Information Theoretic Representation Distillation

Roy Miles[1*] Adrian Lopez-Rodriguez[1*] Krystian Mikolajczyk[1]

[1]Imperial College London

*Both authors contributed equally

Imperial College London

Github Code Link

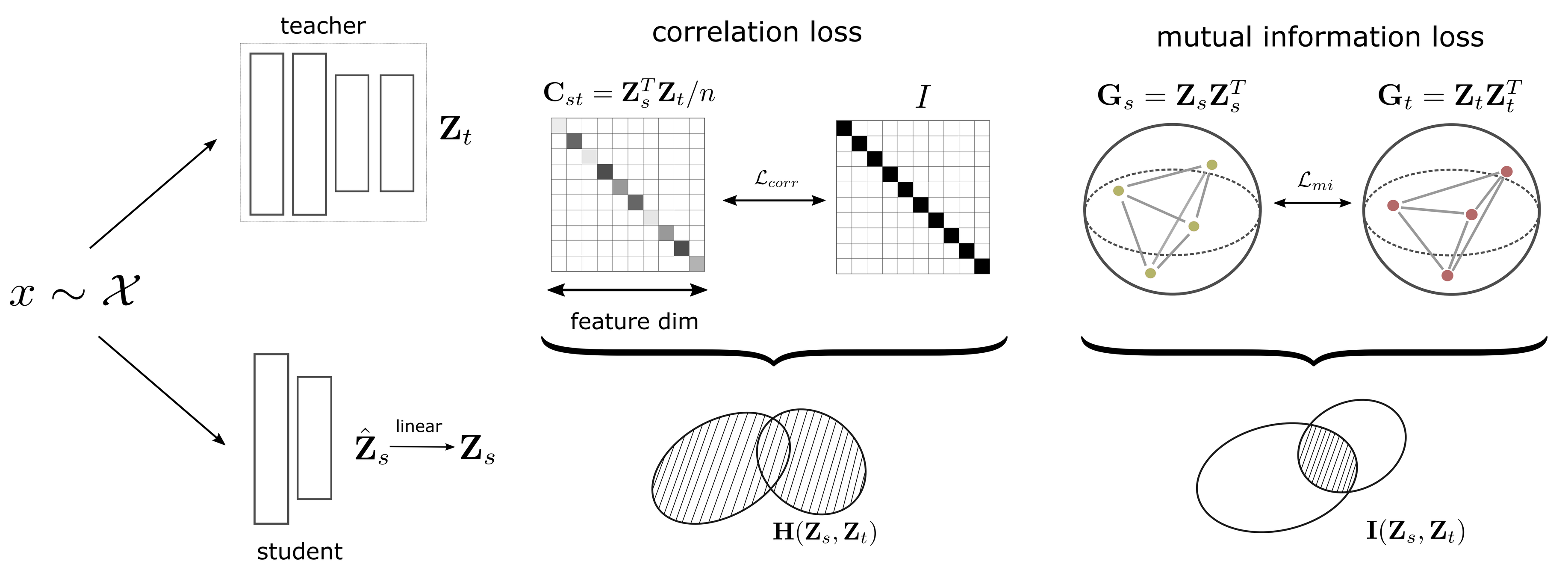
BMVC 2022

Motivation

1. Most SoTA distillation methods are far **too computationally expensive** to adopt in practice, both in terms of their memory consumption, and FLOPs.
2. We introduce a **theoretical framework** for knowledge distillation that is rooted in information theory.
3. In doing so, we derive two complimentary losses that provide a **new SoTA on standard distillation benchmarks**.
4. We also consider two very different task, namely NLP, and binary network classification, thus demonstrating the **flexibility and scope** of our proposed losses.

Method

Our proposed are derived from a set of **cheap** matrix-based estimators [1, 2] resembling rényi's entropy in the single and multi-variate case.



