CRD:
Contrastive Representation Distilation

Yonglong Tian, Dilip Krishnan, Phillip Isola
MIT CSAIL, Google Research
ICLR 2020

Sungman Cho

Contributions

- Contrastive-based objective for transferring knowledge between deep networks.
- Application to model compression, cross-modal transfer, and ensemble distillation.
- New state-of-the-art in many transfer tasks, and sometime even outperforms the teacher network when combined with knowledge distillation.

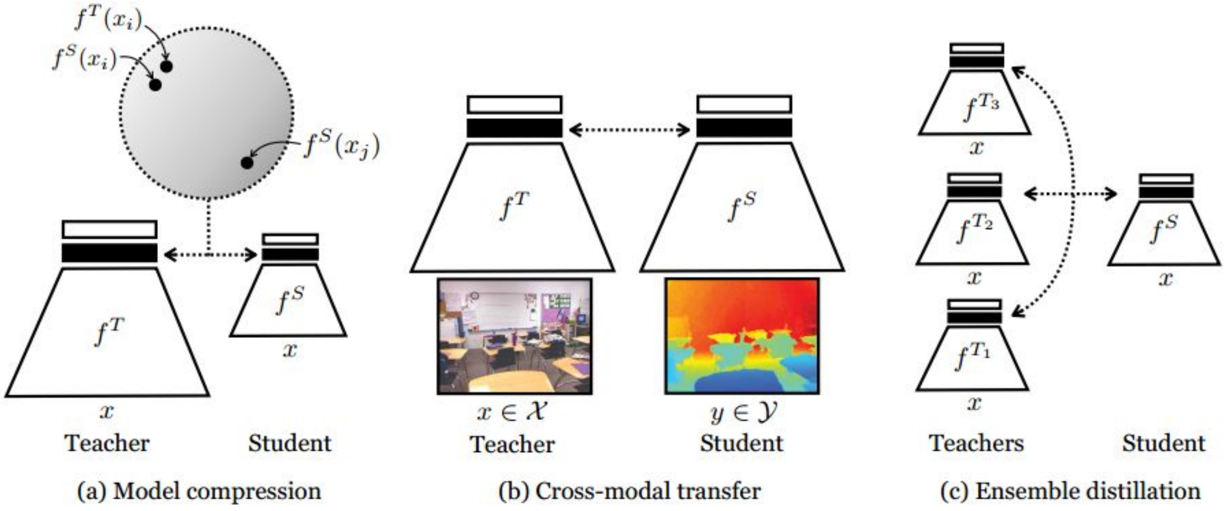
Introduction

- **Knowledge distillation(KD) originally** proposed by **minimizes the KL divergence** between the teacher and student outputs.

  → KL divergence makes intuitive sense **when the output is a distribution**.
    (probability mass function over classes)

Introduction

- However, often we instead wish to transfer **<u>knowledge about a representation.</u>** → **"Cross-modal distillation"** (image to sound or depth)



Introduction

- Representational knowledge is structured.
  (The dimensions exhibit complex interdependencies)

  - Original KD treats all dimensions as independent, conditioned on the input.

    $$\psi(\mathbf{y}^S, \mathbf{y}^T) = \sum_i \phi_i(\mathbf{y}_i^S, \mathbf{y}_i^T)^{*}.$$

    Such a factored objective is insufficient for transferring structural knowledge.
    (i.e. dependencies between output dimensions *i* and *j*)

  - Similar to L2 objective produces blurry results. (Image generation tasks)

Introduction

- **Captures correlations and higher order output dependencies.**

  → We leverage the family of contrastive objectives.

- Contrastive objectives have been used successfully in **representation learning**, **self-supervised** settings

Introduction

- Our objective **maximizes a lower-bound to the mutual information** between the teacher and student representations.

