


# Intriguing Properties of Contrastive Losses

Ting Chen, Calvin Luo, Lala Li

*Google Research, Brain Team*

(with special thanks to Geoffrey Hinton)

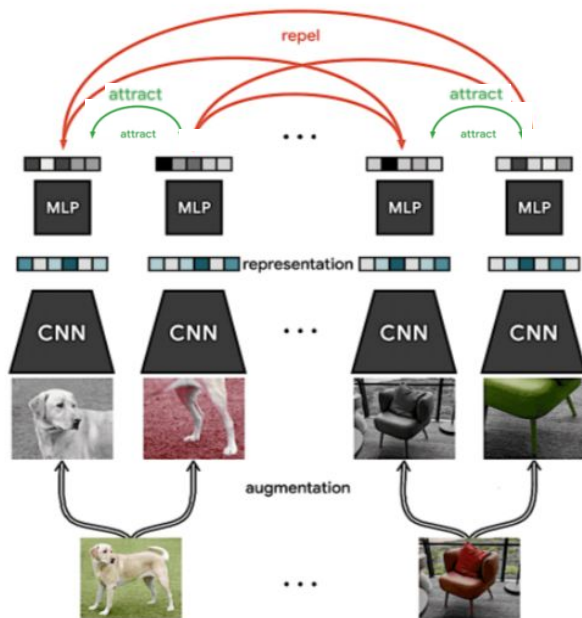




## Three properties studied in this work

1. Do different instantiations of the generalized contrastive loss perform differently?
2. Do instance-based contrastive learning methods learn on images with multiple objects and do they learn good local features?
3. Does feature suppression limit the contrastive learning?

# A common contrastive loss



SimCLR

Contrastive loss based on cross entropy:

$$\text{Let } \text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$$

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

# Generalized contrastive losses

An abstract 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
- $\mathcal{L}_{\text{distribution}}$  : encourages representations (or random subset) to match some prior distribution of high entropy (e.g. Gaussian)

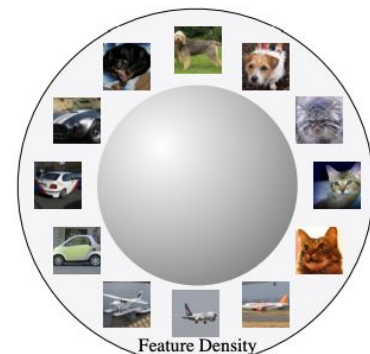
# Example: a common contrastive loss

Contrastive loss based on cross entropy loss with temperature:

$$\mathcal{L}^{\text{NT-Xent}} = -\frac{1}{n} \sum_{i,j \in \mathcal{MB}} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2n} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

By expanding the loss and scaling it by a constant of  $\tau$ :

$$\tau \mathcal{L}^{\text{NT-Xent}} = \underbrace{-\frac{1}{n} \sum_{i,j} \text{sim}(\mathbf{z}_i, \mathbf{z}_j)}_{\mathcal{L}_{\text{alignment}}} + \underbrace{\frac{\tau}{n} \sum_i \log \sum_{k=1}^{2n} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}_{\mathcal{L}_{\text{distribution}}}$$



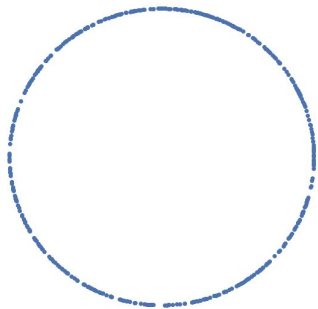
$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$$

[Wang & Isola, 2020]

**Uniformity:** Preserve maximal information

# What about other prior distributions?

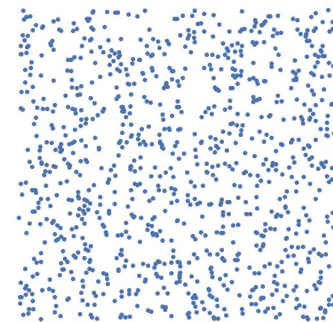
- Is uniform hypersphere prior (via logsumexp) really essential to the effectiveness of contrastive loss?
- Here we explore multiple potential prior distributions:



Uniform hypersphere



Normal / Gaussian



Uniform hypercube

- How do we make hidden vectors match these prior distributions?
  - **Sliced Wasserstein Distance (SWD)** as distribution matching loss

