Sampling Informative Positives Pairs in Contrastive Learning

Anonymous Author(s)

***Abstract*—Contrastive Learning is a paradigm for learning representation functions that recover useful similarity structure in data based on samples of positive (similar) and negative (dissimilar) instances. Typically, positive instances are sampled by randomly perturbing an anchor point using some form of data augmentation, while negative instances are sampled independently from the same distribution as the anchor points. The goal is to learn a representation function that places each anchor near its positive instances and far from its negative instances. However, not all randomly sampled positive instances are equally effective in learning a representation function that captures useful structure in the data. We consider a setting where class structure in the observed data derives from analogous structure in an unobserved latent space. We propose active sampling approaches for positive instances and investigate their role in effectively learning representation functions which recover the class structure in the underlying latent space.**

***Index Terms*—Contrastive Learning, Active Learning**

I. Introduction

Representation learning seeks to identify salient structure in high-dimensional data and compute data representations that are useful for downstream tasks. Representations may be learned from labeled (supervised) or unlabeled training data (unsupervised). In *Contrastive Learning*, a representation function is trained to recover similarity structure in data from pairs of similar and dissimilar observations [van den Oord et al., 2018, Bachman et al., 2019, Chen et al., 2020, Khosla et al., 2020, Zbontar et al., 2021]. Applications in a wide range of domains, including Computer Vision [Bachman et al., 2019, Ma et al., 2021], NLP [Gao et al., 2021, Meng et al., 2021] and Medical Imaging [Fedorov et al., 2021], have demonstrated the promise of this paradigm (see Apx. A for related work).

To train a representation function, we sample positive (similar) and negative (dissimilar) instances. Positives may be generated via data augmentations such as cropping, flipping or blurring of images (Computer Vision) or via sampling of adjacent sentences (NLP). Negatives are often generated by randomly sampling images in a data set or sentences in a corpus. The quality of the resulting representation function depends crucially on the quality of the positives and negatives used for training, i.e., how informative they are about salient structure in the underlying data distribution. Here, we investigate how the choice of sampling strategies impacts the quality of the learned representation function and how we can use such insight to design effective sampling strategies, focusing on the sampling of *positive* instances. We consider a setting where the learner has access to high-dimensional observations generated from unobserved latent variables that are sampled from a hidden latent space. We seek to train a representation function

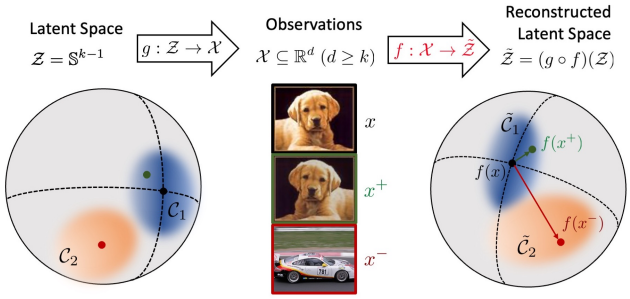

Fig. 1: **Contrastive learning in the latent variable model.**

which approximately recovers low-dimensional structure (latent classes) in the unobservable latent space (Fig. 1). Representation functions are often evaluated with respect to how well simple prediction functions built on top of them perform on downstream tasks. Here, we analyze the quality of our learned representation function with respect to classification in the reconstructed latent space. Our contributions are:

1) We propose an iterative contrastive approach (Alg. 1) which selects positives that uncover weaknesses in the representation function via *active sampling* near the decision boundary of a classifier trained in the reconstructed latent space. We support our approach with theoretical and experimental evidence.
2) We consider a setting where useful structure in the observations depends strongly on an *informative* subset of latents and weakly on the remaining *uninformative* latents. We study sampling strategies (Alg. 2) that generate positive instances with a bias towards capturing variability in the informative subset of latents. We present experimental evidence for the effectiveness of such a strategy.

