Information Theoretic Representation Distillation

Roy Miles*
r.miles18@imperial.ac.uk
Adrian Lopez-Rodriguez*
al4415@imperial.ac.uk
Krystian Mikolajczyk
k.mikolajczyk@imperial.ac.uk

MatchLab
Imperial College London
Department of Electrical and Electronic Engineering
London, UK

Abstract

Despite the empirical success of knowledge distillation, current state-of-the-art methods are computationally expensive to train, which makes them difficult to adopt in practice. To address this problem, we introduce two distinct complementary losses inspired by a cheap entropy-like estimator. These losses aim to maximise the correlation and mutual information between the student and teacher representations. Our method incurs significantly less training overheads than other approaches and achieves competitive performance to the state-of-the-art on the knowledge distillation and cross-model transfer tasks. We further demonstrate the effectiveness of our method on a binary distillation task, whereby it leads to a new state-of-the-art for binary quantisation and approaches the performance of a full precision model. Code: github.com/roymiles/ITRD

1 Introduction

Deep learning has significantly advanced state-of-the-art across a wide range of computer vision tasks. Despite this success, most models are too computationally expensive to deploy on resource-constrained devices. Fortunately, the training of such models is coupled with significant parameter redundancy, which has been explicitly exploited in the pruning and quantisation literature [3, 19, 29, 60]. Knowledge distillation proposes an alternative approach whereby a much larger pre-trained model can provide additional supervision for a smaller model during training. This paradigm removes the restriction of the two models to share the same underlying architecture, thus enabling hand-crafted designs of the target architecture to meet the imposed resource constraints. However, some of the recent state-of-the-art distillation methods, *e.g.* the recent union of self-supervision and knowledge distillation [45, 47], have made it increasingly expensive to train these student models. To this end, we develop a distillation method with a low computational overhead.

Information theory provides a natural lens for quantifying the statistical relationship between these models, and so is a common framework for deriving distillation losses [5, 39].

* The authors contributed equally to this paper

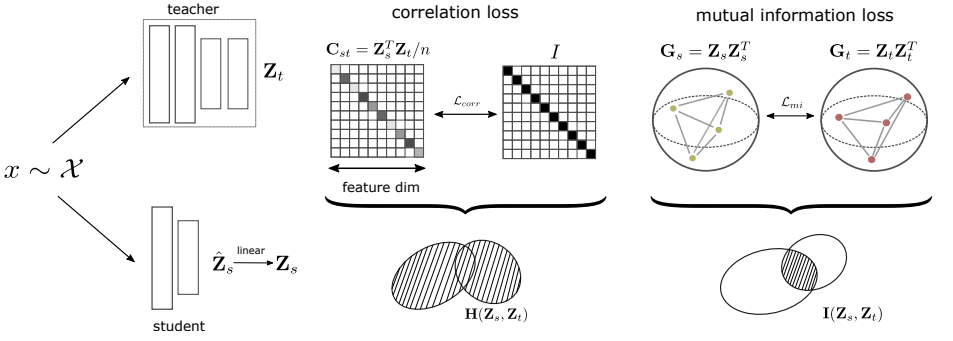

Figure 1: Information theoretic representation distillation (ITRD) involves two distinct losses, namely a correlation loss and a mutual information loss. The former loss maximises the correlation between the student and teacher, while the latter maximises a quantity resembling the mutual information that aims to transfer the intra-batch sample similarity.

Hence, we propose **I**nformation **T**heoretic **R**epresentation **D**istillation (ITRD) as a unified and computationally efficient framework that directly connects information theory with representation distillation. Specifically, this framework is inspired by the generalised Rényi's entropy and makes the training for specific applications more effective. Rényi's entropy is a generalisation of Shannon's entropy and has led to improvements in other areas [23, 36, 51]. As figure 1 shows, we propose to model the distillation task with two distinct loss functions that correspond to maximising the correlation and mutual information between the student and teacher representations. The correlation loss aims to increase the similarity between teacher and student representations across the feature dimension. Conversely, the mutual information loss aims to match the intra-batch sample similarity between the teacher and the student. Our results show a strong accuracy v.s. training cost trade-off in comparison to state-of-the-art across two standard benchmarks, CIFAR100 and ImageNet, for a range of architecture pairings where we achieve up to 24.4% relative improvement. Our loss directly addresses the training efficiency problem, which we believe will encourage its adoption amongst machine learning researchers and practitioners. We further demonstrate the effectiveness of this framework on representation transfer, binary network transfer and NLP architecture transfer, whereby we are able to improve upon the state-of-the-art for all tasks.

