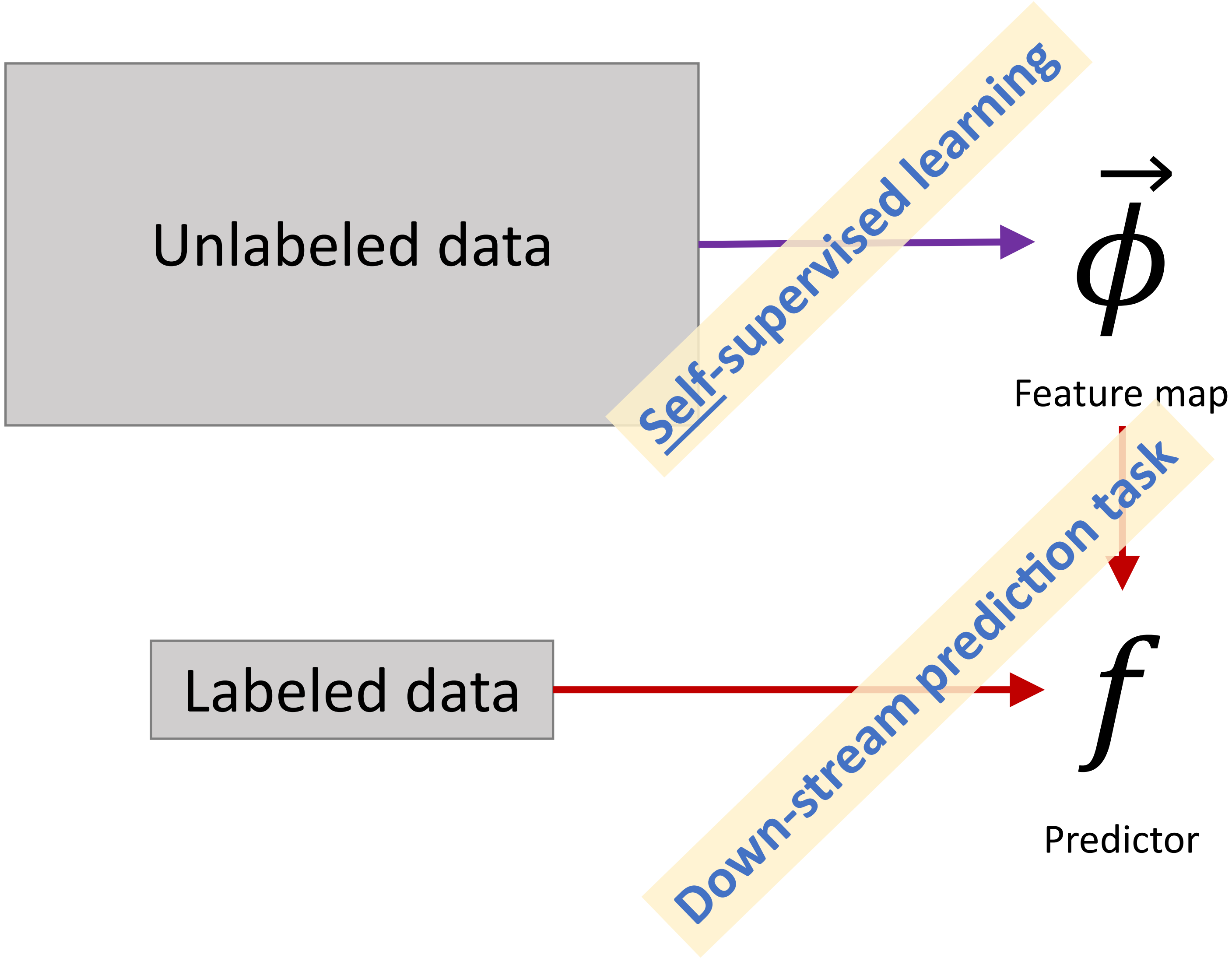


Image credit: towardsdatascience.com

Unsupervised / semi-supervised learning



"Self-supervised learning"

1. Learn to solve artificial prediction problems ("pretext task").
2. Use solution to derive a representation ("feature map") $\vec{\phi}$.

Predict color channel from grayscale channel

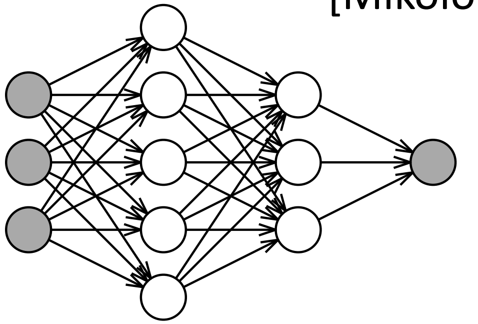
[Zhang, Isola, Efros, 2017]

Predict missing word in a sentence from context

```
The quick brown fox ____ over the lazy dog.

        (a) hops            (c) skips
        (b) jumps           (d) dunks
```