II. Background and Notation

A. Latent Variable Model

Consider a latent variable model (Fig. 1), where $\mathcal{X} \subseteq \mathbb{R}^d$ is the space of observations and $\mathcal{Z} \subseteq \mathbb{R}^k$ $(k \leq d)$ the space of latent variables. We observe $x \in \mathcal{X}$, generated by a map $g : \mathcal{Z} \to \mathcal{X}$, assumed to be smooth and nonzero. Following standard convention, we identify $\mathcal{Z}$ with the unit hypersphere $\mathbb{S}^{k-1} = \{z \in \mathbb{R}^k : z^T z = 1\}$. We assume a latent class structure in $\mathcal{Z}$, i.e., $\mathcal{C} = \{C_1, C_2, \dots\} \subseteq \mathcal{Z}$ (with class labels $c_1, c_2, \dots$), each defined by a certain region on the hypersphere, which, geometrically, corresponds to a spherical cap. Let $c \sim p_c$ denote a distribution over the latent classes and $z \sim p_c(z)$ the uniform distribution over the class c. Furthermore, $(z, z^+) \sim p(z^+|z)p_c(z)$ denotes a positive pair of latents sampled conditioned on the class c: we first sample

an *anchor* z given c and then a positive z^+ conditioned on z. Negatives are sampled from the marginal over $\mathcal{Z}$, i.e., by drawing $c' \sim p_c$, $z^- \sim p_{c'}(z)$. Let $p_{data}(\cdot|c)$ denote a data distribution on $\mathcal{X}$ and $p_+(\cdot|x,c)$ a distribution of positives x^+ on $\mathcal{X}$, conditioned on an anchor x; $p(z|x,c)$ denotes the (unknown) posterior distribution over latents that generated x.

B. Contrastive Loss

We want to learn a representation function $f : \mathcal{X} \to \mathbb{S}^{k-1} \in \mathcal{F}$ that recovers latents from observations; f defines a distribution $p(\tilde{z}|x,c)$ over reconstructed latents given an observation x. To learn f we minimize the *contrastive loss*

$$\mathcal{L}(f;\tau,m) := \mathop{\mathbb{E}}_{\substack{(x,x^+)\sim p_+ \\ x_i^- \sim p_{data}}} \left[-\log \frac{e^{\frac{f(x)^T f(x^+)}{\tau}}}{e^{\frac{f(x)^T f(x^+)}{\tau}} + \sum_{i=1}^m e^{\frac{f(x^-)^T f(x^+)}{\tau}}} \right]$$

which learns representations that are similar for positive pairs and dissimilar for negative pairs. We denote the reconstructed latent space as $\tilde{\mathcal{Z}}$ and set $\tilde{z} := h(z) = (f \circ g)$, where $f \leftarrow \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}(f)$, with the superscript $\sim$ indicating objects in $\tilde{\mathcal{Z}}$. The minimum $f^* \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}(f)$ can be seen as an inversion of the generative map g [Zimmermann et al., 2021], i.e., the composition $h^* := f^* \circ g$ preserves the alignment between latents. Note that this requires that h^* preserves the dot products between positive pairs (z, z^+) up to a constant, i.e., $\kappa z^T z^+ = h^*(z)^T h^*(z^+)$ (with $\kappa > 0$), which is equivalent to requiring that h^* locally reconstructs the latent space up to linear and orthogonal transformation (see also Apx. B).

