CS 839: Theoretical Foundations of Deep Learning		Spring 2022
Lecture 18 Contrastive Representation Learning		
Instructor: Yingyu Liang	*Date: Mar 31st, 2022*	*Scriber: Vasileios Kontonis*

1 Motivation and Model Definition

We consider the problem of unsupervised representation learning. In this setting we use unlabeled data to learn a representation function $f(x)$ of features x so that **classification using $f(x)$ in place of x is easier and requires less labeled data**. For example, such a representation for image classification is the output of the second to last layer of some big deep net trained on a lot of data (e.g., on ImageNet). Another example is text embeddings which are low-dimensional representations of pieces of text that are used in NLP.

To learn the representation f we assume access to pairs of data points (x, x^+) (text or images) that are more similar than randomly sampled points. We try to find f that maps x and x^+ to vectors closer (more parallel), i.e., we want the inner product $f(x)^T f(x^+)$ to be much larger than $f(x)^T f(x^-)$, where x^- is simply a randomly selected datapoint. More concretely, the following loss function is considered (to learn representations for sentences [2]):

$$\mathbf{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right] \tag{1}$$

It seems intuitive that minimizing such loss functions should lead to representations that capture similarity. In this lecture we present a theoretical model introduced in [1] that shows (under assumptions) that the learned representations (using only unlabeled data) should do well on the relevant classification task. They use the term Contrastive Learning to refer collectively to representation learning methods that assume access to similar pairs (x, x^+) and to some x^- dissimilar to x (we will call x^- a negative sample).

We first set up notation and describe the framework for unlabeled data and classification tasks that will be essential for our analysis. Let $\mathcal{X}$ denote the set of all possible data points. Contrastive learning assumes access to *similar* data in the form of pairs (x, x^+) that come from a distribution $\mathcal{D}_{sim}$ as well as k i.i.d. *negative samples* $x_1^-, x_2^-, \ldots, x_k^-$ from a distribution $\mathcal{D}_{neg}$ that are presumably unrelated to x. We denote by $\mathcal{F}$ the class of representation functions $f : \mathcal{X} \to \mathbb{R}^d$, such that $\|f(\cdot)\| \leq R$ for some $R > 0$.

Similarity Distributional Model

To formalize the notion of similar pairs (x, x^+), we introduce the concept of latent classes. Let $\mathcal{C}$ denote the set of all latent classes. Each class $c \in \mathcal{C}$ induces a probability distribution $\mathcal{D}_c$ over $\mathcal{X}$ that represents how relevant x is to class c. For example, $\mathcal{X}$ could be images and c the class "dog" whose associated $\mathcal{D}_c$ assigns high probability to images containing dogs and low probabilities to other images. We notice that in this model, classes can overlap: an image of a baby playing with a dog can be sampled both from $\mathcal{D}_{dog}$ and from $\mathcal{D}_{baby}$. We also assume a distribution ρ over the classes that captures how often each class c occurs in the

unlabeled data. We assume that similar data points x, x^+ are drawn i.i.d. from the same class distribution $\mathcal{D}_c$ for some class c. On the other hand, negative samples are drawn from the marginal of $\mathcal{D}_{sim}$:

$$\mathcal{D}_{sim}(x, x^+) = \mathop{\mathbf{E}}_{c\sim\rho}[\mathcal{D}_c(x)\mathcal{D}_c(x^+)] \tag{2}$$

$$\mathcal{D}_{neg}(x^-) = \mathop{\mathbf{E}}_{c\sim\rho}[\mathcal{D}_c(x^-)] \tag{3}$$

In other words the generative process of a similar pair (x, x^+) and negative sample x^- given in the unsupervised representation learning algorithm is as follows:

1. Sample two classes $c_1 \sim \rho$ and $c_2 \sim \rho$ independently.
2. Sample x, x^+ independently from the same class marginal $\mathcal{D}_{c_1}$, i.e., $(x, x^+) \sim \mathcal{D}_{c_1}^2$.
3. Independently (from all the above) draw a negative sample $x^- \sim \mathcal{D}_{c_2}$.

This is the most important modelling assumption that we are making. It is plausible since it allows for classes to overlap but it is still rather strong.

Testing the Quality of Representations

We formalize the classification tasks that a representation function f will be tested on. For simplicity we will focus on a binary supervised task $\mathcal{T}$ that consists of two classes $\{c^+, c^-\} \subseteq \mathcal{C}$. The labeled dataset for the task $\mathcal{T}$ consists of m i.i.d. draws from the following generative process:

1. A label $c \in \{c^+, c^-\}$ is picked according to a distribution $\mathcal{D}_{\mathcal{T}}$. We will assume that $\mathcal{D}_{\mathcal{T}}$ is simply the uniform distribution over the two labels. We refer to [1] for the more general case.
2. A datapoint x is drawn from the class conditional distribution $\mathcal{D}_c$.