# A wider set of instantiations

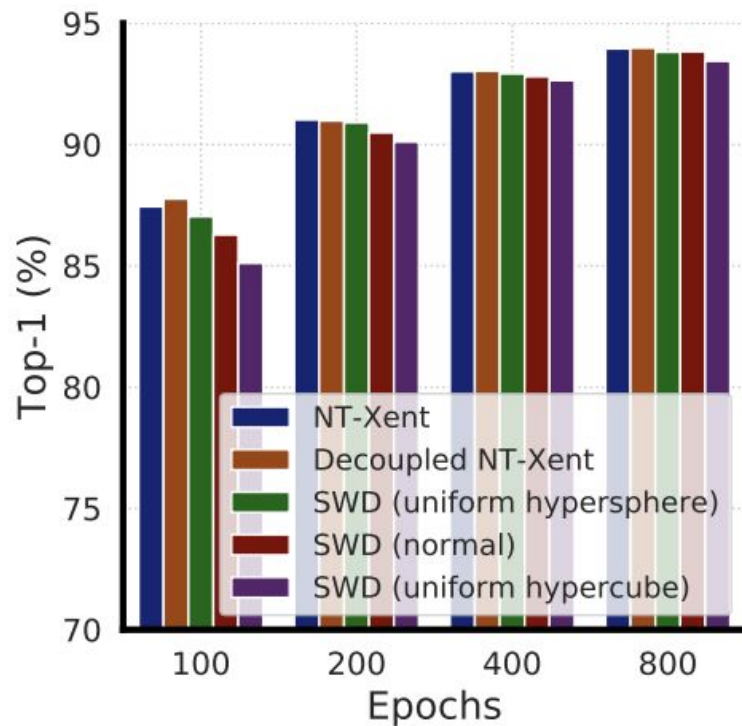
We instantiate generalized contrastive losses with different prior distributions and  $\mathcal{L}_{\text{distribution}}$ .

Table 1: Instantiations of the generalized contrastive loss, i.e.  $\mathcal{L}_{\text{alignment}} + \lambda \mathcal{L}_{\text{distribution}}$ , that we use in this work.  $\tilde{z}$  denotes  $\ell_2$ -normalized  $z \in \mathbb{R}^d$ , and is only used for uniform hypersphere prior.

$\mathcal{L}_{\text{align}}$	Prior distribution	$\mathcal{L}_{\text{distribution}}$
$\frac{1}{nd} \sum_{i,j} \ \tilde{z}_i - \tilde{z}_j\ ^2$	Uniform hypersphere	$\frac{1}{n} \sum_i \log \sum_j \exp(\tilde{z}_i^T \tilde{z}_j / \tau)$
$\frac{1}{nd} \sum_{i,j} \ \tilde{z}_i - \tilde{z}_j\ ^2$	Uniform hypersphere	SWD( $\tilde{Z}, Z^{\text{prior}}$ )
$\frac{1}{nd} \sum_{i,j} \ z_i - z_j\ ^2$	Uniform hypercube	SWD( $Z, Z^{\text{prior}}$ )
$\frac{1}{nd} \sum_{i,j} \ z_i - z_j\ ^2$	Normal distribution	SWD( $Z, Z^{\text{prior}}$ )