III. Informative Positives for Latent Classes

We want to identify classes $\tilde{\mathcal{C}} = \{\tilde{C}_1, \tilde{C}_2, \dots\} \subseteq \tilde{\mathcal{Z}}$ in the reconstructed latent space that recover the true latent classes $\mathcal{C} = \{C_1, C_2, \dots\} \subseteq \mathcal{Z}$. We propose to sample positives that are *informative* about the latent classes, with the goal of training representation functions more efficiently. We first consider a setting where we know the number of classes $\mathcal{C}$ and that they are linearly separable. Our iterative approach (Alg. 1) samples m_t new positives and negatives in round t and trains an updated representation function f_t using initialization f_{t-1} by computing an approximation $\hat{f} \approx \min_{f\in\mathcal{F}} \mathcal{L}$ to the global minimizer f^*. We define the *reconstruction error* of $\hat{f}$ in terms of erroneous class labels: $\hat{f}$ is *α-accurate*, if for $z \sim p_z$, $z \in C_i$ implies $\tilde{z} = (g \circ f)(z) \in \tilde{C}_i$ with probability $1 - \alpha$. We want to sample positives that uncover weaknesses of $\hat{f}$; Fig. 2 illustrates our motivation: due to inaccuracies in the representation function, positive pairs may be correctly labeled as being in the same class in $\mathcal{Z}$, but not in $\tilde{\mathcal{Z}}$. For example, $x^+ \in \mathcal{X}$ generated from latents that are close to the decision boundary of a classifier in $\mathcal{Z}$ (marked in red) could help uncover inaccuracies in $\hat{f}$, if $\hat{f}(x^+)$ is misclassified in $\tilde{\mathcal{Z}}$. We formalize this observation below and propose two sampling strategies for generating "informative" positives.

A. Sampling strategies for positive pairs

1) Random augmentation: In practise, positives are often sampled via random augmentation, with augmentation method highly dependent on the application domain: for images, the anchor might be rotated or cropped to generate positives. In video analysis, adjacent frames may be chosen as positives. In text, positives may be adjacent sentences or generated by dropping different subsets of words from the same sequence. Our baseline emulates these classical augmentation techniques in an abstract setting: we sample an anchor x, generated from a randomly selected latent class; positives are generated via random perturbation of the anchor. We evaluate the quality of our representation function by testing its ability to recover the latent classes by learning a classifier $q : \tilde{\mathcal{Z}} \to \mathbb{R}^{|\mathcal{C}|}$, $q(f(x)) := Wf(x) = W(f \circ g)(z)$, in the reconstructed latent space. We formalize our earlier heuristic description of the class structure: let $\gamma > 0$ denote a margin, such that there exist $w_1, \dots, w_{|\mathcal{C}|} \in \mathbb{R}^d$ with $\sum_{i=1}^{|\mathcal{C}|} \|w_i\|^2 < 1$, such that for latents $z \sim p_c$, we have $\langle z, w_c \rangle \geq \frac{\gamma}{2}$ and $\langle z, w_{c'} \rangle \leq -\frac{\gamma}{2}$ for all $c \neq c'$, i.e., the latent classes $\mathcal{C}$ are separable. Given access to observations $x \in \mathcal{X}$, generated by g, and a contrastively learned representation function f, we can learn an "optimal" matrix $W \in \mathbb{R}^{|\mathcal{C}|\times d}$ to obtain the classifier q in $\tilde{\mathcal{Z}}$. Here, we have the following guarantee for the representation function:

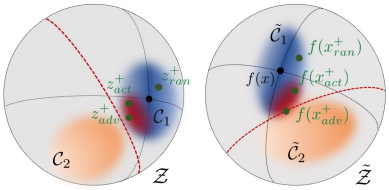

Fig. 2: **Informative positives.** Positives for an anchor (black), sampled actively (x^+_{act}), adversarial (x^+_{adv}), or via random augmentation (x^+_{ran}). The image of active instances $f(x^+_{act})$ lies close to the decision boundary of a classifier q_t (acceptance region in red); that of adversarial instances $f(x^+_{adv})$ is closest to the decision boundary or transgresses it if the reconstructed classes are not separable due to representation error.