- **Contrastive objective** better **transfers all the information in the teacher's representation**, **rather than only transferring knowledge about conditionally independent output class probabilities.**

Related Work

- **Hinton et al. (2015)** : matching output logits

- **Bucilua et al. (2006)** : introduced **the idea of temperature in the softmax** outputs to better represent smaller probabilities in the output of a single sample.
    - Large temperatures : increase entropy

- **Zagoruyko & Komodakis et al. (2016)** : attention transfer
    - Focuses on the feature maps of the network.
    - Limitation : only with same spatial resolution.

Related Work

- **FitNets (Romero et al., 2014)** : regressions to guide the feature activations

- **Zagoruyko & Komodakis (2016)** : weighted form of this regression.

- **CMC (Tian et al., 2019)** : contrastive objective

- **InfoNCE, NCE (Oord et al., 2018; Gutmann & Hyvarinen., 2010)**
  : use contrastive learning in the context of self-supervised representation learning.
  : objective maximizes a lower bound on mutual information.

Method

- **Contrastive Learning**
    - Positive pairs : close
    - Negative pairs : push apart

$$x \sim p_{\text{data}}(x) \quad \triangleleft \quad \textbf{data}$$
$$S = f^S(x) \quad \triangleleft \quad \textbf{student's representation}$$
$$T = f^T(x) \quad \triangleleft \quad \textbf{teacher's representation}$$

Method

Distribution *q* with latent variable *C*

$$q(T, S|C = 1) = p(T, S), \quad q(T, S|C = 0) = p(T)p(S)$$

Suppose in our data, we are given **1 congruent pair for every *N* incongruent pairs.**

1 congruent pair : drawn from the joint distribution, i.e. the same input provided to *T* and *S*
***N* incongruent pairs** : drawn from the product of marginals; independent randomly drawn inputs provided to *T* and *S*

$$q(C = 1) = \frac{1}{N+1}, \quad q(C = 0) = \frac{N}{N+1}$$

Method

By simple manipulation and Baye's rule

$$
\begin{aligned}
q(C=1|T,S) &= \frac{q(T,S|C=1)q(C=1)}{q(T,S|C=0)q(C=0)+q(x,y|C=1)q(C=1)} \\
&= \frac{p(T,S)}{p(T,S)+Np(T)p(S)}
\end{aligned}
$$

Next, we observe a connection to mutual information as follows:

$$
\begin{aligned}
\log q(C=1|T,S) &= \log \frac{p(T,S)}{p(T,S)+Np(T)p(S)} \\
&= -\log(1+N\frac{p(T)p(S)}{p(T,S)}) \leq -\log(N) + \log \frac{p(T,S)}{p(T)p(S)}
\end{aligned}
$$

Method

$$\begin{aligned}\log q(C=1|T,S) &= \log \frac{p(T,S)}{p(T,S)+Np(T)p(S)} \\ &= -\log(1+N\frac{p(T)p(S)}{p(T,S)}) \leq -\log(N)+\log\frac{p(T,S)}{p(T)p(S)}\end{aligned}$$

Taking expectation on both sides

$$I(T;S) \geq \log(N) + \mathbb{E}_{q(T,S|C=1)} \log q(C=1|T,S) \qquad \triangleleft \quad \textbf{MI bound}$$

Fitting data distribution, [0,1]

$$\begin{aligned}\mathcal{L}_{critic}(h) &= \mathbb{E}_{q(T,S|C=1)}[\log h(T,S)] + N\mathbb{E}_{q(T,S|C=0)}[1-\log(h(T,S))] \\ h^* &= \arg\max_h \mathcal{L}_{critic}(h) \qquad \triangleleft \quad \textbf{optimal critic}\end{aligned}$$

