Intriguing Properties of Contrastive Losses

Ting Chen  
Google Research  
iamtingchen@google.com

Calvin Luo  
Google Research  
calvinluo@google.com

Lala Li  
Google Research  
lala@google.com

Abstract

We study three intriguing properties of contrastive learning. First, we generalize the standard contrastive loss to a broader family of losses, and find that various instantiations of the generalized loss perform similarly under the presence of a multi-layer non-linear projection head. Second, we study if instance-based contrastive learning (with a global image representation) can learn well on images with multiple objects present. We find that meaningful hierarchical local features can be learned despite the fact that these objectives operate on global instance-level features. Finally, we study the phenomenon of feature suppression among competing features shared across augmented views, such as “color distribution” vs “object class”. We construct datasets with explicit and controllable competing features and show that, for contrastive learning, a few bits of easy-to-learn shared features can suppress, and even fully prevent, the learning of other sets of competing features. In scenarios where there are multiple objects in an image, the dominant object would suppress the learning of smaller objects. Existing contrastive learning methods critically rely on data augmentation to favor certain sets of features over others, and could suffer from learning saturation for scenarios where existing augmentations cannot fully address the feature suppression. This poses open challenges to existing contrastive learning techniques.

1 Introduction

Contrastive learning [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14] has recently achieved great successes in learning visual representations without supervision. As shown in [13, 14], contrastive learning can learn representations that rival supervised learning, and significantly improve the state-of-the-art in semi-supervised learning on ImageNet. One successful use case of contrastive loss for self-supervised learning is to make augmented views of the same example agree [1, 2, 13]. A widely used contrastive loss to encourage agreement is based on cross entropy [15, 3, 4, 13]. Given an augmented view of an example, the contrastive prediction task aims to classify a set of candidates into the positive example (i.e. the other augmented view of the same example) and negative ones via the cross entropy loss.

In this work, to understand the effectiveness and limitation of existing contrastive learning methods, we study three intriguing aspects. First, we propose a generalization of the standard contrastive loss, and systematically study their performance differences. Second, we study if the instance-based contrastive learning, for which the contrastive loss operates on global representation of an input image, can learn well on images with multiple objects present, and whether or not it leads to meaningful local features. Finally, we systematically study the feature suppression phenomenon in contrastive learning. The suppression effect occurs among competing features shared across augmented views. For example, with random cropping as the augmentation, “color distribution” and “object class” are often competing features as they are likely shared between two augmented views. The suppression effect among competing features can significantly degenerate the representation quality, or even


completely disable the learning of certain features, as shown in our experiments. Existing methods critically rely on hand-crafted data augmentation to favor certain sets of competing features than others.

Our main findings and contributions are summarized below.

- We propose a generalized contrastive loss, and show that differences between contrastive losses are small with a deep projection head.
- We show that the instance-based objective widely used in existing contrastive learning methods can learn on images with multiple objects, and also learn meaningful local features despite operating on global image representation.
- We construct three datasets with explicit and controllable competing features to systematically study the feature suppression effect in contrastive learning.
- We show that a few bits of easy-to-learn shared features can suppress, and even fully prevent, the learning of other sets of competing features. In scenarios where there are multiple objects in an image, the dominant object would suppress the learning of smaller objects. This poses open challenges to existing contrastive learning.

2 Generalized contrastive loss and differences among its instantiations

The common contrastive loss used in most recent work is based on cross entropy \([15, 3, 4]\). Following the notation in \([13]\), the contrastive loss can be defined between two augmented views \((i, j)\) of the same example for a mini-batch of size of \(n\), and can be written as the following.

\[
\mathcal{L}_{\text{NT-Xent}} = - \frac{1}{n} \sum_{i,j \in \mathcal{MB}} \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2n} \mathbb{1}[k \neq i] \exp(\text{sim}(z_i, z_k)/\tau)}
\]

where \(z_i, z_j\) are hidden representations of two augmented views of the same example; \(\text{sim}(u, v) = u^T v / (\|u\| \cdot \|v\|)\) is the cosine similarity between two vectors; \(\tau\) is a temperature scalar and \(\mathcal{MB}\) is a randomly sampled mini-batch consisting of augmented pairs of images. In \([13]\), a MLP projection head is introduced between intermediate layer \(h\) (e.g. output of ResNet encoder) and final output \(z\). It is shown that the projection head is very beneficial and \(h\) is a much better feature representation than \(z\).

