Quantifying Knowledge Distillation Using Partial Information Decomposition

Pasan Dissanayake[1*] **Faisal Hamman**[1] **Barproda Halder**[1] **Ilia Sucholutsky**[2]
Qiuyi Zhang[3] **Sanghamitra Dutta**[1]

[1]University of Maryland   [2]Princeton University   [3]Google Research

Abstract

Knowledge distillation deploys complex machine learning models in resource-constrained environments by training a smaller student model to emulate internal representations of a complex teacher model. However, the teacher's representations can also encode nuisance or additional information not relevant to the downstream task. Distilling such irrelevant information can actually impede the performance of a capacity-limited student model. This observation motivates our primary question: *What are the information-theoretic limits of knowledge distillation?* To this end, we leverage Partial Information Decomposition to quantify and explain the transferred knowledge and knowledge left to distill for a downstream task. We theoretically demonstrate that the task-relevant transferred knowledge is succinctly captured by the measure of *redundant* information about the task between the teacher and student. We propose a novel multi-level optimization to incorporate redundant information as a regularizer, leading to our framework of *Redundant Information Distillation (RID)*. RID leads to more resilient and effective distillation under nuisance teachers as it succinctly quantifies task-relevant knowledge rather than *simply aligning student and teacher representations*.

1 INTRODUCTION

Modern-day machine learning requires large amounts of compute for both training and inference. Knowledge distillation (Hinton, 2015; Romero et al., 2015) can be used to compress a complex machine learning model (the teacher) by distilling it into a relatively simpler model (the student). The term "distillation" in this context means obtaining some assistance from the teacher while training the student so that the student performs much better than when trained alone (see Figure 1). In its earliest forms, knowledge distillation involved the student trying to match the output logits of the teacher (Hinton, 2015). More advanced methods focus on distilling multiple intermediate representations of the teacher to the corresponding layers of the student (Romero et al., 2015; Ahn et al., 2019; Tian et al., 2020; Liang et al., 2023). We also refer to Gou et al. (2021); Sucholutsky et al. (2023) for surveys.

Information theory has been instrumental in both designing (Ahn et al., 2019; Tian et al., 2020) and explaining (Zhang et al., 2022; Wang et al., 2022) knowledge distillation techniques. However, less attention has been given to characterizing the fundamental limits of the process from an information-theoretical perspective. Our goal is to bridge this gap by *first introducing new measures to quantify the "transferred knowledge" and "knowledge to distill" for a teacher and student model given a target downstream task*. We bring in an emerging body of work called Partial Information Decomposition (PID) (Williams and Beer, 2010; Griffith et al., 2014; Bertschinger et al., 2014) to explain knowledge distillation. We define the knowledge to distill using the PID measure of "unique" information about the task that is available only with the teacher but not the student. As it follows, the transferred knowledge is succinctly quantified by the measure of "redundant" information common between the teacher and student.

We propose a multi-level optimization that maximizes redundant information (transferred knowledge) as a regularizer for more effective distillation. While PID has been explored in a few avenues of machine learning, it has remained a challenge to maximize these measures as a regularizer since computing them itself requires solving an optimization. Our optimization

---

leads to a novel knowledge distillation framework – Redundant Information Distillation (RID) – which precisely captures task-relevant knowledge and filters out the task-irrelevant information from the teacher. In summary, our main contributions are as follows:

- **Quantifying transferred knowledge and knowledge to distill:** Given a downstream task and a teacher and a student model, we formally define the knowledge to distill as the unique information in the teacher (Definition 3.1) and the transferred knowledge as the redundant information (Definition 3.2). Through examples and theoretical results (Theorem 3.1), we first show that redundant information succinctly captures task-related knowledge transferred to the student as opposed to existing frameworks which directly align teacher (T) and student (S) representations, e.g., VID (Ahn et al., 2019) maximizes the mutual information $I(T; S)$. Theorem 3.1 points out a fundamental limitation of existing knowledge distillation frameworks for capacity-limited students: they blindly align the student and teacher without precisely capturing task-related knowledge.

- **Maximizing redundant information as a regularizer:** To alleviate this limitation, we propose a strategy to incorporate redundant information as a regularizer during model distillation to precisely maximize the task-relevant transferred knowledge. We first circumvent the challenge of computing the redundant information measure proposed in (Bertschinger et al., 2014) by utilizing the quantity termed intersection information defined in Griffith and Ho (2015), which we prove (in Theorem 4.1) to be a lower-bound for redundant information. The significance of Theorem 4.1 is that it enables us to obtain an optimization formulation to maximize a lower-bound of redundant information without making distributional assumptions, a contribution that is also of independent interest outside the domain of knowledge distillation.