[Mikolov, Sutskever, Chen, Corrado, Dean, 2013]



This talk: Contrastive learning

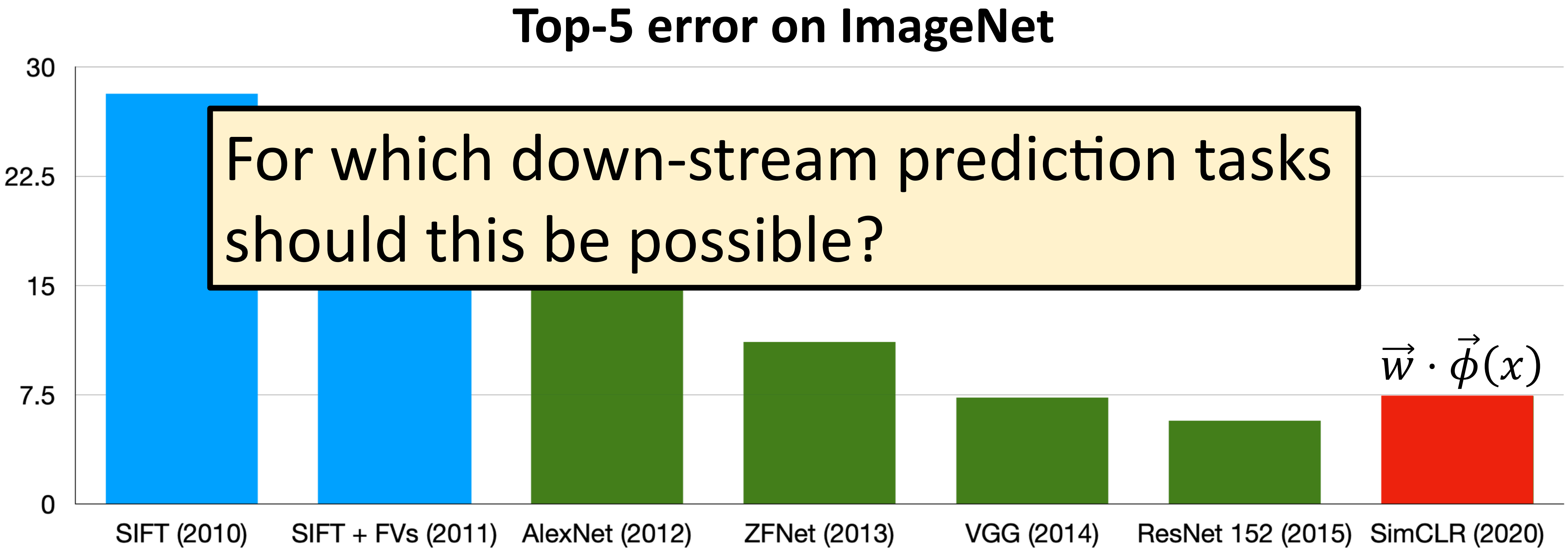
- "Positive" examples: naturally occurring pairs
- "Negative" examples: completely random pairs

Snippets from same article

Snippets from different articles

Contrastive learning appears to work!

Linear models over $\vec{\phi}$, learned with only ~10% of labels, are near SOTA



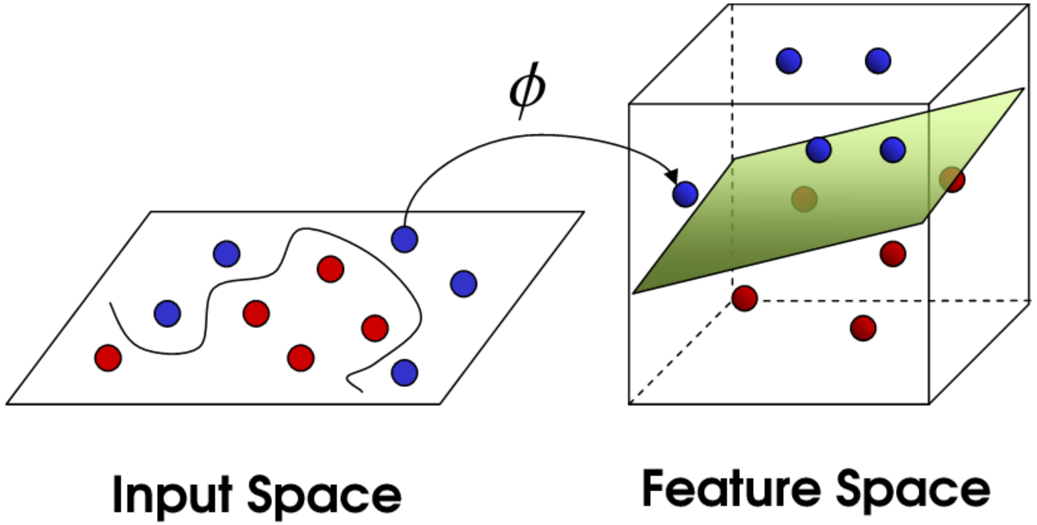
For which down-stream prediction tasks should this be possible?

