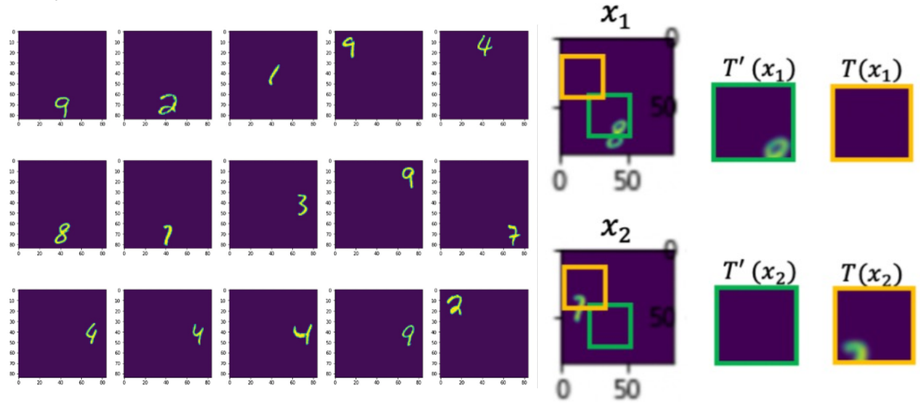
Contrastive Representation Learning with Trainable Augmentation Channel

Masanori Koyama[1], Kentaro Minami[1], Takeru Miyato[1], Yarin Gal[2] (1 Preferred Networks, 2 University of Oxford)

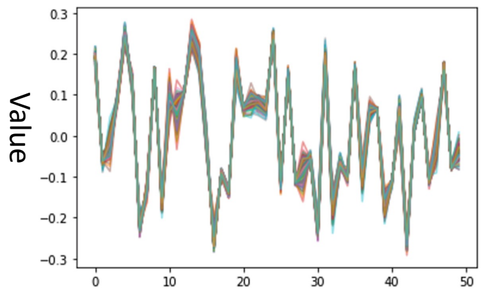
Motivation

In contrastive representation learning, data representation is trained so that it can classify the image instances even when the images are altered by augmentations. However, depending on the datasets, some augmentations can damage the information of the images beyond recognition, and such augmentations can result in collapsed representations. We present a partial solution to this problem by formalizing a stochastic encoding process in which there exist a tug-of-war between the data corruption introduced by the augmentations and the information preserved by the encoder. We show that, with the infoMax objective based on this framework, we can learn a data-dependent distribution of augmentations to avoid the collapse of the representation.



On the dataset like the one we present above(**Regional MNIST**), identifying the crop $T'(x)$ with $T(x)$ in the encoding space for image x would force x_1 to be identified with x_2 via transition-rule of similarity

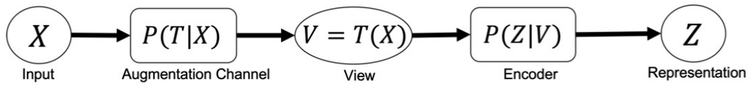
$$T'(x_1) \sim T(x_1) = T'(x_2) \sim T(x_1)$$



(LEFT)Realization of 200 instances of 50 dimensional representations achieved by SimCLR on this example dataset with cropping augmentation. Because of the "blank crop" that attracts all representations via transitivity, all instances look very similar in the latent space.

Proposed Framework

We formalize the encoding process with an input-dependent random augmentation channel and maximize the MI $I(X;Z)$



Proposition 1. Suppose that $P(Z \mid T(X)) = C_\beta \exp(\beta \mathcal{S}(Z, h(T(X))))$ where $\mathcal{S} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is a similarity function on the range of Z and C_β is a constant dependent only on β. Then

$$I(X;Z) = E_{X,Z}\left[\log E_{T'|X}\left[\frac{\exp(\beta \mathcal{S}(Z, h(T'(X))))}{E_{T'',\tilde{X}}[\exp(\beta \mathcal{S}(Z, h(T''(\tilde{X}))))]}\right]\right] \quad (2)$$

Also, when $P(T|X)$ is uniformly distributed on a compact set of view-transformations, the mean approximation of Z and Jensen's inequality on the $E_{T'|X}$ part of (2) recovers the simCLR loss.