Algorithm 1 Active selection of positive pairs

Learn initial $f_0 \in \mathcal{F}$ and classifier q_0 based on $\mathcal{D}_0$.
for $t = 0, 1, \dots, T-1$ **do**
  for $i = 0, \dots, m_t - 1$ **do**
    Sample anchor $c \sim p_c$, $x_i \sim p_{data}(\cdot|c)$.
    (i) **Random:** Sample $x_i^+ \sim p_{pos}(\cdot|x_i, c)$.
    (ii) **Active:** Sample $x_i^+ \sim p_{pos}(\cdot|x_i, c)$.
      If $\left|w_t \cdot f(x_i^+)\right| \geq a_t$: **Continue**.
    (iii) **Adversarial:**
      Set $x_i^+ \leftarrow \operatorname{argmax}_{x\in\mathcal{X}} l\left(\{q_t(f_t(x)) - q_t(f_t(x_i))\}\right)$.
  end for
  $\mathcal{D}_{t+1} := \{(x_i, x_i^+; x_i^- \sim p_{data})\}_{j=1}^{m_t} \cup \mathcal{D}_t$
  Learn $f_{t+1} \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}(f)$ over $\mathcal{D}_{t+1}$, with initialization f_t. Learn q_{t+1}.
end for

Theorem III.1. *Let $\gamma > 0$ (fixed) denote the margin in $\mathcal{Z}$ and $\tilde{\gamma} \leq \gamma$ the margin in $\tilde{\mathcal{Z}}$. For any $\delta > 0$ we have with probability at least $1 - \delta$ that for all $f \in \mathcal{F}$ ($\rho' := 1/(1-\rho)$)*

$$L_{class}(f) \leq \rho' \left(\hat{\mathcal{L}}_{un}(f) - \rho\right) + \rho' \left(4 L_\alpha \mathcal{R}_{\mathcal{D}}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2|\mathcal{D}|}}\right)$$

$\mathcal{R}_{\mathcal{D}}(\mathcal{F})$ denotes the Rademacher complexity of the function class $\mathcal{F}$, $L_\alpha \leq \frac{2}{\tilde{\gamma}}(1 - 2\alpha)$, and ρ the probability of sampling twice from the same class (i.e., a "false" negative instance).

The theorem is an adaption of [Saunshi et al., 2019, Thm. 4.1] to our latent variable model (see Apx. C). Positives (x, x^+) sampled via random augmentation will differ in their effectiveness, with some being more informative about the underlying latent class structure than others: consider a set $\mathcal{D}_0 \subseteq \mathcal{Z}$, which contains samples from two latent classes that are linearly separable with margin γ. Let f_t be a learnt, α-accurate representation function. Due to representation error, $h(\mathcal{D}_0) \subseteq \tilde{\mathcal{Z}}$ will be separable with a smaller margin $\tilde{\gamma} < \gamma$ only. Hence, instances that lie close to the decision boundary will be more challenging to classify in that they preserve the class label in $\mathcal{Z}$, but not in $\tilde{\mathcal{Z}}$. Such positives are particularly informative for learning good representations. Can we encourage the sampling of such positive instances?

2) Active selection of positive instances: Alternatively, we learn a classifier q_{t-1} in $\tilde{\mathcal{Z}}$ in round t, which is consistent with the representations of instances in $\mathcal{D}_{t-1}$. Then we sample a new anchor x and generate x^+ via random augmentation. But in contrast to the baseline approach, we accept only positives that are close to the decision boundary of q_{t-1}. Positives that are far away from the decision boundary are assumed to be less informative and will be rejected (Alg. 1, l. 6). In practise, the error of the representation function f_t (i.e., α) and consequently the margin $\tilde{\gamma}$ in $\tilde{\mathcal{Z}}$ are unknown to us. However, we know that the desired positives lie near the decision boundary of our classifier q_t, so we can encourage our algorithm to sample from this region (Fig. 2, marked red). Below, we present experimental evidence for the advantage of this strategy (Tab. I). We also analyze this advantage theoretically: we pick two classes and learn a separator between them. This reduces the problem to a binary classification task (an extension to the multi-class case follows from analogous arguments). Formally, let $|\mathcal{C}| = 2$ and $C_\pm := \{z | (z, \pm 1) \in \mathbb{S}^{k-1} \times \mathcal{C}\} \subseteq \mathcal{Z}$ and assume that C_+, C_- are linearly separable with margin γ. We want to learn a classifier in $\tilde{\mathcal{Z}}$ that recovers the latent class structure, i.e., that separates $\{\tilde{C}_+, \tilde{C}_-\} \subseteq \tilde{\mathcal{Z}}$. Note that *active selection* resembles classical active learning techniques [Balcan et al., 2007]. To compare the efficiency of the active strategy and the baseline, we compare the number of samples (m_t) that need to be drawn in each round to ensure convergence:

Theorem III.2. *For any $\delta, \epsilon > 0$, we can recover latent classes up to error ϵ with probability $1 - \delta$ with sample complexity (1) $m = O(\frac{d}{\epsilon})$ for random augmentations (Alg. 1(i)), and (2) $m = O\left(d^{3/2} \log(\frac{1}{\epsilon})\right)$ for active selection (Alg. 1(ii), with rejection threshold $a_t = \frac{\pi}{2^{t-1}}$).*

We see that the active approach allows for a much reduced sample size, indicating that representation functions can be trained more efficiently via active selection of positives. We defer all proof details to the Apx. C.

3) Adversarial augmentation: We may also generate positive instances via *adversarial* augmentation by selecting positives that are closest to the decision boundary or that are the most margin-transgressing. In round t, we locally maximize the loss with respect to a classifier q_t learnt based on representations f_t (Alg. 1(iii)), i.e., we solve $\operatorname{argmax}_{x \in \mathcal{X}} \; l\left(\{q_t(f_t(x)) - q_t(f_t(x_i))\}\right)$ with a suitable loss function l. This results in targeted perturbations of the anchor. While this approach may further reduce the number of labels m that need to be add in each iteration (see experiments below), solving this optimization problem for each positive in each iteration is computationally challenging.

Parameters			Baseline	Active	Double-active
n_{lat}	n_{obs}	N			
48	256	200	0.937	**0.951**	0.945
48	512	200	0.946	**0.955**	**0.955**
64	512	200	0.907	**0.916**	0.901
64	256	500	0.903	0.907	**0.927**
64	1024	500	0.922	0.920	**0.927**

TABLE I: Highest classification accuracy after N updates for Alg. 1 (two classes, n_{lat}/n_{obs} latent/ observed dimensions).

B. Experiments

In each iteration t, positives and negatives are sampled for a set of anchors and then added to the training set $\mathcal{D}_t$, which we use to train a representation function f_t using the standard contrastive loss above. Representations are evaluated in terms of their ability to recover latent classes by learning a classifier q_t in $\tilde{\mathcal{Z}}$ and reporting its accuracy. All data is synthetic and sampled from the unit hypersphere. To define class-conditioned distributions in the true latent space, we define a class mean on the unit hypersphere, add isotropic Gaussian noise with a predefined standard deviation and renormalize to unit length. We investigate four scenarios, which employ different sampling strategies (see Apx. D): (1) anchors and positives are sampled from the baseline prior (**baseline**, resembling Alg. 1(i)); (2) anchors and candidate positives sampled from baseline, positives accepted only if close to decision boundary (**active**, Alg. 1(ii)); (3) anchors sampled with a bias towards proximity to the decision boundary of q_{t-1}, candidate positive sampled via perturbation in the observation space and accepted, if their image lies near the decision boundary (**adversarial**, closer to Alg. 1(iii)). The results are given in Table I. We see that the active strategies consistently outperform the baseline. This corroborates Thm. A.7, which shows that active strategies can converge at a faster rate.

IV. Sampling positives via targeted augmentations

In our second setting, we sample positives via augmentation in the true latent space rather than the observation or reconstructed latent space. E.g., imagine an agent interacting with a visual scene where some latents describe useful structure in the scene's appearance (e.g., type, location of objects), while others describe more superficial structure (e.g., surface texture, shading of objects). Here, structure in the former latents is more useful for downstream tasks.