# Different generalized contrastive losses perform similarly

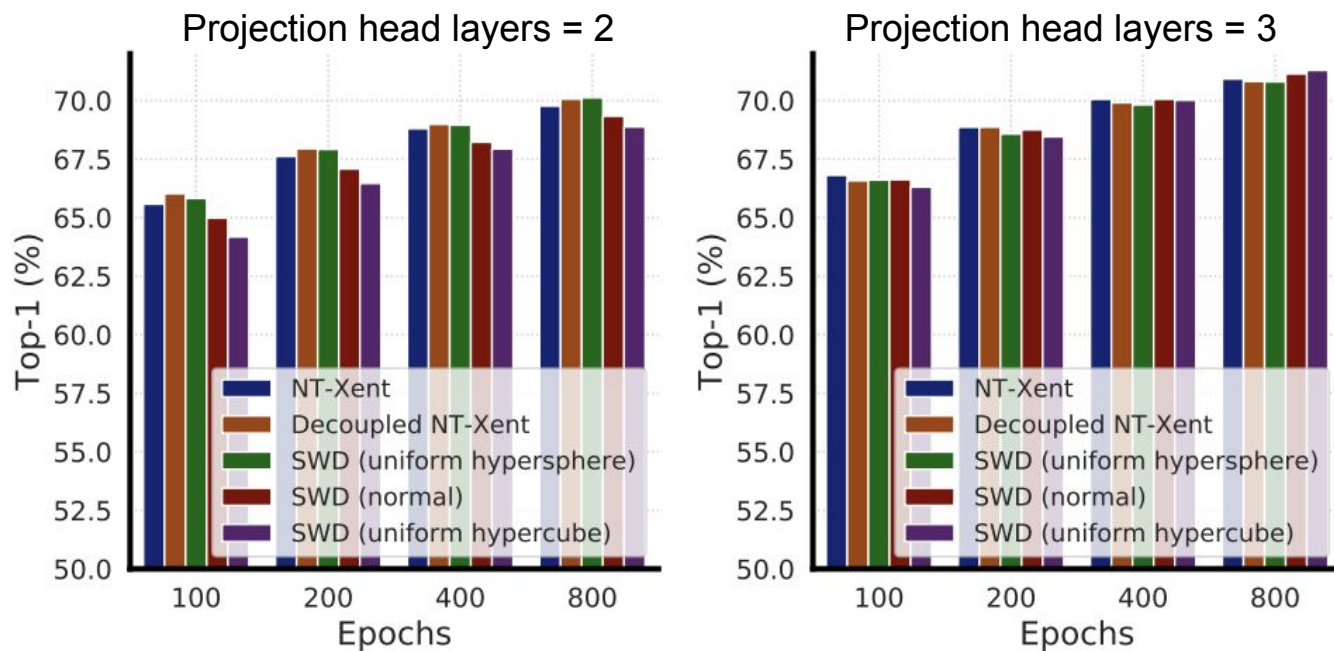
On CIFAR-10, linear evals of ResNet-50 trained with different losses are similar.





# Different generalized contrastive losses perform similarly

On ImageNet, linear evals of ResNet-50 trained with different losses are similar (with a deep projection head).




# The impact of batch size on representation quality is small

With proper hyperparameter tuning, the impact of batch size on representation quality is small.

Table 2: Linear eval accuracy of ResNet-50 on ImageNet.

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



## Three properties studied in this work

1. Do different instantiations of the generalized contrastive loss perform differently?
2. Do instance-based contrastive learning methods learn on images with multiple objects and do they learn good local features?
3. Does feature suppression limit the contrastive learning?

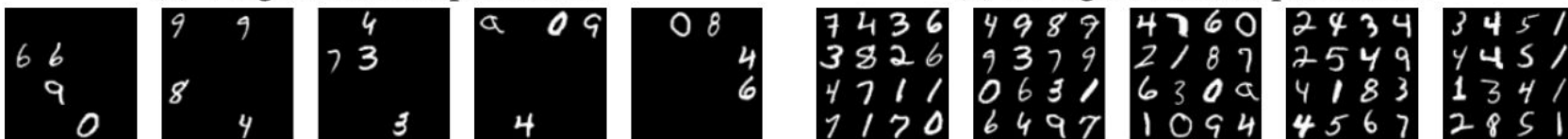
# MultiDigit dataset

Two placement strategies: (1) random, (2) grid.



(a) 4 digits, random placement.

(b) 16 digits, random placement.



(c) 4 digits, in-grid placement.

(d) 16 digits, in-grid placement.

Figure 2: MultiDigit dataset. More digits lead to more overlapping in random placement.

# SimCLR can still learn on images with multiple objects

Training with a given number of digits, but evaluate on a single digit at a time.

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.

		Placing of digits	Number of digits (size $28 \times 28$ )					
			1	2	4	8	12	16
Supervised	Random		99.5	99.5	99.3	99.4	98.9	98.3
	In-grid		99.5	99.6	99.5	99.3	98.6	92.4
SimCLR	Random		98.9	98.9	99.0	98.9	98.2	96.4
	In-grid		98.3	98.6	99.1	99.2	99.1	98.3


# SimCLR learns local features that exhibit hierarchical properties

We apply K-means with various numbers of clusters on the  $l_2$ -normalized hidden features of ResNet before average pooling.