Method

$$I(T;S) \geq \log(N) + \mathbb{E}_{q(T,S|C=1)} \log q(C=1|T,S) \qquad \triangleleft \quad \textbf{MI bound}$$

$$\mathcal{L}_{critic}(h) = \mathbb{E}_{q(T,S|C=1)}[\log h(T,S)] + N\mathbb{E}_{q(T,S|C=0)}[1 - \log(h(T,S))]$$

$$h^* = \arg\max_h \mathcal{L}_{critic}(h) \qquad \triangleleft \quad \textbf{optimal critic}$$

$$h, h^*(T,S) = q(C=1|T,S)$$ **Gibb's inequality**

$$I(T;S) \geq \log(N) + \mathbb{E}_{q(T,S|C=1)}[\log h^*(T,S)]$$

$$f^{S*} = \arg\max_{f^S} \mathbb{E}_{q(T,S|C=1)}[\log h^*(T,S)]$$

Method

$$I(T;S) \geq \log(N) + \mathbb{E}_{q(T,S|C=1)} \log q(C=1|T,S) \quad \triangleleft \quad \textbf{MI bound}$$

$$\mathcal{L}_{critic}(h) = \mathbb{E}_{q(T,S|C=1)}[\log h(T,S)] + N\mathbb{E}_{q(T,S|C=0)}[1 - \log(h(T,S))]$$

$$h^* = \arg\max_h \mathcal{L}_{critic}(h) \quad \triangleleft \quad \textbf{optimal critic}$$

$$\Big\downarrow \quad h, h^*(T,S) = q(C=1|T,S)$$

Gibb's inequality

$$I(T;S) \geq \log(N) + \mathbb{E}_{q(T,S|C=1)}[\log h^*(T,S)]$$

$$f^{S*} = \arg\max_{f^S} \mathbb{E}_{q(T,S|C=1)}[\log h^*(T,S)]$$

Method

$$f^{S*} = \arg\max_{f^S} \mathbb{E}_{q(T,S|C=1)}[\log h^*(T,S)]$$

Weakening the bound

$$\begin{aligned} I(T;S) &\geq \log(N) + \mathbb{E}_{q(T,S|C=1)}[\log h^*(T,S)] + N\mathbb{E}_{q(T,S|C=0)}[\log(1-h^*(T,S))] \\ &= \log(N) + \mathcal{L}_{critic}(h^*) = \log(N) + \max_h \mathcal{L}_{critic}(h) \\ &\geq \log(N) + \mathcal{L}_{critic}(h) \end{aligned}$$

$$\begin{aligned} f^{S*} &= \arg\max_{f^S} \max_h \mathcal{L}_{critic}(h) \qquad \triangleleft \quad \textbf{our final learning problem} \\ &= \arg\max_{f^S} \max_h \mathbb{E}_{q(T,S|C=1)}[\log h(T,S)] + N\mathbb{E}_{q(T,S|C=0)}[\log(1-h(T,S))] \end{aligned}$$

Method

$$
\begin{aligned}
f^{S*} &= \arg\max_{f^S}\max_h \mathcal{L}_{critic}(h) && \triangleleft \quad \textbf{our final learning problem} \\
&= \arg\max_{f^S}\max_h \mathbb{E}_{q(T,S|C=1)}[\log h(T,S)] + N\mathbb{E}_{q(T,S|C=0)}[\log(1-h(T,S))]
\end{aligned}
$$

$$h : \{\mathcal{T}, \mathcal{S}\} \rightarrow [0, 1].$$

Represent : $$h(T,S) = \frac{e^{g^T(T)' g^S(S)/\tau}}{e^{g^T(T)' g^S(S)/\tau} + \frac{N}{M}}$$

Knowledge Distillation Objective

$$\mathcal{L}_{KD} = (1-\alpha)H(y, y^S) + \alpha\rho^2 H(\sigma(z^T/\rho), \sigma(z^S/\rho)) \tag{20}$$

where ρ is the temperature, α is a balancing weight, and σ is softmax function. $H(\sigma(z^T/\rho), \sigma(z^S/\rho))$ is further decomposed into $KL(\sigma(z^T/\rho)|\sigma(z^S/\rho))$ and a constant entropy $H(\sigma(z^T/\rho))$.