Our main results (1)

[Contrastive learning is useful when multi-view redundancy holds.]

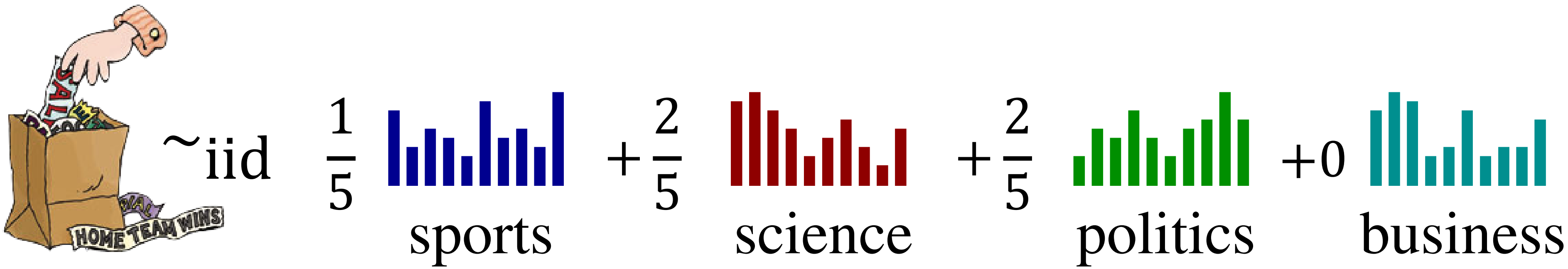
Assume unlabeled data has two views X and Z, each with near-optimal MSE for predicting target Y (possibly via non-linear functions). Then:

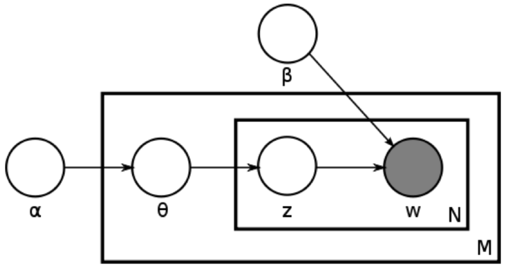
$\exists$ (low-ish dim.) linear function of $\vec{\phi}(X)$ that achieves near-optimal MSE.



Our main results (2)

Assume unlabeled data follow a topic model (e.g., LDA). Then: representation $\vec{\phi}(x)$ = linear transform of topic posterior moments (of order up to document length).





Rest of the talk

1. Contrastive learning & feature map $\vec{\phi}$
2. Multi-view redundancy
3. Interpreting the representation
4. Experimental study

1. Contrastive learning & feature map

Formalizing contrastive learning

[Steinwart, Hush, Scovel, 2005; Abe, Zadrozny, Langford, 2006; Gutmann & Hyvärinen, 2010; Oord, Li, Vinyals, 2018; Arora, Khandeparkar, Khodak, Plevrakis, Saunshi, 2019]

- Learn predictor to discriminate between

$$(x, z) \sim P_{X,Z} \quad \text{[positive example]}$$

and

$$(x, z) \sim P_X \otimes P_Z \quad \text{[negative example]}$$

- Specifically, estimate odds-ratio

$$g^*(x, z) = \frac{\Pr[\text{positive} \mid (x, z)]}{\Pr[\text{negative} \mid (x, z)]}$$

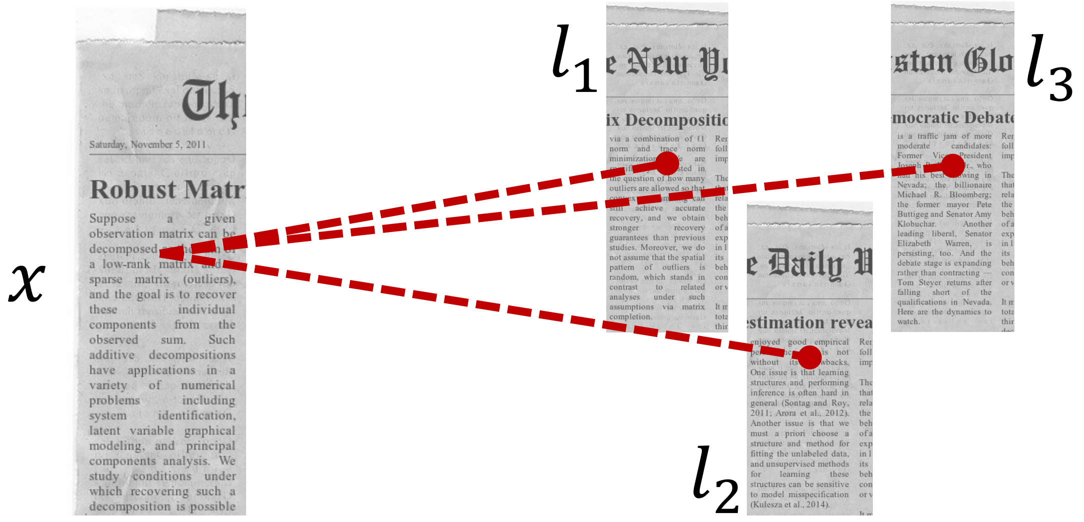
by fitting a model to random positive & negative examples (which are, WLOG, evenly balanced: $0.5\ P_{X,Z} + 0.5\ P_X \otimes P_Z$).