In this work, we generalize the standard contrastive loss to the following form.

\[
\mathcal{L}_{\text{generalized contrastive}} = \mathcal{L}_{\text{alignment}} + \lambda \mathcal{L}_{\text{distribution}}
\]

Both terms are defined on hidden representations. \(\mathcal{L}_{\text{alignment}}\) encourages representations of augmented views to be consistent, while \(\mathcal{L}_{\text{distribution}}\) encourages representations (or a random subset of them) to match a prior distribution (of high entropy). It is not difficult to see that the standard contrastive loss in Eq. 1 is a special case as it can be re-written as follows (scaled by a constant \(\tau\)).

\[
\tau \mathcal{L}_{\text{NT-Xent}} = - \frac{1}{n} \sum_{i,j} \text{sim}(z_i, z_j) + \frac{\tau}{n} \sum_{i} \log \sum_{k=1}^{2n} \mathbb{1}[k \neq i] \exp(\text{sim}(z_i, z_k)/\tau)
\]

This form of factorization in Eq. 3 has been proposed in \([16]\), where the second LogSumExp term is referred to as uniformity since it encourages representation to uniformly distributed in the hypersphere. Different from \([16]\), here we generalize the hypersphere uniform distribution and study a wider set of prior distributions for their effectiveness in learning representations.

**SWD for supporting diverse prior distributions.** One issue of using more diverse set of priors is we cannot rely on LogSumExp for matching the distribution. To this end, we resort to the theory of optimal transport, via Sliced Wasserstein Distance (SWD) \([17, 18, 19]\). For two sets of equal-sized samples from two 1-D distributions, the optimal transport can be obtained by computing two
Comparing different instantiations of generalized contrastive loss. Here we ask: Is it essential to use a uniform hypersphere prior for the effectiveness of contrastive loss? How much difference does it make when distinct generalized contrastive losses are used? To answer this question, we conduct experiments following SimCLR settings [13, 14], and use the linear evaluation protocol. Detailed experimental setup can be found in Appendix A.1.

Figure 1 shows linear evaluation results of models trained with different losses under different training epochs. On CIFAR-10, we see little difference in terms of linear evaluation for variants of the generalized contrastive losses, especially when trained longer than 200 epochs. As for ImageNet, there are some discrepancies between different losses, but they disappear when a deeper 3-layer non-linear projection head is used.
Furthermore, we find that deep projection head not only reduces the differences among different generalized contrastive losses, but has a similar effect for batch size. With proper learning rate scaling across batch sizes (e.g. square root scaling with LARS optimizer [21]), the impact of batch size on representation quality is small. Table 2 demonstrate this phenomenon for the standard contrastive loss, and more results on other losses can be found in Appendix A.3.

<table>
<thead>
<tr>
<th>Projection head</th>
<th>Batch size</th>
<th>Epoch 100</th>
<th>Epoch 200</th>
<th>Epoch 400</th>
<th>Epoch 800</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 layers</td>
<td>512</td>
<td>65.4</td>
<td>67.3</td>
<td>68.7</td>
<td>69.3</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>65.6</td>
<td>67.6</td>
<td>68.8</td>
<td>69.8</td>
</tr>
<tr>
<td></td>
<td>2048</td>
<td>65.3</td>
<td>67.6</td>
<td>69.0</td>
<td>70.1</td>
</tr>
<tr>
<td>3 layers</td>
<td>512</td>
<td>66.6</td>
<td>68.4</td>
<td>70.0</td>
<td>71.0</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>66.8</td>
<td>68.9</td>
<td>70.1</td>
<td>70.9</td>
</tr>
<tr>
<td></td>
<td>2048</td>
<td>66.8</td>
<td>69.1</td>
<td>70.4</td>
<td>71.3</td>
</tr>
<tr>
<td>4 layers</td>
<td>512</td>
<td>66.8</td>
<td>68.8</td>
<td>70.0</td>
<td>70.7</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>67.0</td>
<td>69.0</td>
<td>70.4</td>
<td>70.9</td>
</tr>
<tr>
<td></td>
<td>2048</td>
<td>67.0</td>
<td>69.3</td>
<td>70.4</td>
<td>71.3</td>
</tr>
</tbody>
</table>

3 Instance-based objective can learn on images with multiple objects and learn good local features

Most existing contrastive learning methods [13, 10, 22, 4] define their objectives at the instance level where each image is encoded into a single vector representation (e.g. representations of two random crops of the same image instance are treated as a positive pair). In other words, the objective operates on a global representation of its input rather than on some local regions (of its input). We pose two questions regarding instance-based global objective: 1) when there is only a single (dominant) object in the image, the objective seems reasonable as it encourages the model to learn features relevant to object class, but when there are multiple objects present in the image, can instance-based objective still learn well? 2) Since the instance-based objective uses a global summary of its input, can it still learn good local features (e.g. parts of an object, or multiple objects in the same scheme)? To answer these questions, we use SimCLR as representative for the instance-based objective.

