Blake Bordelon, Cedric Flamant
Salma Abdel Magid

CS 281
November 4, 2019

Sample Efficiency and Generality of Contrastive Representation Learning

Learning representation functions with unlabeled data has been shown to be beneficial for downstream classification tasks [1] [2] [3]. Recent theoretical works have attempted to explain utility of unsupervised representation learning for classification by invoking the concept of latent classes and hypothesizing that similar points are sampled from the same latent class [4] [5]. Particular focus has been given to the set of so-called contrastive representation learning algorithms, which leverage prior assumptions about which data points should have similar embedded representations (*i.e.* fragments of nearby text in a corpus or nearby frames in a video). Although these works have refined the concepts of contrastive learning and the necessity of negative sampling (penalizing similar representations of data points expected to be dissimilar), there has been little work evaluating the sample efficiency of such schemes compared to baseline supervised learning algorithms. Nor has a satisfying theoretical or empirical exploration of the generality of such representations been undertaken. Such questions are pressing given the ubiquity of unlabeled data and the difficulty of acquiring category data, for example, in the case of natural images [6].

0.1 Setup

In contrastive representation learning, the loss for a representation function $f \in \mathcal{F}$ depends on dot product similarities of the representation at different points

$$\mathcal{L}_{un}(f) = \mathbb{E}_{x,x^+,x^-}\ell(f(x)^\top f(x^-) - f(x)^\top f(x^+)) \tag{1}$$

where x and x^+ are assumed to be drawn from the same latent class (since they are sampled nearby etc.) and x^- is sampled with high probabilty from outside of the latent class from which x was sampled. ℓ is a pointwise loss function, taken to be convex (like the hinge or the logistic loss).

For a downstream classification task $\mathcal{T}$, we then train a classifier $C(f(x)) = Wf(x)$ with a fixed learning algorithm on the representation function f. This allows us to have a supervised loss for the representation function $\mathcal{L}_{sup}(f, \mathcal{T})$. Arora *et. al.* show that in a limited sense $\mathcal{L}_{un}(f)$ acts as a surrogate for $\mathcal{L}_{sup}(f, \mathcal{T})$ but also demonstrate that contrastive algorithms do not always pick out the best representation function.

0.2 Objectives and Methods

Our big picture goals on this project are the following

1. Characterize the sample efficiency of this class of algorithms empirically, and if possible, theoretically. Given a budget of m unlabeled data points with a similarity prior and

n data points with class labels, can we give a probabilistic bound on the classification error? How much does the m additional unlabeled data points improve performance compared to directly training a supervised algorithm on the n labeled data points?

2. Attempt to provide a rigorous definition of representational generality in terms of the number of k-way classification tasks that have supervised loss below some value.

3. Empirically measure this two dimensional (m, n) sample efficiency and the generality of learned representations when trained on synthetic toy data sets and real datasets.

4. Compare the gain in performance that using a trained contrastive representation function offers against a baseline of a randomly instantiated embedding to quantify the benefit of contrastive representation learning over reservoir computing.

5. Explore other unsupervised costs to see if a stronger result about downstream classification performance can be derived than that found by Arora *et. al.* [5].

0.3 Implementation

To explore these questions, we plan to train a neural network to implement $f(x)$ by performing empirical risk minimization on an empirical version of the unsupervised cost $\hat{\mathcal{L}}_{un}$ which is determined by m unlabeled data points. Then we will train a linear classifier C on the predictions $\{f(x_i), y_i\}$ for all (x_i, y_i) in the labeled dataset of size n. Errors from the unsupervised loss can be backpropagated to learn $f(x)$ and the classifier C will be learned efficiently with multi-class logistic regression after training $f(x)$. We will then explore the dependence of the final classification test risk on m and n.

0.4 Datasets

We are considering two domains for experimentation: natural images and text data. In the former, we will use DAVIS: Densely Annotated Video Segmentation. DAVIS is used for unsupervised multi-object video object segmentation tasks. Each high-resolution full HD video sequence in this data set is accompanied by densely annotated, pixel-accurate and per-frame ground truth segmentation. The sequences have been carefully captured to cover multiple instances of major challenges typically faced in video object segmentation. Each video is annotated with specific attributes such as occlusions,fast-motion, non-linear deformation and motion-blur [7]. Nearby frames will be assumed to be similar and randomly sampled frames from other clips will be assumed dissimilar.

In the latter domain, we will use the Reddit Sarcasm data set. This data set contains 1.3 million sarcastic comments from the popular social commentary site, Reddit. Reddit is the sixth most popular website in the United States and eighteenth worldwide. The data set was generated by scraping comments which contained the "/s" tag (which is used to denote

a sarcastic comment) [8]. To perform a semantic analysis, we will sample words or phrases within a post as similar and words/phrases from other posts as dissimilar.

0.5 What does success look like? What are the risks?

1. Successfully implementing this two stage learning procedure in PyTorch and get performance above the reservoir computing baseline.
2. Developing an empirical estimate of the 2D (m, n) test risk for classification tasks on natural language and annotated natural video datasets.
3. Developing a measure of generality of learned representations and using it to quantify the utility of learning the representation function.
4. The major risk with this type of algorithm is that it is difficult to know that the samples we suspect are negative samples are in fact drawn from different latent classes. If we had that information in the original dataset, we may as well perform supervised learning on all $n + m$ points. We are curious how much this effect matters in practice and will examine the effect of negative sampling from the wrong classes when the ground truth labels are known.

References

[1] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations, 2018.

[2] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos, 2015.

[3] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

[4] Sanjeev Arora and Andrej Risteski. Provable benefits of representation learning, 2017.

[5] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019.

[6] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[7] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. 2016.

[8] A large self-annotated corpus for sarcasm. 2017.