More visualization (on ImageNet and MS-COCO) can be found:

<https://contrastive-learning.github.io/intriguing>

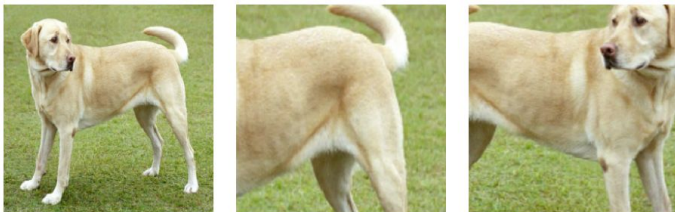


## Three properties studied in this work

1. Do different instantiations of the generalized contrastive loss perform differently?
2. Do instance-based contrastive learning methods learn on images with multiple objects and do they learn good local features?
3. Does feature suppression limit the contrastive learning?

# Feature suppression in contrastive learning

- As studied and shown in SimCLR, contrastive loss requires good design of **data augmentation** to work well.
  - One use of data augmentation is to remove “easy-to-learn” features for the contrastive loss, e.g. color statistics.
- **Competing features** are different features shared between augmented views:



In common: dog class, color distribution, ..



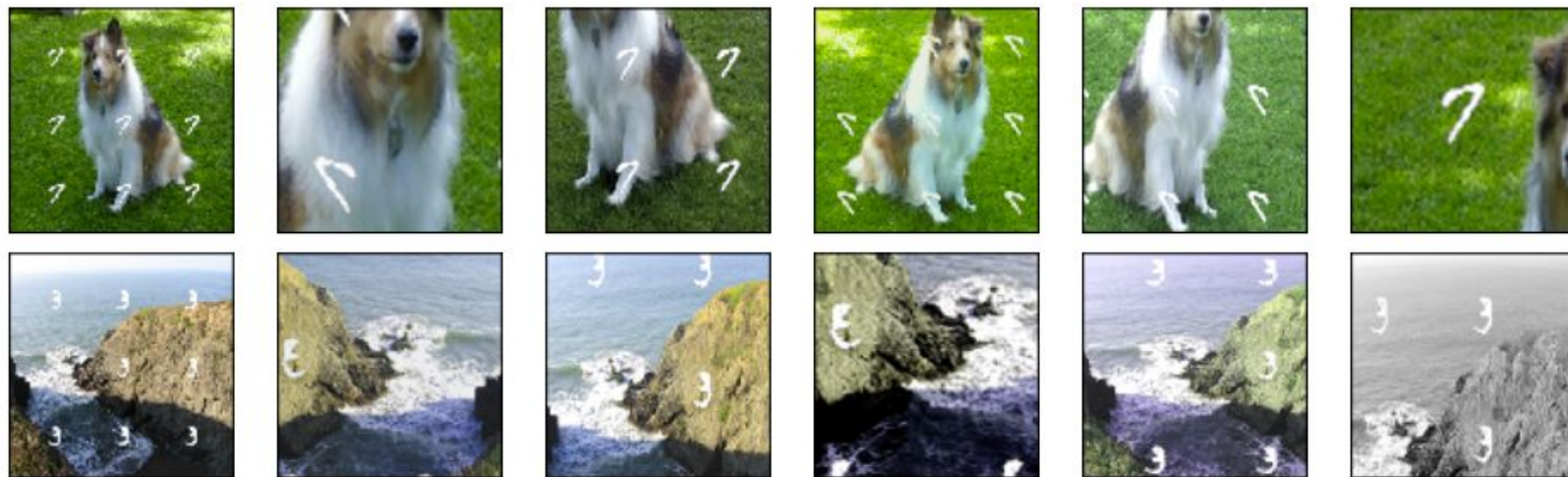
In common: dog class, ..

- Can we **quantitatively** study the impact (suppression effect) of competing features?