2 Related Work

Knowledge Distillation (KD) attempts to transfer the knowledge from a large pre-trained model (teacher) to a much smaller compressed model (student). This was originally introduced in the context of image classification [14], whereby the soft predictions of the teacher can act as pseudo ground truth labels for the student. The soft predictions then provide the student with supervision on the correlations between classes which are not explicitly available from one-hot encoded ground truth labels. Spherical knowledge distillation [11] proposes to re-scale the logits before KD to address the capacity gap problem, while Prime-Aware Adaptive Distillation [58] introduces an adaptive sample weighting. Hinted losses provide a natural extension of KD using an L_2 distance between the student and teacher's intermediate representation [31]. Attention transfer [53] proposed to re-weight the spatial en-

tries before the matching losses, while neuron selectivity transfer [15], similarity-preserving KD [41], and relational KD [24] attempt to transfer the structural similarity. Similarly, FSP matrices [48] attempt to capture the flow of information and Review KD [6] propose the use of attention-based and hierarchical context modules. KD can also be modelled directly within a probabilistic framework [2, 25] through estimating and maximising the mutual information between the student and the teacher. ICKD [21] propose to transfer the correlation between channels of intermediate representations. A natural extension of supervised contrastive learning in the context of knowledge distillation was proposed in CRD [39]. WCoRD [5] also use a contrastive learning objective but through leveraging the dual and primal forms of the Wasserstein distance. CRCD [59] further develop this contrastive framework through the use of both feature and gradient information. Unfortunately, all of these contrastive methods require a large set of negative samples, which are sampled from a memory bank that incurs in additional memory and computational costs, which we avoid altogether.

Additional self-supervision tasks have shown strong performance when coupled with representation distillation. Both SSKD [45] and HSAKD [47] introduce auxiliary tasks for classifying image rotation. However, these added self-supervision tasks incur a high training cost due to augmenting the training batches and adding additional classifiers. Weight sharing through jointly training sub-networks has also been shown to provide implicit knowledge distillation [22, 49, 50] and promising results. In this paper, we propose two distinct distillation losses applied to the features before the final fully-connected layer. Similarly to CRD [39], we posit that the logit representations lack relevant structural information that is necessary for effective distillation through the low dimensional embedding, while using the earlier intermediate representations can hinder the downstream task performance.

Information Theory (IT) provides a natural lens for interpreting and modelling the statistical relationships between intermediate representations of a neural network. This intersection of information theory and deep learning has subsequently led to a rigorous foundation in understanding the dynamics of training [1, 40], while offering fruitful insights into other application domains, such as network pruning and knowledge distillation. In the context of representation distillation, most losses can be modelled as maximising some lower bound on the mutual information between the student and the teacher [5, 39]. In this work, we propose to forge an alternative connection between knowledge distillation and information theory using infinitely divisible kernels [4]. Specifically, we show that maximising both the correlation and mutual information yields two complimentary loss functions that can be related to these entropy-like quantities. We achieve this using a matrix-based function that closely resembles Rényi's α-entropy [33, 34, 44], which is in turn a natural extension of the well-known Shannon's entropy used in IT. More recently, this work has been applied in a representation learning context [53] for parameterising the information bottleneck principle.

3 Preliminaries

Representation Distillation describes the methods that use the representation space that is given as the input to the final fully connected layer of a model. The generalised loss used for representation distillation can be concisely expressed in the following form:

$$\mathcal{L} = \mathcal{L}_{XE}(\mathbf{y}, \text{softmax}(\mathbf{y}_s)) + \beta \cdot d(\mathbf{z}_s, \mathbf{z}_t) \tag{1}$$

where $\mathbf{z}_s \in \mathbb{R}^{d_s}$ and $\mathbf{z}_t \in \mathbb{R}^{d_t}$ are the student and teacher representations, β is a loss weighting, and d is the distillation loss function. The cross entropy $\mathcal{L}_{XE}$ between labels $\mathbf{y}$ and student logits $\mathbf{y}_s$ can be defined as the sum of an entropy and KL divergence term. Furthermore, standard KD [13] uses a further KL divergence as the distillation loss between the student and teacher logits, with a temperature term to soften or sharpen the two distributions.

Following [39], the motivation for using the feature representation space, as opposed to logits or any of the intermediate feature maps is two-fold. Firstly, this space preserves the structural information about the input, which may be lost in the logits. Secondly, intermediate feature matching losses may negatively impact the students' downstream performance in the cross-architecture tasks due to differing inductive biases [39], while also incurring significant computational and memory overheads due to the high dimensionality of these feature maps. In our work, to maximize the information transfer, we propose to express the distillation loss $d(.,.)$ as the weighted sum of a correlation and mutual information term. Below we link these two terms to a general formulation of entropy [34].