- **A novel knowledge distillation framework:** We propose a new framework called Redundant Information Distillation (RID) whose distillation loss is tailored to maximize the redundant information (i.e., the transferred knowledge as per Definition 3.2). We carry out a number of experiments to demonstrate the advantage of this new framework over VID (Ahn et al., 2019), an existing knowledge distillation framework that maximizes the mutual information between the teacher and student representations (Section 5). Experiments are carried out for the CIFAR10 and CIFAR100 datasets, as well as for a transfer learning setting where the teacher is trained on the ImageNet dataset and transferred to students over the CUB-200-2011 dataset. Our framework explains knowl-
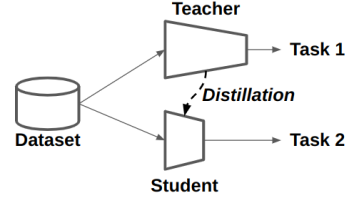

Figure 1: Knowledge Distillation: The teacher (a complex model) assists the student (usually a substantially simpler model) during their training. The learned student can perform much better than an independently trained student without distillation with a similar training setup (i.e., hyper-parameters and data). The teacher may or may not have been trained for the same task as the student.

edge distillation and shows more resilience under less-informative nuisance teachers.

Related Works: Multi-layer knowledge distillation was introduced in FitNets (Romero et al., 2015). Henceforth, a large number of techniques, based on different statistics derived for matching a teacher-student pair, have been proposed. In particular, Ahn et al. (2019); Tian et al. (2020); Chen et al. (2021); Miles et al. (2021) leverage an information-theoretic perspective to arrive at a solution (also see surveys Gou et al. (2021); Sucholutsky et al. (2023)). In this paper, we focus on Variational Information Distillation (VID, Ahn et al. (2019)) as a representative framework of the larger class of distillation frameworks which maximizes $I(T; S)$ as the distillation strategy. We also discuss Task-aware Layer-wise Distillation (TED, Liang et al. (2023)) as a framework that filters out task-related information. Specifically, Liang et al. (2023) highlight the importance of distilling only the task-related information when there is a significant complexity gap between the teacher and the student. Towards this end, Zhu et al. (2022) points out the existence of non-distillable classes due to the unmatched capacity of the student model. We discuss some more related works on knowledge distillation (Kundu et al., 2021; Park et al., 2021; Ben-Baruch et al., 2022; Jiao et al., 2019; Huang et al., 2023) in Appendix A.

Information theory has been instrumental towards explaining the success of knowledge distillation. Wang et al. (2022) utilizes information bottleneck principles (Tishby et al., 2000; Tishby and Zaslavsky, 2015) to explain how a teacher model may assist the student in learning relevant features quickly. Zhang et al. (2022) observes the training process as systematically discarding knowledge from the input. Accordingly, the distillation helps the student to quickly learn what information to discard. Despite these attempts, we observe that
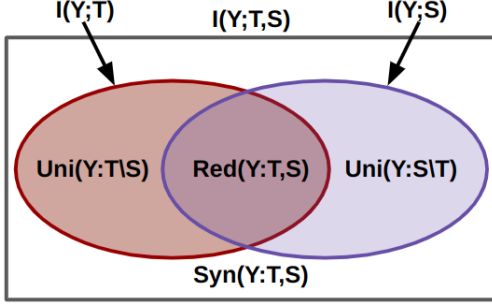
Figure 2: Partial Information Decomposition: The box denotes the total joint mutual information $I(Y;T,S)$, which is decomposed into four non-negative terms named synergistic information $Syn(Y:T,S)$, redundant information $Red(Y:T,S)$ and the two unique information terms $Uni(Y:T\backslash S)$ and $Uni(Y:S\backslash T)$.

there exists a gap in characterizing the fundamental limits of knowledge distillation which we seek to address using the mathematical tool of PID.

PID is also beginning to generate interest in other areas of machine learning (Dutta et al., 2020, 2021; Dutta and Hamman, 2023; Hamman and Dutta, 2024a; Liang et al., 2024a,b; Hamman and Dutta, 2024b; Tax et al., 2017; Ehrlich et al., 2022; Wollstadt et al., 2023; Mohamadi et al., 2023; Venkatesh et al., 2024; Halder et al., 2024; Dewan et al., 2024; Venkatesh and Schamberg, 2022; Venkatesh et al., 2024; Goswami and Merkley, 2024; Kleinman et al., 2021; Lyu et al., 2024). However, it has not been leveraged in the context of knowledge distillation before. Additionally, while most related works predominantly focus on efficiently computing PID, e.g., Kleinman et al. (2021); Liang et al. (2024a); Halder et al. (2024); Pakman et al. (2021) that itself requires solving an optimization over the joint distribution, *there are limited works that further incorporate it as a regularizer during model training.* Dutta et al. (2021) leverages Gaussian assumptions to obtain closed-form expressions for the PID terms, enabling them to use unique information as a regularizer during training for fairness (also see Venkatesh and Schamberg (2022); Venkatesh et al. (2024) for more details on Gaussian PID). Our work makes novel connections between two notions of redundant information and demonstrates how PID can be integrated as a regularizer in a multi-level optimization without Gaussian assumptions, which could also be of independent interest outside the context of knowledge distillation.