Deriving a representation

- Given an estimate $\hat{g}$ of g^*, construct feature map $\vec{\phi}$:

$$\vec{\phi}(x) \coloneqq (\hat{g}(x, l_i) : i = 1, \ldots, M) \in \mathbb{R}^M$$

where $l_1, \ldots, l_M$ are "landmarks", selected from unlabeled data



2. Multi-view redundancy

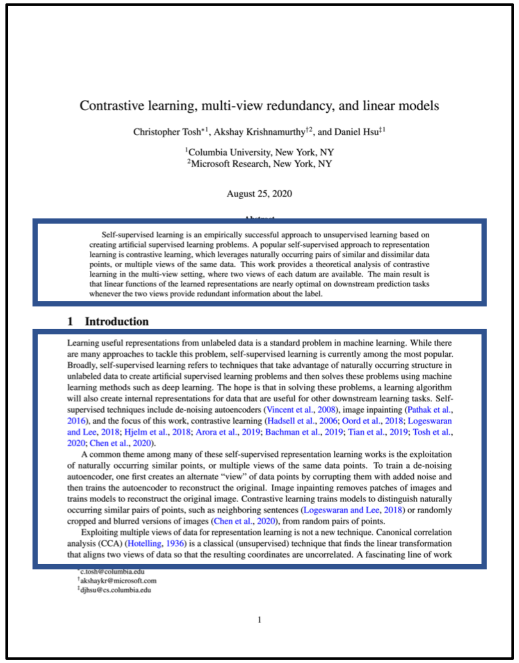
Multi-view data

- Assume (unlabeled) data provides two "views" X and Z, each equally good at predicting a target Y

- **Example**: topic prediction
  - Y = topic of article
  - X = abstract
  - Z = introduction

X

Z



Contrastive learning, multi-view redundancy, and linear models

Christopher Tosh*[1], Akshay Krishnamurthy†[2], and Daniel Hsu‡[1]

[1]Columbia University, New York, NY
[2]Microsoft Research, New York, NY

August 25, 2020

Self-supervised learning is an empirically successful approach to unsupervised learning based on creating artificial supervised learning problems. A popular self-supervised approach to representation learning is contrastive learning, which leverages naturally occurring pairs of similar and dissimilar data points, or multiple views of the same data. This work provides a theoretical analysis of contrastive learning in the multi-view setting, where two views of each datum are available. The main result is that linear functions of the learned representations are nearly optimal on downstream prediction tasks whenever the two views provide redundant information about the label.

1 Introduction

Learning useful representations from unlabeled data is a standard problem in machine learning. While there are many approaches to tackle this problem, self-supervised learning is currently among the most popular. Broadly, self-supervised learning refers to techniques that take advantage of naturally occurring structure in unlabeled data to create artificial supervised learning problems and then solves these problems using machine learning methods such as deep learning. The hope is that in solving these problems, a learning algorithm will also create internal representations for data that are useful for other downstream learning tasks. Self-supervised techniques include de-noising autoencoders (Vincent et al., 2008), image inpainting (Pathak et al., 2016), and the focus of this work, contrastive learning (Hadsell et al., 2006; Oord et al., 2018; Logeswaran and Lee, 2018; Hjelm et al., 2018; Arora et al., 2019; Bachman et al., 2019; Tian et al., 2019; Tosh et al., 2020; Chen et al., 2020).

A common theme among many of these self-supervised representation learning works is the exploitation of naturally occurring similar points, or multiple views of the same data points. To train a de-noising autoencoder, one first creates an alternate "view" of data points by corrupting them with added noise and then trains the autoencoder to reconstruct the original. Image inpainting removes patches of images and trains models to reconstruct the original image. Contrastive learning trains models to distinguish naturally occurring similar pairs of points, such as neighboring sentences (Logeswaran and Lee, 2018) or randomly cropped and blurred versions of images (Chen et al., 2020), from random pairs of points.