Information Theory Rényi's α-entropy [30] provides a natural extension of Shannon's entropy, which has been successfully applied in the context of differential privacy [23], understanding autoencoders [51], and face recognition [36]. For a random variable X with probability density function (PDF) $f(x)$ in a finite set $\mathcal{X}$, the α-entropy $\mathbf{H}_\alpha(X)$ is defined as:

$$\mathbf{H}_\alpha(f) = \frac{1}{1-\alpha}\log_2 \int_{\mathcal{X}} f^\alpha(x)dx \tag{2}$$

Where the limit as $\alpha \to 1$ is the well-known Shannon entropy. To avoid the need for evaluating the underlying probability distributions, a set of entropy-like quantities that closely resemble Renyi's entropy were proposed in [34, 44] and instead estimate these information quantities directly from data. They are based on the theory of infinitely divisible matrices and leverage the representational power of reproducing kernel Hilbert spaces (RKHS), which have been widely studied and adopted in classical machine learning. Since its fruition, this framework has been applied in understanding convolutional neural networks (CNNs) [52], whereby they verify the important data processing inequality in information theory and further demonstrate a redundancy-synergy trade-off in layer representations. We propose to apply these estimators in the context of representation distillation.

We now provide definitions of the entropy-based quantities and their connections with positive semidefinite matrices. This idea then leads to a multi-variate extension using Hadamard products, from which conditional and mutual information can be defined. For brevity, we omit the proofs and connections with Rényi's axioms, which can be found in [34, 44].

Definition 1: Let $X = \{x^{(1)}, \ldots x^{(n)}\}$ be a set of n data points of dimension d and $\kappa : X \times X \to \mathbb{R}$ be a real-valued positive definite kernel. The Gram matrix $\mathbf{K}$ is obtained from evaluating κ on all pairs of examples, that is $K_{ij} = \kappa(x^i, x^j)$. The matrix-based analogue to Rényi's α-entropy for a normalized positive definite (NPD) matrix $\mathbf{A}$ of size $n \times n$, such that $tr(\mathbf{A}) = 1$, can be given by the following functional:

$$\mathbf{S}_\alpha(\mathbf{A}) = \frac{1}{1-\alpha}\log_2(tr(\mathbf{A}^\alpha)) = \frac{1}{1-\alpha}\log_2\left[\sum_{i=1}^{n}\lambda_i(\mathbf{A}^\alpha)\right] \tag{3}$$

where $\mathbf{A}$ is the kernel matrix $\mathbf{K}$ normalised to have a trace of 1 and $\lambda_i(\mathbf{A})$ denotes its i-th eigenvalue. This estimator can be seen as a statistic on the space computed by the kernel κ, while also satisfying useful properties attributed to entropy. In practice, the choice of both

κ and α can be governed by domain-specific knowledge, which we exploit for the task of knowledge distillation. The *log* in these definitions, conventionally taken as base 2, can be interpreted as a data-dependant transformation, and its argument is called the *information potential* [33]. In an optimisation context, the information potential and entropy definitions can be used interchangeably since they are related by a strictly monotonic function.

We are interested in the statistical relationship between two sets of variables, namely the student and teacher representations. To measure this relationship, we introduce the notion of joint entropy, which naturally arises using the product kernel.

Definition 2: Let X and Y be two sets of data points. After computing the corresponding Gram matrices $\mathbf{A}$ and $\mathbf{B}$, the joint entropy is then given by:

$$\mathbf{S}_\alpha(\mathbf{A},\mathbf{B}) = \mathbf{S}_\alpha\left(\frac{\mathbf{A}\circ\mathbf{B}}{tr(\mathbf{A}\circ\mathbf{B})}\right) \tag{4}$$

where $\circ$ denotes the Hadamard product between two matrices. Using these two definitions, the notion of conditional entropy and mutual information can be derived. We focus on the mutual information, which is given by:

$$\mathbf{I}_\alpha(\mathbf{A};\mathbf{B}) = \mathbf{S}_\alpha(\mathbf{A}) + \mathbf{S}_\alpha(\mathbf{B}) - \mathbf{S}_\alpha(\mathbf{A},\mathbf{B}) \tag{5}$$

Both equation 4 and 5 form a foundation for the correlation and mutual information losses respectively, which are proposed in the following section.

4 Information Theoretic Loss Functions

In this section we introduce two distillation losses that use two distinct and complementary similarity measures between the student and teacher representations. The first loss uses a correlation measure which captures the similarity across the feature dimension, while the second loss is derived from a measure of mutual information and captures the similarity between examples within the mini-batch.