2 PRELIMINARIES

Background on PID: Partial Information Decomposition (PID), first introduced in Williams and Beer

(2010), offers a way to decompose the joint information in two sources, say T and S, about another random variable Y (i.e., $I(Y;T,S)$ where $I(A;B)$ denotes the mutual information between A and B (Cover and Thomas, 2006)) into four components as follows:

1. Unique information $Uni(Y : T\backslash S)$ and $Uni(Y : S\backslash T)$: information about Y that each source uniquely contains
2. Redundant information $Red(Y : T, S)$: the information about Y that both T and S share
3. Synergistic information $Syn(Y : T, S)$: the information about Y that can be recovered only by using both T and S.

See Figure 2 for a graphical representation. These PID components satisfy the relationships given below:

$$I(Y;T,S) = Uni(Y : T\backslash S) + Uni(Y : S\backslash T)$$
$$+ Red(Y : T, S) + Syn(Y : T, S) \quad (1)$$
$$I(Y;T) = Uni(Y : T\backslash S) + Red(Y : T, S) \quad (2)$$
$$I(Y;S) = Uni(Y : S\backslash T) + Red(Y : T, S). \quad (3)$$

While this system of equations cannot be solved to arrive at a deterministic definition for each PID term, defining only one of the terms is sufficient to define the rest. Consequently, a wide array of definitions exists, each based on different desired properties (Williams and Beer, 2010; Bertschinger et al., 2014; Griffith et al., 2014; Griffith and Ho, 2015). Among these, the definition proposed in Bertschinger et al. (2014) is motivated with an operational interpretation of unique information from decision theory. Moving on to the context of knowledge distillation, we map T to be the teacher representation, S to be the student representation, and Y to be the downstream task that the student is being trained for. That makes $I(Y;T)$ and $I(Y;S)$ be the total knowledge about Y that is in the teacher and in the student, respectively.

Notation and Problem Setting: We consider a layer-wise distillation scheme where the teacher representation $T(X)$ is distilled into the student representation $S_{\eta_s}(X)$, where X is the input. The target of the student is to predict the task Y from X. Both $T(\cdot)$ and $S_{\eta_s}(\cdot)$ are deterministic functions of X and the randomness is due to the input being random. Note that the student representation depends on the parameters of the student network denoted by η_s and hence written as S_{η_s}. However, when this parameterization and dependence on X is irrelevant or obvious, we may omit both and simply write T and S. We denote the supports of Y, T and S by $\mathcal{Y}, \mathcal{T}$ and $\mathcal{S}$, respectively. In general, upper-case letters denote random variables, except P and Q which represent probability distributions, C, H, W which stand for the representation dimensions, and K which indicates the number of layers distilled.

Lowercase letters are used for vectors unless specified otherwise. Lowercase Greek letters denote the parameters of neural networks.

Knowledge distillation is achieved by modifying the student loss function to include a distillation loss term in addition to the ordinary task-related loss as follows:

$$\mathcal{L}(\eta_s) = \lambda_1 \mathcal{L}_{\text{ordinary}}(Y, \hat{Y}(X))$$
$$+ \lambda_2 \mathcal{L}_{\text{distill}}(Y, \hat{Y}(X), S_{\eta_s}, T). \quad (4)$$

Here, $\lambda_1, \lambda_2 > 0$. When the task at hand is a classification, Y denotes the true class label and we use the cross entropy loss defined as $\mathcal{L}_{CE}(Y, \hat{Y}) = -\mathbb{E}_{P_X}\left[\log P_{\hat{Y}(X)}(Y)\right]$ as the ordinary task-related loss for the student. Here, $\hat{Y}(X)$ is the student's final prediction of Y. The teacher network is assumed to remain unmodified during the distillation process.

3 EXPLAINING KNOWLEDGE DISTILLATION

In this section, we propose information theoretic metrics to quantify both the task-relevant information that is available in the teacher for distillation, and the amount of information that has already been transferred to the student. We mathematically demonstrate favorable properties of our proposed measures in comparison to other candidate measures. Our mathematical results highlight the limitations of existing knowledge distillation frameworks that naively align the student with the teacher with no regard for task-relevance.