A. Sampling strategies

Our interventions on the data generation process (e.g., an agent interacts with an object in the scene) are analogous to data augmentations in our earlier settings, where the values of some latent variables are fixed and others are perturbed. In practise, useful structure in the observations does not depend equally

Algorithm 2 Active perturbation in latent space

Initialize $f_0 \in \mathcal{F}$.
Define index set $I \subset [k]$ of informative latent variables.
for $t = 0, 1, \ldots, T-1$ **do**
  for $i = 0, \ldots, m-1$ **do**
    Sample anchor $c \sim p_c$ and $z, \bar{z} \sim p_{data}(\cdot|c)$.
    Calibrate τ.
    Perturb informative latents: $z_I^+ \leftarrow p_\tau(\cdot|z_I)$
    Randomize other latents: $z_{\neg I}^+ \leftarrow \bar{z}_I$
    $(z_i, z_i^+) \leftarrow (z, z^+)$
  end for
  $\mathcal{D}_t := \{(x_i, x_i^+; x_i^- \sim p_{data})\}_{j=1}^m$
   with $x_i = g(z_i), x_i^+ = g(z_i^+)$.
  Learn $f^{t+1} \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}(f)$, with initialization f^t.
end for

on all latents, but on a particularly *informative* subset. Here, it can be helpful to sample positives via perturbations that focus on informative latents, rather than affecting all latents uniformly. Scenarios with an informative subset of latents arise in Reinforcement Learning: suppose an agent receives complex observations, which are functions of a large number of latents, but can only influence a subset of them through its actions. Characterizing variability in the controllable subset is critical for choosing optimal actions, while other latents evolve independently of the agent's actions and are thus less important for planning. In our framework, latents influenced by the agent's actions are *informative* with respect to a control task, whereas the remaining latents are *uninformative*. While we may not know what the latents encode and which are informative, we know that the agent can only manipulate the informative latents. Hence, we assume that the agent performs targeted perturbations of informative latents. We investigate sampling strategies for positive pairs $p(z^+|z)$ (anchor $z \sim p(z)$) and evaluate their effectiveness with respect to the encoder's ability to learn a posterior distribution $p(\tilde{z}|x, c) =: q_h(z|x, c)$, such that structure in $\mathcal{Z}$ (encoded in $p(z|x, c)$) is recovered in $\tilde{\mathcal{Z}}$:

Theorem IV.1. *Consider the minimizer* $h^* = \operatorname{argmin}_{(z,z^+)} \mathbb{E}\left[H(p(z^+|z, c), q_h(z^+|z, c))\right]$ *with* $q_h(z^+|z, c) = C_h(z^+)^{-1} e^{h(z^+)^T h(z)/\tau}$ *denoting the conditional distributions (where* $C_h(z) := \int e^{h(z^+)^T h(z)/\tau} dz$*) over reconstructed latents and* $H(p, q_h)$ *denoting the cross entropy between distributions* p *and* q_h*. Then* h^* *locally reconstructs the latent space up to linear and orthogonal transformation.*

The proof adapts a result by Zimmermann et al. [2021] (see Apx. B). Let an index set $I \subset [k]$ characterize the informative latents (k: dimension of the latent space). Alg. 2 applies targeted perturbations to the informative latents and randomize the others (*active perturbation*).

B. Experiments

Class-conditioned distributions in the true latent space are obtained by concatenating samples from the surface of two unit hyperspheres (of dimensions n_{info}, representing informative latents, and n_{noise} for uninformative latents). To sample informative latents, we define a class mean on the first hypersphere, add isotropic Gaussian noise with a predefined standard deviation and then renormalize to unit length. With that, the distribution of informative latents conditioned on a class resembles a von Mises-Fisher distribution (Eq.(1.1.) in Apx. B). The standard deviation is calibrated to generate little overlap between the classes in order to simulate separability. The uninformative latents are sampled uniformly from the second hypersphere, the scale can be varied to control the degree to which variability in the observation space is due to the informative or uninformative latents. The effect of the latents on the observations is entangled and non-linear due to the function $g(z)$, so identifying informative vs. uninformative variability via observations $x = g(z)$ is non-trivial. We investigate four strategies for sampling positives given an anchor (details in Apx. E): (1) equal-sized random perturbations of all latents (**baseline**); (2) small, targeted perturbations of informative latents and randomization of others (**info-active**); (3) small, targeted perturbations of informative latents and larger, targeted perturbations of others (**active**); (4) generate positives from the same class, assuming access to class-preserving transformation (**control**). Our results confirm that targeted perturbation of the informative latents generates more effective positives than the baseline; even if the difference in the perturbation scales of informative and uninformative latents is small (Tab. II, see also Apx. E). A strong advantage of Alg. 2 over Alg. 1 is that it *does not require prior knowledge of classes or class structure*.