Exploiting multiple views of data for representation learning is not a new technique. Canonical correlation analysis (CCA) (Hotelling, 1936) is a classical (unsupervised) technique that finds the linear transformation that aligns two views of data so that the resulting coordinates are uncorrelated. A fascinating line of work

*c.tosh@columbia.edu
†akshaykr@microsoft.com
‡djhsu@cs.columbia.edu

1

Multi-view learning methods

- **Co-training** [Blum & Mitchell, COLT 1998]:
  - If $X \perp Z \mid Y$, then bootstrapping methods "work"
- **Canonical Correlation Analysis** [Kakade & Foster, COLT 2007]:
  - Suppose there is redundancy of views via linear predictors:
    for each $V \in \{X, Z\}$

$$R^2_{V,Y} \;\geq\; R^2_{(X,Z),Y} - \epsilon$$

  - Then CCA-based dimension reduction preserves linear predictability of Y
  - (No assumption of conditional independence!)

Q: What if views are redundant only via non-linear predictors?

Multi-view redundancy

ϵ-multi-view redundancy assumption:

$$\mathbb{E}[(\mathbb{E}[Y \mid V] - \mathbb{E}[Y \mid X, Z])^2] \leq \epsilon \text{ for each } V \in \{X, Z\}.$$

Surrogate predictor: $\mu(x) \coloneqq \mathbb{E}[\mathbb{E}[Y \mid Z] \mid X = x]$

Best (possibly non-linear) prediction of Y using Z

Lemma: If ϵ-multi-view redundancy holds, then

$$\mathbb{E}[(\mu(X) - \mathbb{E}[Y \mid X, Z])^2] \leq 4\epsilon.$$

We'll show:

Learned feature map $\vec{\phi}(x)$ satisfies $\mu(x) \approx$ linear function of $\vec{\phi}(x)$

Approximating the surrogate predictor

$$g^*(x,z) = \frac{\Pr[\text{pos} \mid x,z]}{\Pr[\text{neg} \mid x,z]} = \frac{P_{X,Z}[x,z]}{P_X[x]P_Z[z]}$$

$$\mu(x) = \mathbb{E}[\,\mathbb{E}[\,Y \mid Z\,] \mid X = x\,]$$

$$= \mathbb{E}[\mathbb{E}[Y \mid Z]g^*(x,Z)]$$ since $g^*(x,z)P_Z(dz) = P_{Z|X=x}(dz)$

$$\approx \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[\,Y \mid Z = l_i\,]g^*(x,l_i)$$ with $l_1, \dots, l_M \sim_{iid} P_Z$

$$= \vec{w} \cdot \vec{\phi}^*(x)$$ using $\vec{\phi}^*(x) \coloneqq \left(g^*(x,l_1), \dots, g^*(x,l_M)\right)$

Theorem: Under ϵ-multi-view redundancy assumption, w.h.p.,

$$\min_{\vec{w}} \mathbb{E}\left[\left(\vec{w} \cdot \vec{\phi}^*(X) - \mathbb{E}[\,Y \mid X,Z\,]\right)^2\right] \leq 4\epsilon + O(1/M)$$

Error transform theorem

The learned $\vec{\phi}$ is based on odds-ratio estimate $\hat{g}$ that only approximately solves contrastive learning problem (say, with respect to cross entropy loss).

Theorem: Under ϵ-multi-view redundancy assumption, w.h.p.,

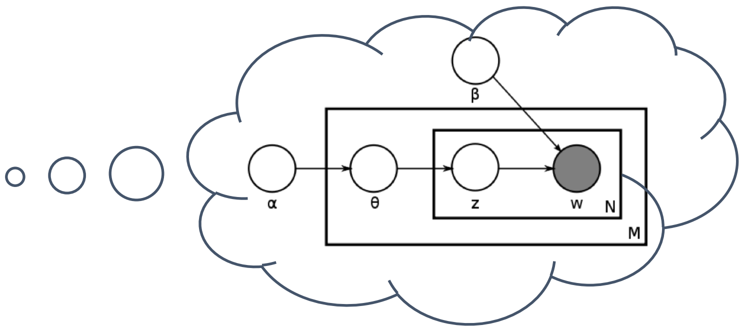
$$\min_{\vec{w}} \mathbb{E}\left[\left(\vec{w} \cdot \vec{\phi}(X) - \mathbb{E}[Y \mid X, Z]\right)^2\right] = O\left(\text{error}(\hat{g})\right) + 4\epsilon + O(1/M)$$

Error in down-stream prediction task

Contrastive learning error (excess cross entropy loss)

3. Interpreting the representation

What's in the representation?