3.1 SimCLR can learn on images with multiple objects

Commonly used self-supervised learning datasets, such as MNIST, CIFAR-10, ImageNet, are object centered, i.e. the image is mainly occupied by a single (dominant) object. To experiment with multiple objects in a controllable setting, we propose a new dataset setting by composing multiple digits as follows.

**MultiDigits dataset.** We place MNIST digits (28 × 28 size) on a shared canvas (112 × 112 size). We vary the number of digits placed on the canvas. One factor that could interfere with learning of multiple digits is overlapping digits, therefore we use two placement strategies: random vs in-grid (Figure 2). Random placement of digits incurs no constraint on where digits can be placed on the
canvas, whereas in-grid placement puts each digit in one of the $4 \times 4$ grid cells the canvas is divided into, and no two digits can fall in the same cell. In-grid placement ensures no overlapping of digits.

We first pretrain a ResNet-18 with SimCLR or supervised learning with the same augmentation policy (random cropping and resize) on MultiDigits dataset. To access the representation quality, we then train linear classifiers for images with a single digit of size $28 \times 28$ on the canvas. Similarly during evaluation, we place only one digit of size $28 \times 28$ on the canvas.

As shown in Table 3, representations learned using supervised loss maintains its quality when up to 8 digits are placed in the image. After that the representation becomes worse as the canvas gets more crowded. Notably, representations learned using SimCLR display a similar phenomenon. Regardless of placement strategy, top-1 accuracy stays at the same level up to 8 digits, demonstrating that SimCLR can learn from images with multiple objects. In addition, the increased performance gap between the two placement strategies with increased number of digits shows that object overlapping makes it harder for contrastive losses to learn from multiple objects.

Table 3: Top-1 linear evaluation accuracy (%) for pretrained ResNet-18 on the MultiDigits dataset. We vary the number of digits placed on the canvas during training from 1 to 16. During evaluation only 1 digit is present. As a baseline, a network with random weights gives 18% top-1 accuracy.

<table>
<thead>
<tr>
<th>Placing of digits</th>
<th>Number of digits (size $28 \times 28$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Supervised</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>99.5</td>
</tr>
<tr>
<td>In-grid</td>
<td>99.5</td>
</tr>
<tr>
<td>SimCLR</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>98.9</td>
</tr>
<tr>
<td>In-grid</td>
<td>98.3</td>
</tr>
</tbody>
</table>

### 3.2 SimCLR learns local features that exhibit hierarchical properties

To understand the local features learned by SimCLR, we apply K-means on intermediate features of the pretrained ResNet with SimCLR, and see how local regions of an image are grouped together. For good representations, we expect that regions of similar objects or object parts should be grouped together.

Specifically, we take a pretrained Resnet-50 $2 \times$ on ImageNet, and run inference on images (from ImageNet validation set and COCO [23]) of size $448 \times 448$. We run K-means with various numbers of clusters on the l2-normalized hidden features from middle layers of the network (e.g. block group 2,3,4 of the ResNet). We also compare SimCLR learned features with supervised learned features, as well as the raw pixel (RGB) features extracted from each $14 \times 14$ patch.

Figure 3a shows that as the number of clusters increases, the learned representations tend to group image regions based on parts of the object (i.e. facial components of the dog). This phenomenon appears in both SimCLR and supervised learned features, but not with raw pixel features, indicating meaningful local features learned by SimCLR and supervised learning. In Figure 3b, we compare ResNet intermediate features at different layers, and it suggests that earlier layers contain more edge-related features, while later layers contain more object/part features.
4 Feature suppression limits the potential of contrastive learning

Contrastive learning requires good design of data augmentation to work well. As shown in [13], without color augmentations that randomly shift color distribution (while maintaining information regarding object class), the quality of learned representations are significantly worse. In other words, the presence of “color distribution” features suppresses their competing feature of “object class”, and is addressed by color augmentation. However, there may be scenarios where the known augmentations cannot fully address this feature suppression effect, and it can thus limit the potential of contrastive learning. Here we quantitatively study the feature suppression phenomenon by constructing datasets with explicit and controllable competing features, and see how well contrastive learning method could learn.
4.1 Datasets with explicit and controllable competing features

To construct datasets with controllable competing features, we leverage two strategies: *channel addition* that adds different feature information in a shared canvas, and *channel concatenation* that expand the RGB channels to include additional features. With these strategies, we construct three datasets below.