4.1 Maximising correlation

This first loss attempts to correlate the student and teacher representations. The intuition is that if the two sets of representations are perfectly correlated then the student is at least as discriminative as the teacher. Let $\mathbf{Z}_s \in \mathbb{R}^{n\times d}$ and $\mathbf{Z}_t \in \mathbb{R}^{n\times d}$ [1] denote a batch of representations from the student and teacher respectively. These matrices are computed before the final fully-connected layer to preserve the structural information of the data, thus enabling a strong distillation signal for the student. We first normalise these representations to zero mean and unit variance across the batch dimension and then propose to construct a cross-correlation matrix, $\mathbf{C}_{st} = \mathbf{Z}_s^T\mathbf{Z}_t/n \in \mathbb{R}^{d\times d}$. Perfect correlation between the two sets of representations is achieved if all of the diagonal entries $v_i = (\mathbf{C}_{st})_{ii}$ are equal to one. To formulate this as a minimization problem, we propose the following loss:

$$\mathcal{L}_{corr} = \log_2 \sum_{i=1}^{d} |v_i - 1|^{2\alpha} \tag{6}$$

[1] For clarity, we omit a linear embedding layer used on the student representations to match its dimensionality with the teacher.

This general objective is motivated by the recent work on Barlow Twins [56] for self-supervised learning, however, there are several distinct differences. Firstly, we drop the redundancy reduction term, which minimizes the off-diagonal entries in the cross correlation matrix, since we are not jointly learning both representations, *i.e.*, the teacher is fixed. In fact we observed that this objective significantly hurts the performance of the student. This performance degradation was similarly observed when decorrelating the off-diagonal entries in the self-correlation matrix $\mathbf{C}_{ss}$, and is likely a consequence of the limited model capacity. Secondly, we introduce an α parameter, which provides a natural generalisation to emphasise low or highly correlated features. Finally, the $\log_2$ transformation was empirically shown to improve the performance by reducing spurious variations within a batch. These modifications were not only empirically justified, but also provide a closer relationship with the matrix-based entropy function in equation 3 (see Supplementary).

4.2 Maximising mutual information

The correlation loss aims to match the information present in each feature dimension between the teacher and student representations. 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 natural choice for achieving this through the lens of information theory is to maximise the mutual information between the two representations. Maximising the mutual information has been successfully applied in past distillation methods [7], following the idea that a high mutual information indicates a high dependence between the two models and thus resulting in a strong student representation. Most other works relate their distillation losses to some lower bound on mutual information [39], however, using an alternative cheap entropy-like estimator, we propose to maximise this quantity directly:

$$\mathcal{L}_{mi} = -\mathbf{I}_\alpha(\mathbf{G}_s;\mathbf{G}_t) = \mathbf{S}_\alpha(\mathbf{G}_s,\mathbf{G}_t) - \mathbf{S}_\alpha(\mathbf{G}_s) - \cancel{\mathbf{S}_\alpha(\mathbf{G}_t)} \tag{7}$$

where $\mathbf{G}_s \in \mathbb{R}^{n\times n}$ and $\mathbf{G}_t \in \mathbb{R}^{n\times n}$ are the student and teacher Gram matrices (*i.e.*, $\mathbf{A}$ and $\mathbf{B}$ in equation 5). These matrices are constructed using a batch of normalised features $\mathbf{Z}_s$ and $\mathbf{Z}_t$ with a polynomial kernel of degree 1. The resulting matrix is subsequently normalised to have a trace of one. The teacher entropy term in this loss is omitted since the teacher weights are fixed during training. Substituting the marginal and joint entropy definitions from equations 3 and 4, with $\mathbf{G}_{st} = \mathbf{G}_s \circ \mathbf{G}_t$ (normalised to have a trace of one), leads to

$$\mathcal{L}_{mi} = \frac{1}{1-\alpha}\log_2 \sum_{i=1}^{n} \lambda_i\left(\mathbf{G}_{st}^{\alpha}\right) - \frac{1}{1-\alpha}\log_2 \sum_{i=1}^{n} \lambda_i\left(\mathbf{G}_{s}^{\alpha}\right) \tag{8}$$

Where $\mathbf{G}_{st}$ is also normalised to have unit trace. Since computing the eigenvalues for lots of large matrices can be computationally expensive during training [16], we restrict our attention to $\alpha = 2$. This allows us to use the Frobenius norm as a proxy objective and one of which has a connection with the eigenspectrum - $\|\mathbf{A}_F\|^2 = tr(\mathbf{A}\mathbf{A}^H) = \sum_{i=1}^{n} \lambda_i(\mathbf{A}^2)$ since $\mathbf{A}$ is symmetric.

$$\mathcal{L}_{mi} = \log_2 \|\mathbf{G}_s\|_F^2 - \log_2 \|\mathbf{G}_{st}\|_F^2 \tag{9}$$

In practice, we observed that removing the *log* transformations improved the performance, thus resulting in a slight departure from the connection to mutual information. Specifically, the loss instead minimises the distance between the marginal and joint *information potential*, rather than the mutual information (see Supplementary).