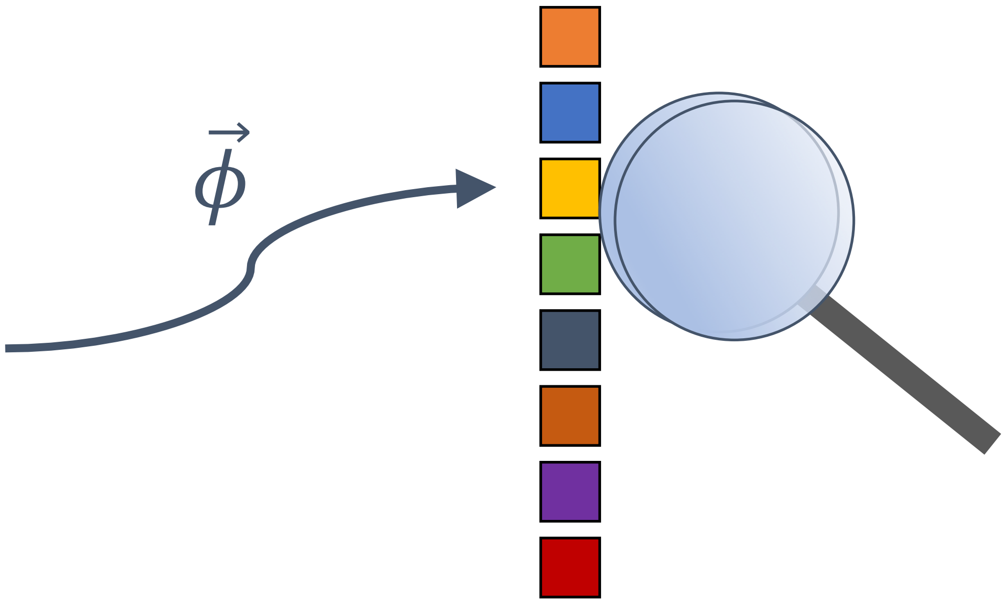
To interpret the representations, we look to probabilistic models...

The Daily

Friday, February 28, 2020

Contrastive estimation revea

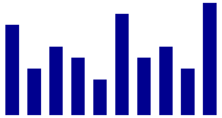
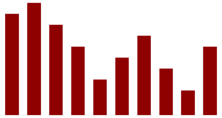
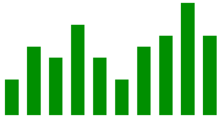
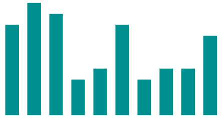
Using unlabeled data to find useful embeddings is a central challenge in the field of representation learning. Classical approaches to this task often start by fitting some type of structure to the unlabeled data, such as a generative model or a dictionary, and then embed future data by performing inference using the fitted structure (Blei et al., 2003; Raina et al., 2007). While this approach has sometimes enjoyed good empirical performance, it is not without its drawbacks. One issue is that learning structures and performing inference is often hard in general (Sontag and Roy, 2011; Arora et al., 2012). Another issue is that we must a priori choose a structure and method for fitting the unlabeled data, and unsupervised methods for learning these structures can be sensitive to model misspecification (Kulesza et al., 2014).



Topic model [Hofmann, 1999; Blei, Ng, Jordan, 2003; …]

- K topics, each specifies a distribution over the vocabulary
- A document is associated with its own distribution w over K topics
- Words in document (BoW): i.i.d. from induced mixture distribution
  - Assume they are arbitrarily partitioned into two halves, x and z

E.g.,

$\sim$iid $\frac{1}{5}$  $+\frac{2}{5}$  $+\frac{2}{5}$  $+0$ 

sports science politics business

For now, assume document is about <u>single topic</u> (one of $\{t_1, t_2, \ldots, t_K\}$)

Interpreting the density ratio…

Conditional independence assumption + Bayes rule

Density ratio →

$$\frac{P_{X,X}(x,z)}{P_X(x)P_Z(z)} = \sum_{k=1}^{K} \frac{\Pr(t_k \mid x)\Pr(z \mid t_k)}{P_Z(z)}$$

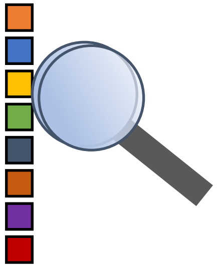
$$= \frac{\vec{\pi}(x) \cdot \vec{\lambda}(z)}{P_Z(z)}$$

Posterior over topics given x

Likelihoods of topics given z

Inside the feature map

$L\,\vec{\pi}(x)$



- **Embedding**: $\vec{\phi}^*(x) = (g^*(x, l_i) : i = 1, \ldots, M)$ where

$$g^*(x, z) \propto \vec{\pi}(x) \cdot \vec{\lambda}(z)$$

- Therefore

$$\vec{\phi}^*(x) = D\left[\vec{\lambda}(l_1) \quad \cdots \quad \vec{\lambda}(l_M)\right]^\top \vec{\pi}(x)$$

(for some diagonal matrix D)

Likelihoods of topics given l_i's

Posterior over topics given x

Interpretation

- In the "one topic per document" case, document feature map is a linear transformation of the posterior over topics

$$\vec{\phi}^*(x) = L\,\vec{\pi}(x)$$

- **Theorem**: If L is full-rank, every linear function of topic posterior can be expressed as a linear function of $\vec{\phi}^*(\cdot)$

For more general models, get theorem in terms of posterior moments.