**DigitOnImageNet dataset.** We overlay MNIST digits on ImageNet images via channel addition/summation (Figure 5a). For each ImageNet image, we assign a unique MNIST digit and replicate it in nine fixed locations before the standard SimCLR augmentations [13] are applied to create augmented views. Therefore the original ImageNet images and added MNIST digits are competing features. Although it is difficult to quantify information in MNIST digits, we can manually control the number of unique MNIST digits used. Ideally, we want the model to learn both set of features so that it could perform well for both MNIST digit and ImageNet object recognition.

**MultiDigits dataset (varying the size of one digit).** This dataset is modified from MultiDigits introduced above. Here we only consider two digits, and vary the size of one of them (Figure 5b). In this work, we place two digits on a canvas of size $112 \times 112$. We fix the size of one of the digits to be $20 \times 20$ while varying the other from $20 \times 20$ to $80 \times 80$. Digits of different sizes can be considered as competing features. Ideally, we want the model to learn features for digits of all sizes appeared during training.

**RandBit dataset.** We concatenate a real image with an image of a random integer in the channel dimension (Figure 5c). The random integer is randomly sampled from range of $[1, \log_2(n)]$. The integer, shared between two views, is replicated for spatial dimension and represented as $n$ binary channels. RGB channels and random bits are competing features.

Figure 5: Probing datasets with explicit and controllable competing features.
number to make it easily learnable. Furthermore, unlike RGB channels, these additional channels of random bits will not be altered by augmentation, so they are identical for both augmented views of the same image. The RGB channels and the added channels of random bits are competing features, and this construction allows us to control the amount of information in the added competing feature, which is \( n \) bits. Also, we know that the mutual information between two views given this construction is at least \( \log_2(n) \).

### 4.2 Easy-to-learn features (MNIST digit) suppress the learning of other features (ImageNet object class)

![Graphs showing supervised and unsupervised learning](image)

Figure 6: (a) Supervised learning accuracy on ImageNet classification. (b) Linear evaluation of learned features for both MNIST classification and ImageNet classification on the DigitOnImageNet dataset. Batch size of 1024 and 2-layer projection head is used. Different batch sizes and projection head layers have negligible influence on the trade-off between ImageNet vs MNIST accuracy.

On DigitOnImageNet datasets, we vary the number of unique MNIST digits used in the training set, and all MNIST digits are used in the validation/test set. As a baseline, we train supervised ResNet-50 on the created datasets with ImageNet labels, and the number of unique MNIST digits has little impact on the top-1 ImageNet classification accuracy (Figure 6a).

We then train SimCLR on the datasets with different temperatures. As shown in Figure 6b, when we increase the number of unique MNIST digits, the linear evaluation performance of the learned features for MNIST classes increases accordingly, while the accuracy for ImageNet classes decreases dramatically. The trade-off between digit recognition ability and object recognition ability shows that simple features suppress the learning of difficult features, when both are shared between two augmented views. Different batch sizes and projection head depths have negligible influence to the outcome we observe here. Therefore, it is difficult to learn both of the competing features using existing contrastive losses (e.g. SimCLR).

### 4.3 The presence of dominant object suppresses the learning of features of smaller objects

On the MultiDigits dataset, as mentioned, we fix one digit to be size of \( 20 \times 20 \) while varying the other from \( 20 \times 20 \) to \( 80 \times 80 \), on a canvas of \( 112 \times 112 \). We first pretrain a ResNet-18 with SimCLR or supervised learning with the same augmentation policy (random cropping and resize) and batch size of 1024. To access the representation quality, we then train linear classifiers for each of the digit sizes that appeared during pretraining. For training of the linear classifier, we only place a single digit at a time on the canvas of the same size as during pretraining.

The results are summarized in Table 4. For supervised learning, the learned representations for the smaller digit do not change much as the other digit increases its size, and the model perform well for both small and large digits (accuracy \( > 99\% \)). However, for SimCLR, the learned representations of the smaller digit degenerate significantly when the size of the other digit increases, almost to the level of a random untrained network. The dominant object can be learned very well (accuracy \( > 99\% \)) while suppressing the learning of the smaller object. Although tuning temperature has some effects on reducing the feature suppression, the trend stays unchanged.
### 4.4 Extra channels with a few bits of easy-to-learn mutual information suppress the learning of all features in RGB channels

In the RandBit datasets, we add additional channels (identical across pixels) of random bits to MNIST and ImageNet. As mentioned above, SimCLR augmentation is only applied to RGB channels so extra added channels will be shared among two view.