Correlation Loss The correlation loss aims to match the information present in each **feature dimension** between the teacher and student representations. The parameter α is related to Rényi's entropy order.

$$\mathcal{L}_{corr} = \log_2 \sum_{i=1}^{d} |v_i - 1|^{2\alpha} \tag{1}$$

Mutual Information Loss The mutual information loss provides an additional complimentary objective whereby we transfer the **intra-batch similarity** (i.e., the relationship between samples) from the teacher representations to the student representations. The second loss transfers to relationship between different data points within the batch.

$$\mathcal{L}_{mi} = \log_2 \|\mathbf{G}_s\|_F^2 - \log_2 \|\mathbf{G}_{st}\|_F^2 \tag{2}$$

The final loss is then just a weighted sum of these two. Experiments found that the model performance is **robust to the choice in loss weights**.

$$\mathcal{L}_{ITRD} = \mathcal{L}_{XE} + \beta_{corr}\mathcal{L}_{corr} + \beta_{mi}\mathcal{L}_{mi} \tag{3}$$

Experiments & Results

CIFAR-100 test *accuracy* (%) of student networks trained with a number of distillation methods. The best results are highlighted in **bold**, while the second best results are underlined. ITRD achieves the best performance for 10 out of 13 of the architecture pairs, with a **6.8% and 24.4% relative improvement** over ReviewKD and WCoRD respectively.

Teacher Student	W40-2 W16-2	W40-2 W40-1	R56 R20	R110 R20	R110 R32	R32x4 R8x4	V13 V8	V13 MN2	R50 MN2	R50 V8	R32x4 SN1	R32x4 SN2	W40-2 SN1
Teacher	75.61	75.61	72.32	74.31	74.31	79.42	74.64	74.64	79.34	79.34	79.42	79.42	75.61
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36	64.60	64.60	70.36	70.50	71.82	70.50
KD	74.92	73.54	70.66	70.67	73.08	73.33	72.98	67.37	67.35	73.81	74.07	74.45	74.83
CRD	75.64	74.38	71.63	71.56	73.75	75.46	74.29	69.94	69.54	74.58	75.12	76.05	76.27
WCoRD	76.11	74.72	**71.92**	71.88	74.20	76.15	74.72	70.02	70.12	74.68	75.77	76.48	76.68
ReviewKD	**76.12**	75.09	71.89	-	73.89	75.63	74.84	70.37	69.89	-	**77.45**	**77.78**	77.14
Ours: $\mathcal{L}_{corr}$	75.85 ±0.12	74.90 ±0.29	71.45 ±0.21	71.77 ±0.34	74.02 ±0.27	75.63 ±0.09	74.70 ±0.27	69.97 ±0.33	**71.41** ±0.41	**75.71** ±0.02	76.80 ±0.28	77.27 ±0.25	**77.35** ±0.25
Ours: $\mathcal{L}_{corr} + \mathcal{L}_{mi}$	**76.12** ±0.04	**75.18** ±0.22	71.47 ±0.07	**71.99** ±0.46	**74.26** ±0.05	**76.19** ±0.22	**74.93** ±0.12	**70.39** ±0.39	71.34 ±0.33	75.49 ±0.32	76.91 ±0.19	77.40 ±0.06	77.09 ±0.08

Experiments are also performed on **ImageNet**, where ITRD is comparable to state-of-the-art. The only two methods with a higher accuracy are 2× and 20× more computationally expensive respectively.

	Model	EM	F1
	Teacher (BERT)	81.5	88.6
T6	DistilBERT	79.1	86.9
	TextBrewer	80.8	88.1
	ITRD	**81.5**	**88.5**
T3	TextBrewer	76.3	84.8
	ITRD	**77.7**	**85.8**

To show the wide applicability of our method, we consider distillation on a reading comprehension task. ITRD **outperforms both NLP-specific distillation methods TextBrewer and DistilBert** in both the Exact Match (EM) metrics and in F1 score.

ITRD can be used to **reduce the gap between binary and full-precision (FP) networks**. Both CRD and ReviewKD degrade the BNN performance and, in contrast, ITRD improves upon the original ReCU by 1.3%, which is only 0.7% shy of the FP model.

Network	Method	W/A	Top-1 (%)
ResNet-18	FP	32/32	94.8
	RBNN	1/1	92.2
	ReCU	1/1	92.8
	ReCU + CRD	1/1	92.1
	ReCU + ReviewKD	1/1	92.6
	ReCU $+\mathcal{L}_{corr} + \mathcal{L}_{mi}$	1/1	**94.1**

ITRD losses can be implemented in **as few as 15 lines of code**.
We have also publicly releases the complete training and inference pipelines.

References

[1] Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C. Principe. Measures of entropy from data using infinitely divisible Kernels. IEEE Transactions on Information Theory, 2015.
[2] Paul L. Williams and Randall D. Beer. Nonnegative Decomposition of Multivariate Information. 2010.