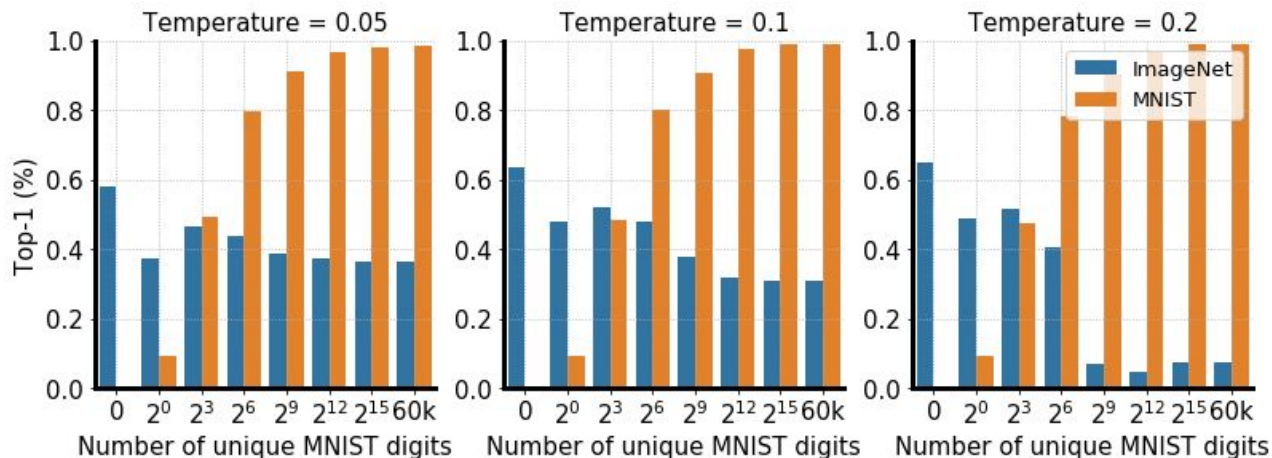
# Datasets with controllable competing features

1. Adding competing features using channel addition: overlay a controlled number of unique MNIST digits on ImageNet images.

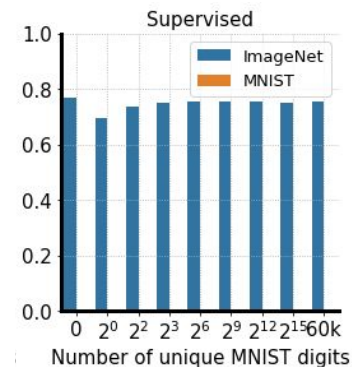


## Easy-to-learn features (MNIST digit) suppress the learning of other features (ImageNet object class)

Standard SimCLR couldn't learn features that are good for linear evaluation on both MNIST digits and ImageNet classes.

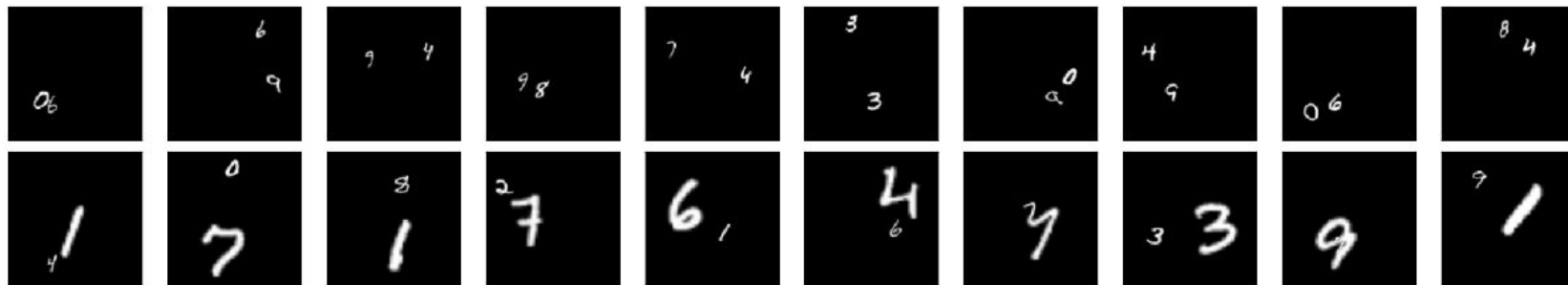


However, supervised learning of ImageNet classes is fine →



# Datasets with controllable competing features

2. Adding competing features using channel addition: place digits of different sizes on the same canvas. We fix the size of one digit and vary the other.