Cross-Modal Transfer Loss

Features of teacher network are still valuable to help with learning of the student on another domain.

$$\mathcal{L}_{critic}(h) = \mathbb{E}_{q(T,S|C=1)}[\log h(T,S)] + N\mathbb{E}_{q(T,S|C=0)}[1 - \log(h(T,S))]$$

$$h^* = \arg\max_{h} \mathcal{L}_{critic}(h) \qquad \triangleleft \quad \textbf{optimal critic}$$

Ensemble Distillation Loss

We have *M > 1* teacher networks,

$$\mathcal{L}_{CRD-EN} = H(y, y^S) - \beta \sum_i \mathcal{L}_{critic}(T_i, S)$$

Experiments : Accuracy (CIFAR 100)

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	resnet56 resnet20	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	vgg13 vgg8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD*	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet*	73.58 (↓)	72.24 (↓)	69.21 (↓)	68.99 (↓)	71.06 (↓)	73.50 (↑)	71.02 (↓)
AT	74.08 (↓)	72.77 (↓)	70.55 (↓)	70.22 (↓)	72.31 (↓)	73.44 (↑)	71.43 (↓)
SP	73.83 (↓)	72.43 (↓)	69.67 (↓)	70.04 (↓)	72.69 (↓)	72.94 (↓)	72.68 (↓)
CC	73.56 (↓)	72.21 (↓)	69.63 (↓)	69.48 (↓)	71.48 (↓)	72.97 (↓)	70.71 (↓)
VID	74.11 (↓)	73.30 (↓)	70.38 (↓)	70.16 (↓)	72.61 (↓)	73.09 (↓)	71.23 (↓)
RKD	73.35 (↓)	72.22 (↓)	69.61 (↓)	69.25 (↓)	71.82 (↓)	71.90 (↓)	71.48 (↓)
PKT	74.54 (↓)	73.45 (↓)	70.34 (↓)	70.25 (↓)	72.61 (↓)	73.64 (↑)	72.88 (↓)
AB	72.50 (↓)	72.38 (↓)	69.47 (↓)	69.53 (↓)	70.98 (↓)	73.17 (↓)	70.94 (↓)
FT*	73.25 (↓)	71.59 (↓)	69.84 (↓)	70.22 (↓)	72.37 (↓)	72.86 (↓)	70.58 (↓)
FSP*	72.91 (↓)	n/a	69.95 (↓)	70.11 (↓)	71.89 (↓)	72.62 (↓)	70.23 (↓)
NST*	73.68 (↓)	72.24 (↓)	69.60 (↓)	69.53 (↓)	71.96 (↓)	73.30 (↓)	71.53 (↓)
CRD	**75.48** (↑)	**74.14** (↑)	**71.16** (↑)	**71.46** (↑)	**73.48** (↑)	**75.51** (↑)	**73.94** (↑)
CRD+KD	75.64 (↑)	74.38 (↑)	71.63 (↑)	71.56 (↑)	73.75 (↑)	75.46 (↑)	74.29 (↑)

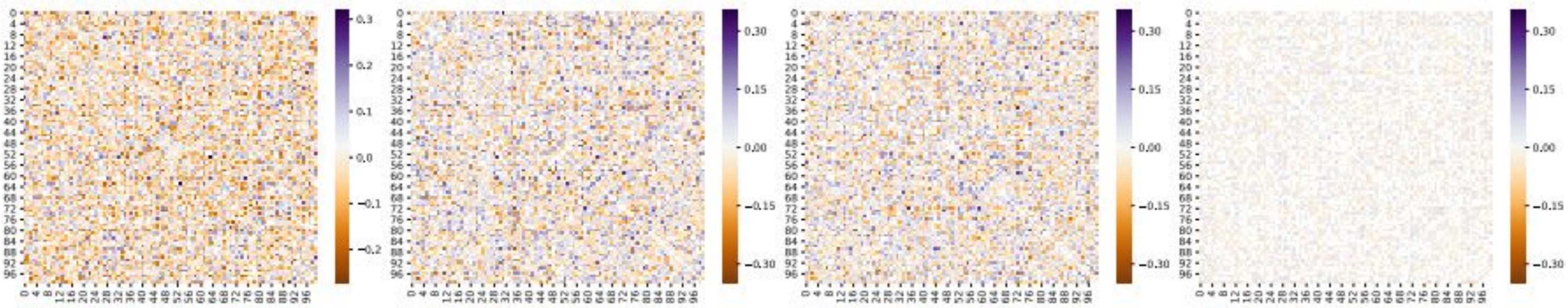
Table 1: Test *accuracy* (%) of student networks on CIFAR100 of a number of distillation methods (ours is CRD); see Appendix for citations of other methods. ↑ denotes outperformance over KD and ↓ denotes underperformance. We note that CRD is the *only* method to always outperform KD (and also outperforms all other methods). We denote by * methods where we used our reimplementation based on the paper; for all other methods we used author-provided or author-verified code. Average over 5 runs.