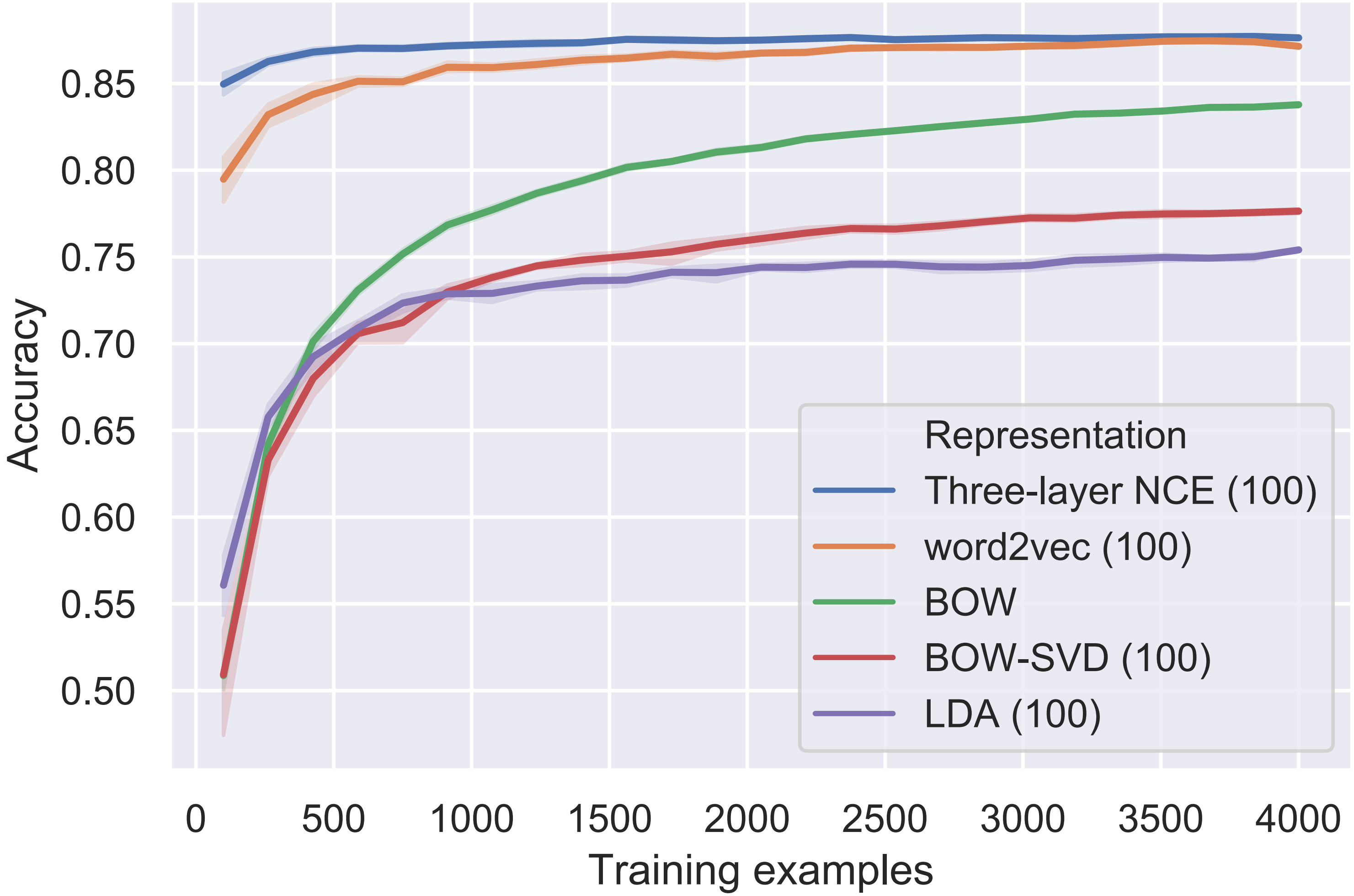
4. Experimental study

Study dataset and comparisons

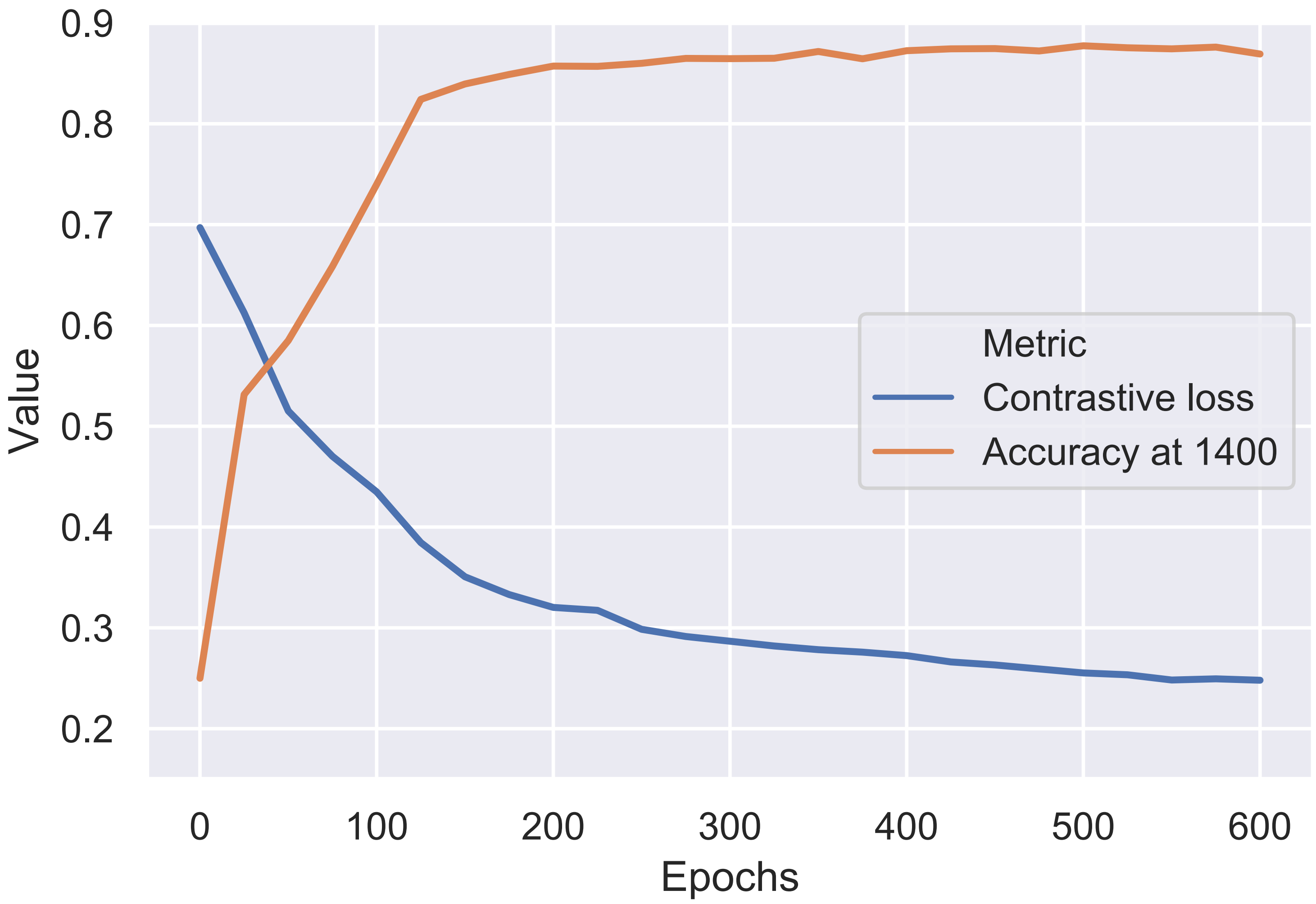
- **AG News** [Del Corso, Gulli, Romani, 2005; Zhang, Zhao, LeCun, 2015]: Four categories (world, sports, business, sci/tech) of news articles
  - 16,700 words in vocabulary after removing rare words; avg. ~45 words/document
  - Use 4 x 29,000 unlabeled examples for contrastive learning to get $\vec{\phi}$
  - Use (up to) 4 x 1,000 labeled examples to train linear classifier (multi-class logreg)
  - Use 4 x 1,900 labeled examples for test set
- Our feature map $\vec{\phi}$ (called "NCE" for Noise Contrastive Embedding):
  - Three-layer ReLU networks with ~300 nodes/layer
  - Dropout regularization, batch normalization, PyTorch initialization
  - Trained using RMSProp
- Baseline feature maps $\vec{\phi}$:
  - word2vec [Mikolov *et al*, 2013], Latent Dirichlet Allocation [Blei *et al*, 2003], BoW

Accuracy on supervised task vs # sample size

$\vec{\phi}(x) \in \mathbb{R}^M$ for $M = 100$

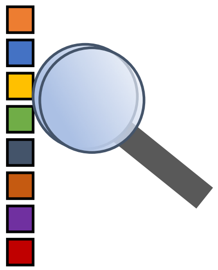


Performance on contrastive task vs accuracy



In closing…

Broader theme: Study "deep learning"-style representation learning through the lens of probabilistic models



- Multi-view redundancy (à la CCA)
- Topic models and other multi-view mixture models
- …

Acknowledgements

- Thanks to Miro Dudík for initial discussions & suggestions
- Support from NSF CCF-1740833 & JP Morgan Faculty Award

Thanks!

Related / complementary analyses

- Steinwart, Hush, Scovel (2005), Abe, Zadrozny, Langford (2006)
  - Use NCE to for estimating density level sets / outlier detection
- Gutmann & Hyvärinen (2010)
  - Use NCE to fit statistical models with intractable partition functions
- Arora, Khandeparkar, Khodak, Plevrakis, Saunshi (2019)
  - If X, Z are **conditionally independent given class label**, then contrastive learning gives linearly useful representations