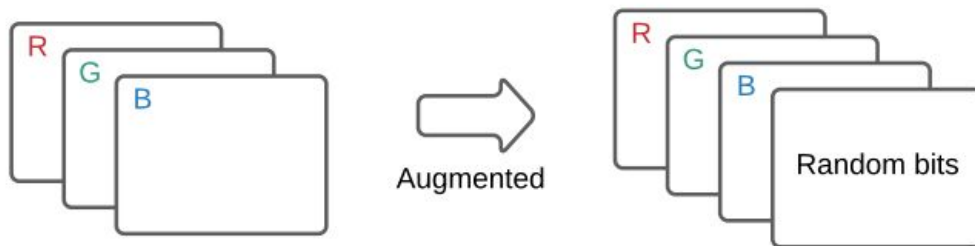
# The presence of dominant object suppresses the learning of features of smaller objects

Table 4: Top-1 linear evaluation accuracy (%) for pretrained ResNet-18 on the MultiDigits dataset. We fix the size of 1st digit while increasing the size of the 2nd digit. For SimCLR, results are presented for two temperatures. Accuracies suffered from a significant drop when increasing 2nd digit size are red colored.

		2nd digit size (1st digit is kept the same size of $20 \times 20$ )						
		$20 \times 20$	$30 \times 30$	$40 \times 40$	$50 \times 50$	$60 \times 60$	$70 \times 70$	$80 \times 80$
Supervised	1st digit	99.1	99.2	99.2	99.2	99.1	99.1	99.0
	2nd digit	99.1	99.5	99.5	99.6	99.5	99.5	99.6
SimCLR ( $\tau = 0.05$ )	1st digit	97.8	97.6	96.2	96.5	88.5	74.5	39.9
	2nd digit	97.8	97.9	97.8	98.3	98.2	97.7	98.2
SimCLR ( $\tau = 0.2$ )	1st digit	98.7	98.8	98.3	87.5	24.9	19.8	20.3
	2nd digit	98.7	99.2	99.2	99.0	99.1	98.9	99.4
Random net (untrained)	1st digit	16.5	16.7	16.6	16.6	16.6	16.9	16.5
	2nd digit	16.5	19.1	21.9	24.1	26.5	28.1	29.0

# Datasets with controllable competing features

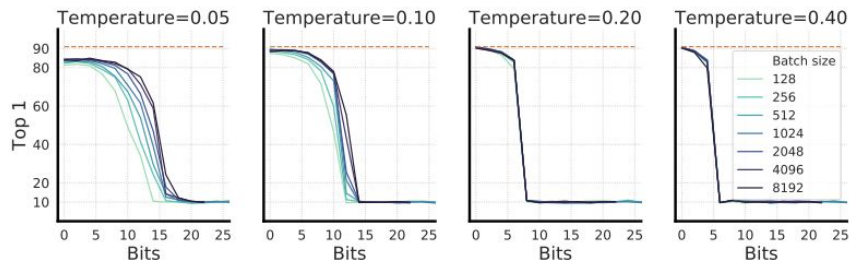
3. Adding competing features using channel concatenation: extra channels are controllable random bits that are shared between views.



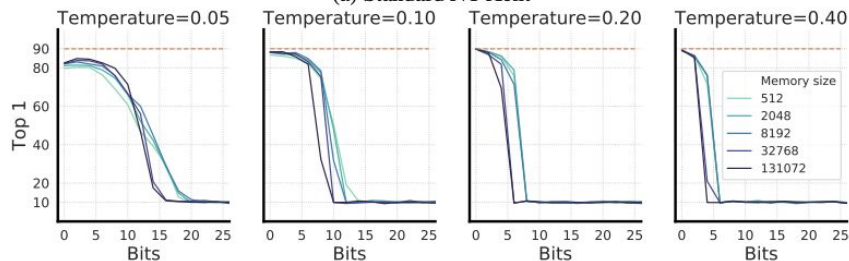
# A few random bits completely disable learning

This phenomenon persists for different batch sizes, losses ( $\tau/\lambda$ ), and the use of EMA network (from MoCo).

