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

Learned from data

Input Space  Feature Space
Unsupervised / semi-supervised learning

Unlabeled data $\phi$

Labeled data $f$

Feature map

Down-stream prediction task

Self-supervised learning
"Self-supervised learning"

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

**Predict color channel from grayscale channel**

**Predict missing word in a sentence from context**

The quick brown fox ____ over the lazy dog.

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

---

[Zhang, Isola, Efros, 2017]

[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 $\tilde{\phi}$, learned with only $\sim10\%$ of labels, are near SOTA Top-5 error on ImageNet

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 $\phi(X)$ that achieves near-optimal MSE.
Our main results (2)

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

$\sim \text{iid} \quad \frac{1}{5} \quad \text{sports} \quad + \quad \frac{2}{5} \quad \text{science} \quad + \quad \frac{2}{5} \quad \text{politics} \quad + \quad 0 \quad \text{business}$
Rest of the talk

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

• 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 \( \tilde{\phi} \):

\[
\tilde{\phi}(x) := (\hat{g}(x, l_i) : i = 1, \ldots, M) \in \mathbb{R}^M
\]

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
  • $Y = \text{topic of article}$
  • $X = \text{abstract}$
  • $Z = \text{introduction}$
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: \(\mu(x) := \mathbb{E}[^{\text{\square}} \mathbb{E}[Y \mid Z] \mid X = x]\)

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 \(\tilde{\phi}(x)\) satisfies \(\mu(x) \approx \text{linear function of } \tilde{\phi}(x)\)
Approximating the surrogate predictor

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

$$= \mathbb{E}[\mathbb{E}[Y \mid Z]g^*(x, Z)]$$

$$= \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[Y \mid Z = l_i]g^*(x, l_i)$$

$$= \vec{w} \cdot \vec{\phi}^*(x)$$

since $$g^*(x, z)P_Z(dz) = P_{Z\mid X=x}(dz)$$

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

with $$l_1, \ldots, l_M \sim \text{iid } P_Z$$

**Theorem:** Under $$\epsilon$$-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 $\epsilon$-multi-view redundancy assumption, w.h.p.,

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

- 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...
Topic model

• $K$ 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.,

\[ \sim i i d \quad \frac{1}{5} \text{ sports } + \frac{2}{5} \text{ science } + \frac{2}{5} \text{ politics } + 0 \text{ business} \]

For now, assume document is about single topic (one of $\{t_1, t_2, \ldots, t_K\}$)
Interpreting the density ratio...

Conditional independence assumption + Bayes rule

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

\[
= \frac{\tilde{\pi}(x) \cdot \tilde{\lambda}(z)}{P_Z(z)}
\]

Posterior over topics given \(x\) \hspace{1cm} \text{Likelihoods of topics given } z
Inside the feature map

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

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

- Therefore

  $$\tilde{\phi}^*(x) = D[\tilde{\lambda}(l_1) \cdots \tilde{\lambda}(l_M)]^\top \bar{\pi}(x)$$

  (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
  \[ \phi^*(x) = L \tilde{\pi}(x) \]

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

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 $\phi$
  • 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 $\phi$ (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 $\hat{\phi}$:
  • word2vec [Mikolov et al, 2013], Latent Dirichlet Allocation [Blei et al, 2003], BoW
Accuracy on supervised task vs # sample size

$\tilde{\phi}(x) \in \mathbb{R}^M$ 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