4.3 Combining correlation and mutual information

Both the proposed losses provide two different learning objectives. Maximising the correlation is applied across the feature dimension, thus ensuring that the students average representation across the batch is perfectly correlated with the teacher. On the other hand, maximising the mutual information encourages the same similarity between samples as from the teacher. These two losses operate distinctly over the two dimensions of the representations, namely the *feature*-dim and the *batch*-dim. The final loss we aim to minimise is given as follows:

$$\mathcal{L}_{ITRD} = \mathcal{L}_{XE} + \beta_{corr}\mathcal{L}_{corr} + \beta_{mi}\mathcal{L}_{mi} \tag{10}$$

where $\mathcal{L}_{XE}$ is a cross-entropy loss, while β_{corr} and β_{mi} are hyperparameters to weight the losses. To demonstrate the simplicity of our proposed method, and similarly to past works [56], we provide the PyTorch-based pseudocode in algorithm 1.

```
# f_s: Student network
# f_t: Teacher network
# y: Ground-truth labels
# y_s, y_t: Student and teacher logits
# z_s, z_t: Student and teacher representations (n x d)
for x in loader:
  # Forward pass
  z_s, y_s = f_s(x)
  z_t, y_t = f_t(x)
  z_s = embed(z_s)
  # Cross entropy loss
  loss = cross_entropy(y_s, y)

  # Normalise representations
  z_s_norm = (z_s - z_s.mean(0)) / z_s.std(0)
  z_t_norm = (z_t - z_t.mean(0)) / z_t.std(0)
  # Compute cross-correlation vector
  v = einsum('bx,bx→x', z_s, z_t) / n
  # Compute correlation loss
  dist = torch.pow(v - torch.ones_like(v), 2)
  h_st = torch.log2(torch.pow(dist, alpha).sum())
  loss += h_st.mul(beta_corr)

  # Compute Gram matrices
  z_s_norm = normalize(z_s, p=2)
  z_t_norm = normalize(z_t, p=2)
  g_s = einsum('bx,dx→bd', z_s_norm, z_s_norm)
  g_t = einsum('bx,dx→bd', z_t_norm, z_t_norm)
  g_st = g_s * g_t
  # Normalize Gram matrices
  g_s = g_s / torch.trace(g_s)
  g_st = g_st / torch.trace(g_st)
  # Compute the mutual information loss
  p = g_s.pow(2) - g_st.pow(2)
  loss += p.sum().mul(beta_mi)

  # Optimisation step
  loss.backward()
  optimizer.step()
```

Algorithm 1: PyTorch-style pseudocode for ITRD

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 [18]	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
FitNet [34]	73.58	72.24	69.21	68.99	71.06	73.50	71.02	64.14	63.16	70.69	73.59	73.54	73.73
AT [53]	74.08	72.77	70.55	70.22	72.31	73.44	71.43	59.40	58.58	71.84	71.73	72.73	73.32
SP [48]	73.83	72.43	69.67	70.04	72.69	72.94	72.68	66.30	68.08	73.34	73.48	74.56	74.52
CC [46]	73.56	72.21	69.63	69.48	71.48	72.97	70.71	64.86	65.43	70.25	71.14	71.29	71.38
RKD [27]	73.35	72.22	69.61	69.25	71.82	71.90	71.48	64.52	64.43	71.50	72.28	73.21	72.21
PKT [25]	74.54	73.45	70.34	70.25	72.61	73.64	72.88	67.13	66.52	73.01	74.10	74.69	73.89
FT [17]	73.25	71.59	69.84	70.22	72.37	72.86	70.58	61.78	60.99	70.29	71.75	72.50	72.03
NST [15]	73.68	72.24	69.60	69.53	71.96	73.30	71.53	58.16	64.96	71.28	74.12	74.68	74.89
CRD [39]	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 [6]	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 [6]	**76.12**	75.09	71.89	-	73.89	75.63	74.84	70.37	69.89	-	**77.45**	**77.78**	77.14
$\mathcal{L}_{corr}$	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
$\mathcal{L}_{corr} + \mathcal{L}_{mi}$	**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

Table 1: 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. The mean and standard deviation was estimated over 3 runs. Same-architecture transfer experiments are highlighted in blue, whereas cross-architectural transfer is shown in red.

5 Experiments

We evaluate our proposed distillation across two standard benchmarks, namely the CIFAR-100 and ImageNet datasets. To further demonstrate the effectiveness of our loss, we perform additional experiments on the transferability of the students representations (see Supplementary), distilling from a full-precision model to a binary network, and on an NLP reading comprehension task. For all of these experiments, we jointly train the student model with an additional linear embedding for the student representation. This embedding is used for the correlation loss and is shared by the mutual information loss when there is a mismatch in dimensions between the student and the teacher.