Experiments : Accuracy (ImageNet)

	Teacher	Student	AT	KD	SP	CC	Online KD *	CRD	CRD+KD
Top-1	26.69	30.25	29.30	29.34	29.38	30.04	29.45	28.83	**28.62**
Top-5	8.58	10.93	10.00	10.12	10.20	10.83	10.41	9.87	**9.51**

Table 3: Top-1 and Top-5 error rates (%) of student network ResNet-18 on ImageNet validation set. We use ResNet-34 released by PyTorch team as our teacher network, and follow the standard training practice of ImageNet on PyTorch except that we train for 10 more epochs. We compare our CRD with KD (Hinton et al., 2015), AT (Zagoruyko & Komodakis, 2016a) and Online-KD (Lan et al., 2018). "*" reported by the original paper Lan et al. (2018) using an ensemble of online ResNets as teacher, no pretrained ResNet-34 was used.

Experiments : Correlation (CIFAR 100)



(a) Student: vanilla (b) Student: AT (c) Student: KD (d) Student: ours (CRD)

Experiments : Transfer

	Student	KD	AT	FitNet	CRD	CRD+KD	Teacher
CIFAR100→STL-10	69.7	70.9	70.7	70.3	71.6	**72.2**	68.6
CIFAR100→TinyImageNet	33.7	33.9	34.2	33.5	**35.6**	35.5	31.5

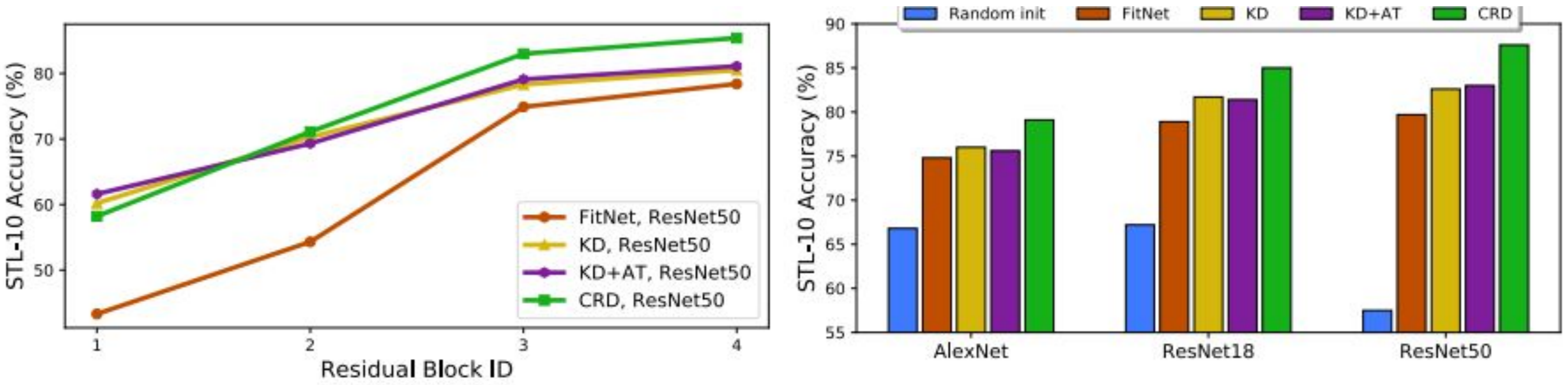
Table 4: We transfer the representation learned from CIFAR100 to STL-10 and TinyImageNet datasets by freezing the network and training a linear classifier on top of the last feature layer to perform 10-way (STL-10) or 200-way (TinyImageNet) classification. For this experiment, we use the combination of teacher network WRN-40-2 and student network WRN-16-2. Classification accuracies (%) are reported.