Figure 7 shows the linear evaluation accuracy of models trained on MNIST (with additional random bits added). We observe that the linear evaluation accuracy quickly drops with a few bits of competing feature added. This detrimental effect on the representation quality persists on bigger datasets like CIFAR-10 and ImageNet as well, and cannot be avoided by using different contrastive losses, batch sizes, or memory mechanism based on momentum contrast (details in Appendix B.1). We believe the fact that just a few bits of easy-to-learn features can completely disable the good representation learning is related to the saturation of the distribution matching loss. As shown in Appendix B.2, the linear increase in bits requires an exponential increase in batch size, which is not sustainable as the required batch size can quickly go beyond the size of the dataset size. In practice, we rely on using data augmentation to remove those uninformative easy-to-learn features so that contrastive learning can learn useful representations. Interestingly, the extra bits do not affect a generative model, variational autoencoder [24, 25], nearly as much, despite other settings such as model size are held the same, prompting a potential direction of addressing the issue.

### 5 Related Work

Our work studies the contrastive loss based on cross entropy loss [15, 3, 4, 13]. This loss is widely used in recent successful contrastive learning methods [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14].
terms of the contrastive loss, our work is perhaps most related to [16], which shows that formulating contrastive loss as alignment and uniformity in the hypersphere gives similar performance as the standard contrastive loss. We further generalize this factorization, and show other distribution matching losses can be used, and they could achieve similar results. Other than standard contrastive loss that directly utilize negative examples, BYOL [22] demonstrates another way to maintain representation distribution/entropy without directly relying on distribution matching, and SWAV [26] shows clustering-based method equipped with proper data augmentations could also achieve similar performance. We conduct preliminary experiments of BYOL on RandBit and found that it also suffers from feature suppression as generalized contrastive loss. It is expected that SWAV would exhibit similar behaviors on RandBit as those random bits could fuel representations for perfect clustering.

The connection between contrastive loss and mutual information has been studied before [3, 20]. We show that for the generalized contrastive loss, it can also be related to mutual information. Despite the connection between contrastive loss and mutual information, it has been pointed out that mutual information estimation may suffer from certain limitations [27, 28]. Moreover, [29, 12] show that higher mutual information learned by the network does not warrant better representation quality. In our work, we find adding mutual information bits between two views which are irrelevant to downstream tasks can be harmful for the quality of learned representations. Data augmentation plays an important role at favoring certain bits of mutual information than others.

There is a growing number of recent work on the topic of understanding contrastive learning, both theoretically [30, 31, 32, 33, 34] and empirically [16, 12, 35, 36]. However, little work has been done to study the phenomenon of feature suppression. To our knowledge, we are the first one to quantitatively and systematically study this problem. We believe this is still a very open question and could benefit from more future investigation. Finally, the feature suppression effect in unsupervised contrastive learning that we study in this work may also exist in standard supervised learning (“contrastive loss” between examples and class labels), as suggested by [37, 38], though the specific form would be different.

6 Conclusion

In this work, we study three intriguing properties of contrastive losses. In particular, our results highlight that feature suppression is still an open challenge in contrastive learning. While there is a plethora of work on improving contrastive learning, few of them directly aim to address feature suppression. This limitation of contrastive learning becomes a bottleneck for scenarios where existing augmentation cannot fully address the feature suppression phenomenon, and learning would saturate at a level of dissatisfaction.

We would also like to point out some limitations of our study. Firstly, we focus mostly on contrastive learning with explicit negatives (e.g. SimCLR and MoCo). We believe other methods based on clustering and/or without negative pairs would exhibit similar phenomenon but we leave that as future work. Secondly, many of our proposed image datasets are not fully realistic despite being composed from some (challenging) natural image datasets such as ImageNet. We admit it is very hard to explore competing features or multiple objects in a controllable fashion on realistic large scale image datasets.

Acknowledgements

We specially thank Geoffrey Hinton for many inspiring discussions and helpful advice. We would also like to thank David Fleet, Simon Kornblith, Mohammad Norouzi, Kevin Swersky and Katherine Hermann for insightful discussions. In addition, we are thankful to William Chan and Sara Sabour for ideas on implementation of sorting on TPUs. We also thank the anonymous reviewers for their constructive feedback.
References