5.1 Model compression

Experiments on CIFAR-100 classification [18] consist of 60K 32×32 RGB images across 100 classes with a 5:1 training/testing split. The results are shown in table 1 for multiple student-teacher pairs. For a fair comparison, we include those methods that use the standard CRD [39] teacher weights. The model abbreviations in the results table are given as follows: Wide residual networks (WRNd-w) [54], MobileNetV2 [10] (MN2), ShuffleNetV1 [57] / ShuffleNetV2 [38] (SN1 / SN2), and VGG13 / VGG8 [37] (V13 / V8). R32x4, R8x4, R110, R56 and R20 denote **CIFAR**-style residual networks, while R50 denotes an **ImageNet**-style ResNet50 [12]. CRCD [59] is not shown in table 1 since it uses different and not publicly available teacher weights[2]. Although both SSKD and HSAKD do provide official implementations and teacher weights, their use of self-supervision and additional auxiliary tasks is much more computationally expensive and orthogonal to our work. However, we do include these methods in the ImageNet experiment since the same teacher weights are used.

[2] In addition, using the unofficial code released by the authors, we were unable to replicate their reported results.

v.s.	ReviewKD	WCoRD	$\mathcal{L}_{corr}$
$\mathcal{L}_{corr}$	-3.7%	+16.2%	-
$\mathcal{L}_{corr} + \mathcal{L}_{mi}$	+6.8%	+24.4%	+10.5%

Table 2: Relative performance improvement (averaged over all architecture pairs in table 1) of the correlation and mutual information losses against ReviewKD, WCoRD and $\mathcal{L}_{corr}$ only.

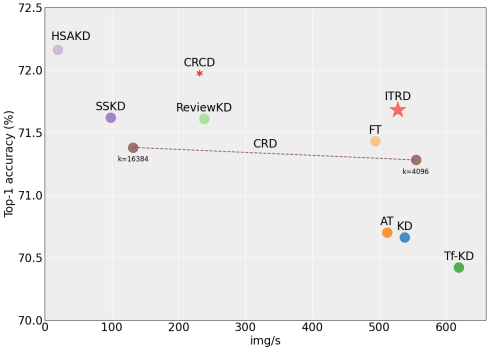

Figure 2: Top-1 Accuracy on ImageNet *vs* training efficiency with a ResNet-18 as the student and a pre-trained ResNet-34 as the teacher. For CRCD, the training efficiency was evaluated using the authors unofficial implementation, while the accuracy is reported in their paper.

For all experiments in table 1, we set $\beta_{corr} = 2.0$ and $\beta_{mi} = 1.0$ (or $\beta_{mi} = 0.0$ when only using $\mathcal{L}_{corr}$). For the correlation loss α, we use a value of 1.01 for the same architectures and 1.50 for the cross-architectures. ITRD achieves the best performance for 10 out of 13 of the architecture pairs, with a 6.8% and 24.4% relative improvement[3] over ReviewKD and WCoRD respectively. The addition of $\mathcal{L}_{mi}$ is also shown to complement the $\mathcal{L}_{corr}$ loss through a 10.5% average relative improvement over all pairs, as shown in table 2.

Experiments on ImageNet classification [32] involve 1.3 million images from 1000 different classes. In this experiment, we set the input size to 224×224, and follow a standard augmentation pipeline of cropping, random aspect ratio and horizontal flipping. We use the *torchdistill* library with standard settings, *i.e.*, 100 epochs of training using SGD with an initial learning rate of 0.1 that is divided by 10 at epochs 30, 60 and 90. The results are shown in figure 2 against the total training efficiency, which is measured in *img/s* and is inversely proportional to the total training time. This metric is evaluated using the official *torchdistill* implementations where possible. In the case of HSAKD, we used their official implementation and for CRCD we used the unofficial implementation provided by the authors. For a fair comparison, the batch sizes were scaled to ensure the training would fit within a predetermined memory constraint of 8GB, and we used for training an RTX 2080Ti GPU.

In terms of accuracy, ITRD achieves an error of 28.32%, being only behind CRCD and HSAKD, which are much more computationally costly through the use of either negative contrastive sampling and a gradient-based loss, or additional augmented training data. Conversely, ITRD is computationally efficient, with only a small overhead coming from a single linear layer that embeds the student and teacher representations to the same space, and from

[3]For clarity, we use the same definition for relative improvement as provided in WCoRD [5]. This is given by $\frac{X-Y}{X-KD}$, where the X method is compared to Y relative to standard KD with KL divergence.