Experiments : Transfer

	Student	KD	AT	FitNet	CRD	CRD+KD	Teacher
CIFAR100→STL-10	69.7	70.9	70.7	70.3	71.6	**72.2**	68.6
CIFAR100→TinyImageNet	33.7	33.9	34.2	33.5	**35.6**	35.5	31.5

Table 4: We transfer the representation learned from CIFAR100 to STL-10 and TinyImageNet datasets by freezing the network and training a linear classifier on top of the last feature layer to perform 10-way (STL-10) or 200-way (TinyImageNet) classification. For this experiment, we use the combination of teacher network WRN-40-2 and student network WRN-16-2. Classification accuracies (%) are reported.

Experiments : Cross-modal Transfer



(a) Accuracy on STL-10 with linear probing

(b) Accuracy on STL-10 with fully finetuning

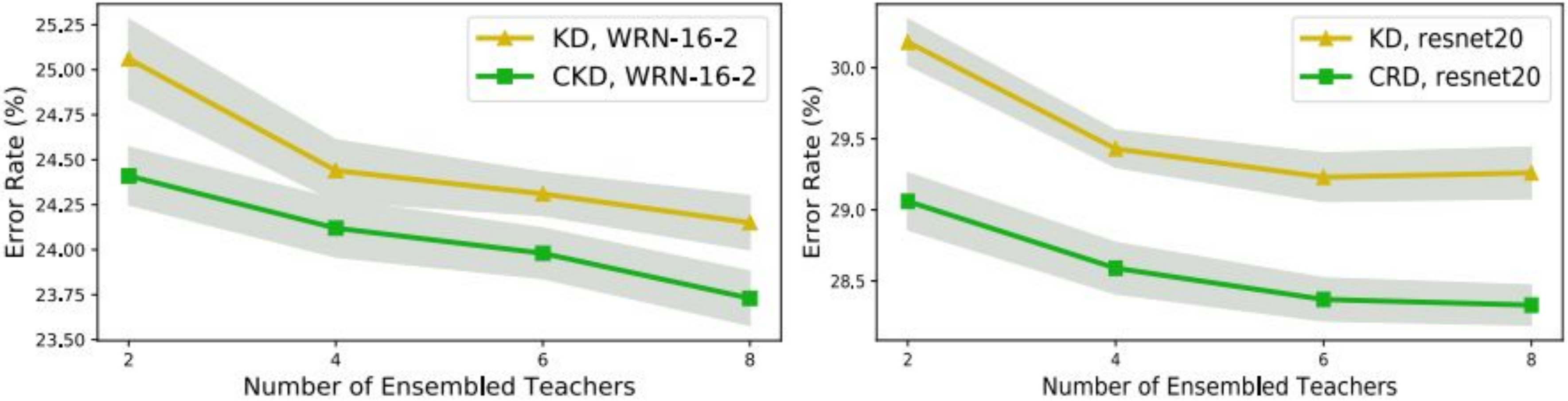
Figure 3: Top-1 classification accuracy on STL-10 using *chrominance* image (*ab* channel in *Lab* color space). We initialize the *chrominance* network randomly or by distilling from a *luminance* network, trained with large-scale labeled images. We evaluate distillation performance by (a) linear probing and (b) fully finetuning.