Therefore, a labeled pair (x, c) has joint distribution

$$\mathcal{D}_{\mathcal{T}}(x, c) = \mathcal{D}_{\mathcal{T}}(c)\mathcal{D}_c(x) \tag{4}$$

We observe that the data distributions $\mathcal{D}_c$ of the classification task are the same as in the unlabeled data. This allows for theoretically justifying that capturing similarity in unlabeled data leads to quantitative guarantees on the classfication tasks.

The quality of the representation function f is evaluated by its performance on a classification task $\mathcal{T}$ using linear classification. A binary classifier for $\mathcal{T}$ is a function $g : \mathcal{X} \to \mathbb{R}^2$ whose output coordinates are indexed by the classes c in task $\mathcal{T}$ The loss incurred by g on point $(x, y) \in \mathcal{X} \times \mathcal{T}$ is defined as $L(g(x)_y - g(x)_{y'})$ with $y' \neq y$. Since we are focusing on binary classification given a label y there always one option for $y' \neq y$ and therefore we can think of the loss function $L(z)$ as a one-dimensional function. For example, $L(z)$ may be logistic loss $L(z) = \ln(1 + \exp(-z))$ for $z \in \mathbb{R}$. Therefore, when the observed label is $y = c^+$ we look in the difference $g(x)_{c^+} - g(x)_{c^-}$. If this difference is large (which means that our

classifier g is putting a lot of weight on the true class) then its logistic loss will be small. We refer to [1] for the corresponding definitions of the more general multiclass setting. We shall use the notation $\{g(x)_c - g(x)_{c'}\}_{c' \neq c} \in \mathbb{R}$ to denote the difference between the observed class c and the other class c'. The supervised loss of the classifier g is then:

$$L_{sup}(\mathcal{T}, g) \coloneqq \mathop{\mathbf{E}}_{(x,c)\sim\mathcal{D}_{\mathcal{T}}} \left[L\big(\{g(x)_c - g(x)_{c'}\}_{c' \neq c}\big) \right]$$

To use a representation function f with a linear classifier, we train linear weight layer $W \in \mathbb{R}^{k \times d}$ on top of the representation $f(x)$. In other words, the final classifier is $g(x) = Wf(x)$. For our simpler case of binary classification, we have that the matrix W has only two rows, i.e., $W \in \mathbb{R}^{2 \times d}$. Since the best W can be found by fixing f and training a linear classifier, we abuse notation and define the *supervised loss* of f on $\mathcal{T}$ to be the loss when the best W is chosen for f:

$$L_{sup}(\mathcal{T}, f) = \inf_{W \in \mathbb{R}^{2 \times d}} L_{sup}(\mathcal{T}, Wf) \tag{5}$$

Since contrastive learning has access to data with latent class distribution ρ, it is natural to have better guarantees for tasks involving classes that have higher probability in ρ. We can take the average over a pair of independent labels (conditional that they are different) and define the following average loss of a representation f:

$$L_{sup}(f) \coloneqq \mathop{\mathbf{E}}_{c^+, c^- \sim \rho^2} \left[L_{sup}(\{c^+, c^-\}, f) \mid c^+ \neq c^- \right]$$

2 Contrastive Learning and Main Theorem

We describe the training objective for contrastive learning. We choose the same loss function L as in the supervised problem above. Let $(x, x^+) \sim \mathcal{D}_{sim}$ be a similar pair and $x^- \sim \mathcal{D}_{neg}$ a negative sample. The population **unsupervised loss** used by the contrastive algorithm is

$$L_{un}(f) \coloneqq \mathbf{E}\left[L\left(f(x)^T \big(f(x^+) - f(x^-) \big) \right) \right] \tag{6}$$

and its empirical version is $\widehat{L}_{un}(f) = \frac{1}{M} \sum_{j=1}^{M} L\left(f(x_j)^T \left(f(x_j^+) - f(x_j^-) \right) \right)$.

Note that, by the assumptions of the framework described above (see the generative process of similar and negative samples), we can express the unsupervised loss as

$$L_{un}(f) = \mathop{\mathbf{E}}_{c^+, c^- \sim \rho^2} \mathop{\mathbf{E}}_{\substack{x, x^+ \sim \mathcal{D}^2_{c^+} \\ x^- \sim \mathcal{D}_{c^-}}} \left[L\left(f(x)^T \left(f(x^+) - f(x^-) \right) \right) \right]$$

In order to learn a representation function from $\mathcal{F}$ we try to find a function that minimizes the empirical unsupervised loss. The representation $\widehat{f}$ will be then used to improve the performance of supervised linear classification tasks. The ideal result would be to show that the supervised error of the learned representation $\widehat{f}$ is better than the error of any other f.

$$L_{sup}(\widehat{f}) \leq \inf_{f \in \mathcal{F}} L_{sup}(f)\,.$$

We cannot prove the above but we can prove the following theorem on the performance of the learned representation $\widehat{f}$.