Network	Method	Top-1 (%)
ResNet-18	Full Precision	94.8
	RAD [8]	90.5
	IR-Net [27]	91.5
	RBNN [20]	92.2
	ReCU [46]	92.8
	ReCU + CRD	92.1
	ReCU + ReviewKD	92.6
	ReCU + KD	93.3
	ReCU $+\mathcal{L}_{corr}$	93.9
	ReCU $+\mathcal{L}_{corr}+\mathcal{L}_{mi}$	**94.1**

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

Table 3: **Left:** Binary Network classification on CIFAR-10. **Right:** Question Answering on SQuAD 1.1. The teacher architecture, BERT, contains 12 layers, whereas the students, *T6* and *T3*, follow the same architecture as BERT but with 6 and 3 layers respectively.

computing the gram and cross-correlation matrices. The results show the applicability of ITRD to large-scale datasets, while being significantly more efficient and simple to adopt.

Binary neural networks (BNNs) [8, 20, 27, 46] are an extreme case of quantisation, where the weights can only represent two values. BNNs can obtain a significant model size reduction and increase of inference speed on CPUs [29] and FPGAs [42], with only a small drop in accuracy. We now show that ITRD can be used to reduce the gap between binary and full-precision (FP) networks. We use the state-of-the-art method ReCU [46] as our base model, and we distill the information from a FP teacher to our BNN student, which share the same architecture apart from the quantisation modules in the student. Table 3 shows the results, where for all distillation methods we used the same hyperparameters as in the previous experiments. 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.

NLP Question Answering. To show the wide applicability of our method, Table 3 shows the results of ITRD in a distillation task on the SQuAD 1.1 [28] reading comprehension task, using the transformer-based [43] BERT [7] as a teacher and modified versions of BERT with fewer layers as the students. For this experiment, we use the same hyperparameters used in the previous experiments, and following TextBrewer we apply ITRD to the output of each of the student transformer layers, and also use a standard KD [13] loss between the teacher and students logits. Table 3 shows that we outperform both NLP-specific distillation methods TextBrewer [9] and DistilBert [35] in both the Exact Match (EM) metrics and in F1 score.

6 Conclusion

In this work, we proposed an information-theoretic setting for representation distillation. Using this framework, we introduce novel distillation losses that are very simple and computationally inexpensive to adopt into most deep learning pipelines. Each of the proposed losses aims to extract complementary information from the teacher network. The correlation loss guides the student to match the teacher representation on a feature level. Conversely, the mutual information loss transfers the intra-batch similarity between samples from the teacher to the student. We have shown the superiority of our approach compared to methods of similar computational costs on standard classification benchmarks. Furthermore, we have shown the wide applicability of our method by reducing the gap between full-precision and binary networks, and also improving upon NLP-specific distillation methods.

Acknowledgement. This research was supported by UK EPSRC project EP/S032398/1.

References

[1] Madhu Advani, Artemy Kolchinsky, and Brendan D Tracey. On the information bottleneck theory of deep learning. 2019.

[2] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. *CVPR*, 2019.

[3] Joseph Bethge, Christian Bartz, Haojin Yang, Ying Chen, and Christoph Meinel. MeliusNet : Can Binary Neural Networks Achieve MobileNet-level Accuracy ?

[4] Rajendra Bhati. Infinitely Divisible Matrices. *Transactions of the American Mathematical Society*, 1969.

[5] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein Contrastive Representation Distillation. *CVPR*, 2020.

[6] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling Knowledge via Knowledge Review.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[8] Ruizhou Ding, Ting Wu Chin, Zeye Liu, and DIana Marculescu. Regularizing activation distribution for training binarized deep networks. *CVPR*, 2019.

[9] Yang et al. TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing. In *ACL System Demonstrations*, 2020.

[10] Michael H. Fox, Kyungmee Kim, and David Ehrenkrantz. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CVPR*, 2018.

[11] Jia Guo, Minghao Chen, Yao Hu, Chen Zhu, Xiaofei He, and Deng Cai. Reducing the Teacher-Student Gap via Spherical Knowledge Distillation. *arXiv preprint*, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. ResNet - Deep Residual Learning for Image Recognition. *CVPR*, 2015.

[13] Geoffrey Hinton, Sutskever Ilya, James Martens, and George Dahl. On the importance of initialization and momentum in deep learning. *ICML*, 2013.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NeurIPS*, 2015.

[15] Zehao Huang and Naiyan Wang. Like What You Like: Knowledge Distill via Neuron Selectivity Transfer. 2017.

[16] Andrew Kerr, Dan Campbell, and Mark Richards. QR decomposition on GPUs. *Proceedings of 2nd Workshop on General Purpose Processing on Graphics Processing Units, GPGPU-2*, 2009.