Experiments : Cross-modal Transfer

ImageNet → NYU-Depth

Metric (%)	Random Init.	KD	KD+AT	FitNet	CRD
Pix. Acc.	56.4	58.9	60.1	60.8	**61.6**
mIoU	35.8	38.0	39.5	40.7	**41.8**

Experiments : Ensemble distillation



(a) WideResNet-16-2 (b) ResNet20

Figure 4: Distillation from an ensemble of teachers. We vary the number of ensembled teachers and compare KD with our CRD by using (a) WRN-16-2 and (b) ResNet20. Our CRD consistently achieves lower error rate.

Ablation Study

sampling	objective	WRN-40-2 WRN-16-2	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	vgg13 vgg8
$i \neq j$	InfoNCE	74.78	70.56	72.67	74.69	73.24
	Ours	74.48	70.64	72.64	74.67	73.39
$y_i \neq y_j$	InfoNCE	75.15	71.39	**73.53**	75.22	73.74
	Ours	**75.48**	**71.46**	73.48	**75.51**	**73.94**

Table 6: Ablative study of different contrastive objectives and negative sampling policies on CIFAR100. For contrastive objectives, we compare our objective with InfoNCE (Oord et al., 2018); For negative sampling policy, when given an anchor image x_i from the dataset, we consider either randomly sample negative x_j such that (a) $i \neq j$, or (b) $y_i \neq y_j$ where y represents the class label. Average over 5 runs.

Ablation Study

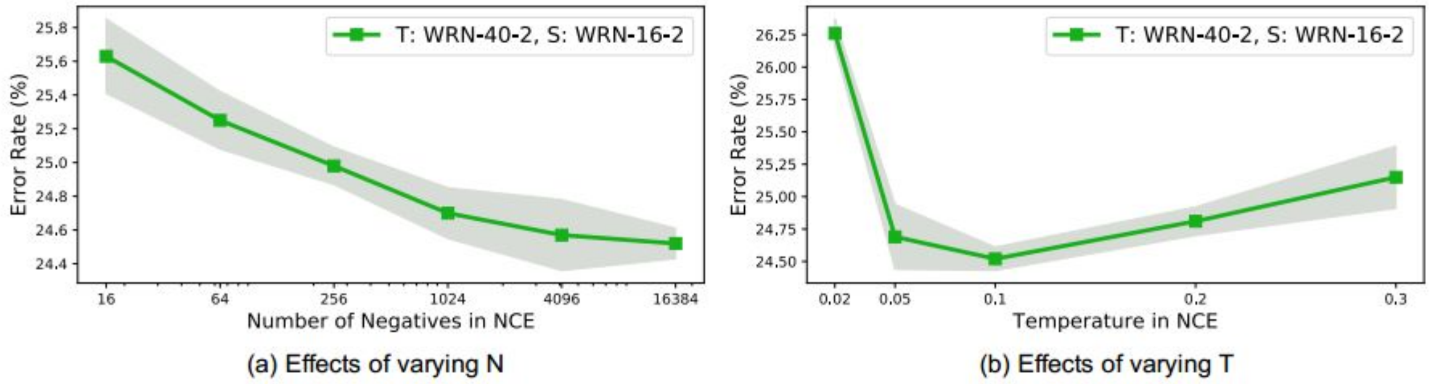


Figure 5: Effects of varying the number of negatives, shown in (a), or the temperature, shown in (b).

Computational Cost :

Original : 2GFLOPs

CRD : 260 MFLOPs (12% of the original)

The memory bank for storing all 128-d features of ImageNet only costs around 600MB

Conclusion

- Developed a novel technique for neural network **distillation**, **using the concept of contrastive objectives**, which are usually used for representation learning.
- Experimented with our objective on a number of applications such as **model compression, cross-modal transfer and ensemble distillation**.
- Contrastive learning is a simple and effective objective with practical benefits.