Theorem 1. Let $\tau = \mathbf{P}_{c,c' \sim \rho^2}[c = c']$ be the probability that two random classes coincide (hitting probability). Then it holds that:

$$L_{sup}(\widehat{f}) \leq \frac{1}{1-\tau}(L_{un}(f) - \tau) + \frac{1}{1-\tau} Gen \qquad \forall f \in \mathcal{F}$$

The above theorem essentially translates the unsupervised error of f into a supervised error guarantee. To prove the theorem we first we need to handle the generalization error. We can show that with probability at least $1-\delta$ over the training set $\mathcal{S}$, of size M, for all $f \in \mathcal{F}$ it holds that

$$L_{un}(\widehat{f}) \leq L_{un}(f) + \epsilon_M \,.$$

The standard generalization argument to prove the above is to first show that for every $f \in \mathcal{F}$ it holds $|\widehat{L}_{un}(f) - L_{un}(f)| \leq \epsilon_M/2$. We can do that using some complexity measure (for example Rademacher complexity). We then have that for every $f \in \mathcal{F}$ it holds

$$L_{un}(\widehat{f}) \leq \widehat{L}_{un}(\widehat{f}) + \epsilon_M/2 \leq \widehat{L}_{un}(f) + \epsilon_M/2 \leq L_{un}(f) + \epsilon_M \,.$$

To keep this presentation simple we shall ignore the generalization error (for more details see [1]) and focus on the following key lemma. We shall define the following notion of the mean classifier using a specific W where the rows are the means of the representations of each class.

Definition 2 (Mean Classifier)**.** For a function f and task $\mathcal{T} = (c^+, c^-)$, the mean classifier is W^μ whose c^{th} row is the mean μ_c of representations of inputs with label c: $\mu_c \coloneqq \mathop{\mathbf{E}}_{x \sim \mathcal{D}_c}[f(x)]$. We denote its loss $L_{sup}(\mathcal{T}, W^\mu f)$ by $L^\mu_{sup}(\mathcal{T}, f)$. The average supervised loss of its *mean classifier* is

$$L^\mu_{sup}(f) \coloneqq \mathop{\mathbf{E}}_{c^+, c^- \sim \rho^2} \left[L^\mu_{sup}(\{c^+, c^-\}, f) \mid c^+ \neq c^- \right]$$

We now prove the following key lemma that bounds by above the supervised loss of the mean classifier.

Lemma 3. For all $f \in \mathcal{F}$ it holds that

$$L^\mu_{sup}(f) \leq \frac{1}{(1-\tau)} \left(L_{un}(f) - \tau \right)$$

Proof. The key idea in the proof is the use of Jensen's inequality. Unlike the unsupervised loss which uses a random point from a class as a classifier, using the mean of the class as the classifier should only make the loss lower. Let $\mu_c = \mathop{\mathbf{E}}_{x \sim \mathcal{D}_c} f(x)$ be the mean of the class c. Using the definitions of $\mathcal{D}_{sim}$ and $\mathcal{D}_{neg}$ we have

$$L_{un}(f) = \mathop{\mathbf{E}}_{\substack{(x,x^+) \sim \mathcal{D}_{sim} \\ x^- \sim \mathcal{D}_{neg}}} \left[L(f(x)^T (f(x^+) - f(x^-))) \right] = \mathop{\mathbf{E}}_{\substack{c^+, c^- \sim \rho^2 \\ x \sim \mathcal{D}_{c^+}}} \mathop{\mathbf{E}}_{\substack{x^+ \sim \mathcal{D}_{c^+} \\ x^- \sim \mathcal{D}_{c^-}}} \left[L(f(x)^T (f(x^+) - f(x^-))) \right] .$$

Using the convexity of the loss L and Jensen's inequality, we obtain

$$\begin{aligned} L_{un}(f) &\geq \mathop{\mathbf{E}}_{c^+,c^-\sim\rho^2} \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{c^+}} \left[L(f(x)^T(\mu_{c^+} - \mu_{c^-}))\right] \\ &= (1-\tau) \mathop{\mathbf{E}}_{c^+,c^-\sim\rho^2} [L^{\mu}_{sup}(\{c^+, c^-\}, f) | c^+ \neq c^-] + \tau \\ &= (1-\tau) L^{\mu}_{sup}(f) + \tau \end{aligned}$$

where the first equality follows by splitting the expectation into the cases $c^+ = c^-$ and $c^+ \neq c^-$, and the final equality follows by using the symmetry in c^+ and c^- since we assumed that classes in tasks are uniformly distributed. □

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *CoRR*, abs/1902.09229, 2019.

[2] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *CoRR*, abs/1803.02893, 2018.