Difference from previous MI based approach

Previous MI approach uses *variational approximation of* $I\big(T(X); T'(X)\big)$ with equally distributed T and T' that holds for any encoding map and any f

$$I(V_1; V_2) \geq E_{V_1,V_2}\left[\frac{\exp(f(V_1, V_2))}{E_{V_1'}[\exp(f(V_1', V_2))]}\right]$$

Our approach uses equality relation and proposes a formulation that connects SimCLR and MI with difference resulting only from Jensen's inequality

Connection to Uniformity and Alignment

In the equation (2), the Numerator $H(Z|X)$ and the denominator $H(Z)$ mathematically corresponds directly to the alignment term and the uniformity term in Isola et al (2020 ICLR).

Algorithm

Alternate training of $P(T|X)$ and Encoder with the goal of optimizing the MI

Algorithm 1 Contrastive Representation learning with trainable augmentation Channel(CRL-TAC)

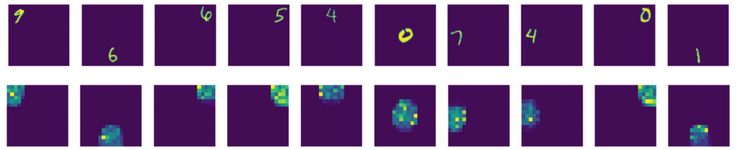
Require: A batch of samples $\{x_k\}$, an encoder model $h_\theta : x \to z$, the number of transformation samples m, a model for conditional random augmentation distribution $x \to P(T|x, \eta)$

1: **for** each iteration i **do**
2: **Update phase for** h
3: Sample $T_{jk} \sim P(T|x_k, \eta)$, $j = 1, ..., m$
4: Apply $\{T_{jk}; j = 1, ..., m\}$ to each x_k, producing a total of $m \times k$ samples of $T_{jk}(x_k)$.
5: Empirically compute the objective (2) or its lower bound, and update θ
6: **Update phase for** $P(T|X)$
7: Sample $T_{jk} \sim Uniform$
8: Evaluate (2) with $P(t_j|x_k, \eta)$ weights, and update η
9: **end for**

Extra Regularization term of to increase $H(T|X)$ worked in favor of both performance and

Results

Trained $P(T|X)$, *where T is a cropping Augmentation*



Linear Classification Protocol (Regional MNIST)

Method	Ours	Ours(topn)	SimCLR	simCLR(oracle)
Projection Head	0.95505 ± 0.0023	0.9552 ± 0.0037	0.3156 ± 0.0044	0.5144 ± 0.011
f output	0.9729 ± 0.0014	0.9748 ± 0.0012	0.4598 ± 0.0056	0.9354 ± 0.0029

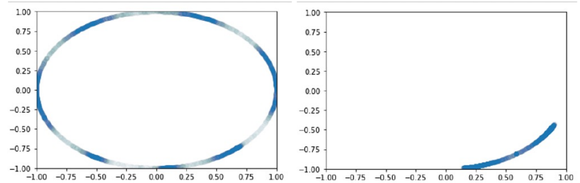
SimCLR(Oracle) applies the representation map trained with uniform cropping augmentation to the crop obtained from the true digit location

Linear Classification Protocol (MNIST)

Method	ours	ours(topn)	SimCLR
Projection Head	0.9642 ± 0.0025	0.9674 ± 0.0015	0.9273 ± 0.0044
f output	0.9805 ± 0.0006	0.9859 ± 0.0004	0.9806 ± 0.0056

Notice that for both Regional MNIST and MNIST, competitive representation is learned at projection head.

Uniformity of the representation



Left: The scatter plot of 2 dimensional representations trained together with P(T|X).
Right: The scatter plot of 2 dimensional representations trained with uniform P(T|X).

(RIGHT)Realization of 200 instances of 50 dimensional representations achieved with trainable $P(T|X)$. Compare this figure to the figure in the first column of this poster.