[17] Jangho Kim, Seong Uk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *NeurIPS*, 2018.

[18] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009.

[19] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning Filters For Efficient Convnets. *ICLR*, 2017.

[20] Mingbao Lin, Rongrong Ji, Zihan Xu, Baochang Zhang, Yan Wang, Yongjian Wu, Feiyue Huang, and Chia Wen Lin. Rotated binary neural network. *NeurIPS*, 2020.

[21] Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring Inter-Channel Correlation for Diversity-preserved Knowledge Distillation. *ICCV*, 2021.

[22] Roy Miles and Krystian Mikolajczyk. Cascaded channel pruning using hierarchical self-distillation. *BMVC*, 2020.

[23] Ilya Mironov. Rényi Differential Privacy. *Proceedings - IEEE Computer Security Foundations Symposium*, 2017. ISSN 19401434.

[24] Wonpyo Park, Kakao Corp, Dongju Kim, and Yan Lu. Relational Knowledge Distillation. *CVPR*, 2019.

[25] Nikolaos Passalis and Anastasios Tefas. Learning Deep Representations with Probabilistic Knowledge Transfer. *ECCV*, 2018.

[26] Baoyun Peng, Xiao Jin, Dongsheng Li, Shunfeng Zhou, Yichao Wu, Jiaheng Liu, Zhaoning Zhang, and Yu Liu. Correlation congruence for knowledge distillation. *CVPR*, 2019.

[27] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and Backward Information Retention for Accurate Binary Neural Networks. *CVPR*, 2020.

[28] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, 2016.

[29] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. *ECCV*, 2016.

[30] Alfréd Rényi. On Measures of Entropy and Information. *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 1960.

[31] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints For Thin Deep Nets. *ICLR*, 2015.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2014.

[33] Luis G. Sanchez Giraldo and Jose C. Principe. Information theoretic learning with infinitely divisible kernels. *ICLR*, 2013.

[34] 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.

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *NeurIPS*, 2019.

[36] B. H. Shekar, M. Sharmila Kumari, Leonid M. Mestetskiy, and Natalia F. Dyshkant. Face recognition using kernel entropy component analysis. *Neurocomputing*, 2011.

[37] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks For Large-scale Image Recognition. *ICLR*, 2015.

[38] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. MnasNet: Platform-Aware Neural Architecture Search for Mobile. *CVPR*, 2018.

[39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2019.

[40] Naftali Tishby. Deep Learning and the Information Bottleneck Principle.

[41] Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. *ICCV*, 2019.

[42] Yaman Umuroglu, Nicholas J. Fraser, Giulio Gambardella, Michaela Blott, Philip Leong, Magnus Jahre, and Kees Vissers. FINN: A framework for fast, scalable binarized neural network inference. In *FPGA 2017*, 2017.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] Paul L. Williams and Randall D. Beer. Nonnegative Decomposition of Multivariate Information. 2010.

[45] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge Distillation Meets Self-supervision. *ECCV*, 2020.

[46] Zihan Xu, Mingbao Lin, Jianzhuang Liu, Jie Chen, Ling Shao, Yue Gao, Yonghong Tian, and Rongrong Ji. ReCU: Reviving the Dead Weights in Binary Neural Networks. *ICCV*, 2021.

[47] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Hierarchical Self-supervised Augmented Knowledge Distillation. *IJCAI*, 2021.

[48] Junho Yim. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. *CVPR*, 2017.

[49] Jiahui Yu and Thomas Huang. Universally Slimmable Networks and Improved Training Techniques. *ICCV*, 2019.

[50] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable Neural Networks. *ICLR*, 2018.

[51] Shujian Yu and José C. Príncipe. Understanding autoencoders with information theoretic concepts. *Neural Networks*, 2019.

[52] Shujian Yu, Kristoffer Wickstrom, Robert Jenssen, and Jose C. Principe. Understanding Convolutional Neural Networks With Information Theory: An Initial Exploration. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[53] Xi Yu, Shujian Yu, and José C. Príncipe. Deep deterministic information bottleneck with matrix-based entropy functional. *ICASSP*, 2021.

[54] Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. *BMVC*, 2016.

[55] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2019.

[56] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *ICML*, 2021.

[57] Xiangyu Zhang, Xinyu Zhou, and Mengxiao Lin. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *CVPR*, 2018.

[58] Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, and Yan Bai. Prime-Aware Adaptive Distillation. pages 1–17.

[59] Jinguo Zhu, Shixiang Tang, Dapeng Chen, and Shijie Yu. Complementary Relation Contrastive Distillation.

[60] Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, and Jinhui Zhu. Discrimination-aware Channel Pruning for Deep Neural Networks. *NeurIPS*, 2018.