Parameters			Baseline	Info-active	Active	Control
n_c	n_i	n_n				
4	4	4	0.777	**0.956**	0.936	0.986
8	4	4	0.609	**0.867**	0.817	0.900
4	2	6	0.728	**0.967**	0.942	0.974
8	2	6	0.408	**0.692**	0.611	0.687
4	2	16	0.759	**0.969**	0.949	0.975

TABLE II: kNN accuracy of Alg. 2 averaged over 10 resamples of the train/ test sets after 1000 updates, with n_c classes, n_i informative latents and n_n noisy latents).

V. Conclusions

This paper studied contrastive learning approaches for recovering low-dimensional structure in a latent variable model. We proposed strategies for generating effective positives, which we analyzed both theoretically and in exploratory experiments. Our results confirm the intuition that encouraging the sampling of positives that are *informative* with respect to the underlying class structure or that explicitly uncover weaknesses in the representation function can improve the efficiency of the contrastive approach. We investigate sampling strategies in an abstract setting modeled after commonly used approaches: our *random augmentation* strategy is an abstraction of approaches used in image and video analysis, which rely on cropping or blurring [Chen et al., 2020]. Future work could explore our active sampling strategies in applications (e.g., Computer Vision, NLP), which would allow to evaluate their efficiency in a more practical setting. We also investigate a setting where salient structure in the observations is driven by a subset of *informative* latents, motivated by Reinforcement Learning, where an agent's actions only influence a subset of latents. Furture work includes testing our approach on such a task.

REFERENCES

M. Anthony, L.M.M. Anthony, P.L. Bartlett, P.L. Bartlett, and Cambridge University Press. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 7187–7209, 28–30 Mar 2022.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32*, pages 15535–15545. 2019.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

Alex Fedorov, Lei Wu, Tristan Sylvain, Margaux Luck, Thomas P. DeRamus, Dmitry Bleklov, Sergey M. Plis, and Vince D. Calhoun. On self-supervised multimodal representation learning: An application to alzheimer's disease. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1548–1552, 2021.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv:2104.08821*, 2021.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Tyler L Hayes, Maximilian Nickel, Christopher Kanan, Ludovic Denoyer, and Arthur Szlam. Can i see an example? active learning the long tail of attributes and relations. *arXiv preprint arXiv:2203.06215*, 2022.

Chih-Hui Ho and Nuno Nvasconcelos. Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, 33:17081–17093, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2021.

Yu Meng, Chenyan Xiong, Payal Bajaj, saurabh tiwary, Paul N. Bennett, Jiawei Han, and Xia Song. COCO-LM: Correcting and contrasting text sequences for language model pretraining. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 1196–1204, 2013.

Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9587, 2021.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

Li S. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4, 01 2011.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, 09–15 Jun 2019.

Anshul Shah, Suvrit Sra, Rama Chellappa, and Anoop Cherian. Max-margin contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8220–8230, 2022.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

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

Feng Wang, Huaping Liu, Di Guo, and Sun Fuchun. Unsupervised representation learning by invariance propagation.

In *Advances in Neural Information Processing Systems*, volume 33, pages 3510–3520, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv:2203.13457*, 2022.

Seunghan Yang, Debasmit Das, Simyung Chang, Sungrack Yun, and Fatih Porikli. Distribution estimation to automate transformation policies for self-supervision. *arXiv:2111.12265*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10042–10051, October 2021.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.