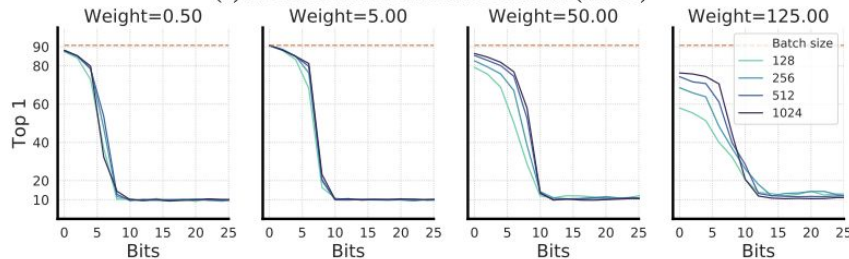
CIFAR-10:



(a) Standard NT-Xent



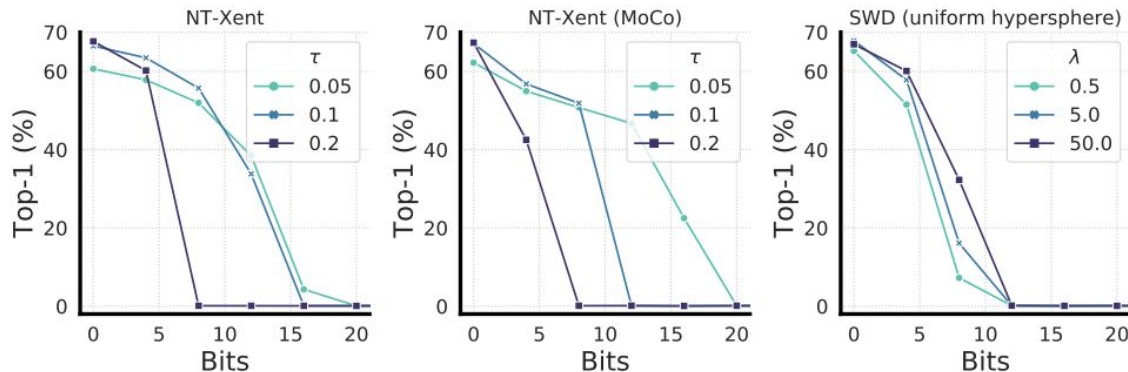
(b) NT-Xent with Momentum Contrast (MoCo)



(c) SWD (uniform hypersphere)

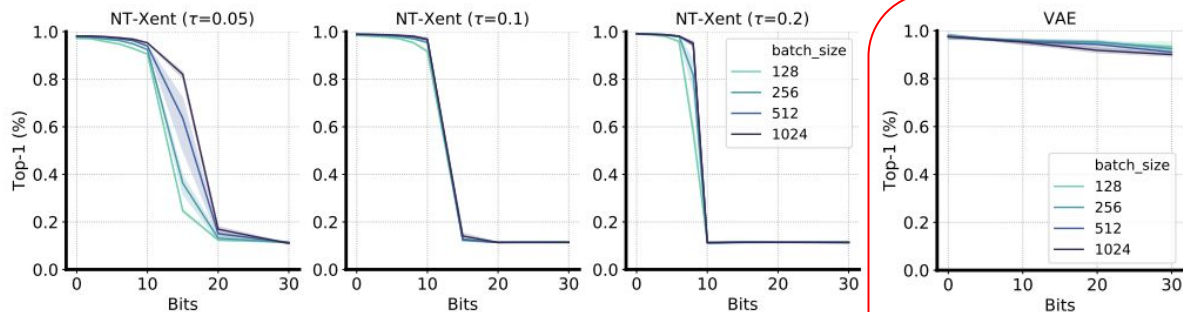
# A few random bits completely disable learning

ImageNet:



(c) ImageNet

MNIST:



(a) MNIST.

(b) MNIST using VAE.

# Conclusion

- We propose and study a generalization of contrastive losses
  - $\mathcal{L}_{\text{generalized contrastive}} = \mathcal{L}_{\text{alignment}} + \lambda \mathcal{L}_{\text{distribution}}$
  - With a multi-layer projection head, various instantiations perform similarly.
- We show instance-based contrastive learning methods can learn on images with multiple objects and also learn meaningful local features.
- In particular, we show feature suppression poses an open challenge
  - So far the most effective method is handcrafted / heuristic-based data augmentation to favor certain features than the others
  - Are there other alternatives?

Code and visualization at <https://contrastive-learning.github.io/intriguing>



# Thank You!



Q&A?