Definition 3.1 (Knowledge to distill). *Let Y, S, and T be the target variable, the student's intermediate representation, and the teacher's intermediate representation, respectively. The knowledge to distill from T to S is defined as $Uni(Y : T\backslash S)$, the unique information about Y that is in T but not in S.*

With the knowledge to distill is defined as the unique information $Uni(Y : T\backslash S)$, we see that the more the distillation happens, the more the $Uni(Y : T\backslash S)$ shrinks. Note that under the knowledge distillation setting, the total knowledge of the teacher $I(Y : T)$ is constant since the teacher is not modified during the process. Since $I(Y; T) = Uni(Y : T\backslash S) + Red(Y : T, S)$ we therefore propose $Red(Y : T, S)$ as a measure for knowledge that has already been transferred.

Definition 3.2 (Transferred knowledge). *Let Y, S, and T be the target variable, student's intermediate representation, and the teacher's intermediate representation, respectively. The transferred knowledge from T to S is defined as $Red(Y : T, S)$, the redundant information about Y between T and S.*

We leverage the unique and redundant information definitions from Bertschinger et al. (2014) for an exact quantification of these quantities.

Definition 3.3 (Unique and redundant information (Bertschinger et al., 2014)). *Let P be the joint distribution of Y, T and S, and Δ be the set of all joint distributions over $\mathcal{Y} \times \mathcal{T} \times \mathcal{S}$. Then,*

$$Uni(Y : T\backslash S) := \min_{Q \in \Delta_P} I_Q(Y; T|S) \quad (5)$$

$$Red(Y : T, S) := I(Y; T) - \min_{Q \in \Delta_P} I_Q(Y; T|S) \quad (6)$$

where $\Delta_P = \{Q \in \Delta : Q(Y = y, T = t) = P(Y = y, T = t), Q(Y = y, S = s) = P(Y = y, S = s) \forall y \in \mathcal{Y}, t \in \mathcal{T}$ and $s \in \mathcal{S}\}$ i.e., Δ_P is the set of all joint distributions with marginals of the pairs (Y, T) and (Y, S) equal to that of P.

Comparison to Existing Approaches for Knowledge Distillation: A multitude of knowledge distillation frameworks exist which are based on maximizing the mutual information between the teacher and the student (i.e., $I(T; S)$) Ahn et al. (2019); Tian et al. (2020); Chen et al. (2021); Miles et al. (2021). While a distillation loss that maximizes $I(T; S)$ can be helpful to the student when the teacher possesses task-related information, we show that it creates a tension with the ordinary loss when the teacher has little or no task-relevant information. Moreover, even though the teacher contains task-related information, the limited capacity of the student may hinder a proper distillation when this kind of framework is used. The following examples provide critical insights, exposing the limitation of $I(T; S)$. Our proposed measure $Red(Y : T, S)$ resolves these cases in an explainable manner by succinctly capturing task-relevant knowledge.

Example 1: (Uninformative teacher) An uninformative teacher representation (i.e., T with $I(Y; T) = 0$) gives $Uni(Y : T\backslash S) = Red(Y : T, S) = 0$ for any S, agreeing with the intuition. Hence, an algorithm that maximizes exactly the transferred knowledge $Red(Y : T, S)$ will have a zero gradient over this term. In contrast, algorithms that maximize the similarity between S and T quantified by $I(T; S)$ will force S to mimic the uninformative teacher, causing a performance worse than ordinary training without distillation. As a simplified example, let $U_1, U_2 \sim Ber(0.5)$ and $Y = U_1, T = U_2$. Then, the teacher cannot predict the intended task Y. Note that, in this case, $I(T; S)$ is not maximized when the student representation is $S = Y$. Instead, it is maximized when $S = U_2$.

Example 2: (Extra complex teacher) Let $U_1 \sim Ber(0.2), U_2 \sim Ber(0.5)$ and $Y = U_1, T = (U_1, U_2)$. Then, the teacher can completely predict the intended task Y. Assume the student is simpler than the

teacher and has only one binary output. In this situation, $I(T; S)$ is not maximized when $S = U_1$ because $I((U_1, U_2) : U_1) \approx 0.72 < 1 = I((U_1, U_2) : U_2)$ where the right-hand side is achieved when $S = U_2$. However, $S = U_1$ is a maximizer for $Red(Y : T, S)$ (i.e., $Red(Y : T, S) = Red(U_1 : T, U_1) = I(Y; T)$). Theorem 3.1 presents a more general case.

Theorem 3.1 formally exposes the limitations of maximizing $I(T; S)$ for capacity-limited students as $I(T; S)$ does not emphasize task-related knowledge.

Theorem 3.1 (Teacher with nuisance). *Let $T = (Z, G)$ where Z contains all the task-related information (i.e., $I(Y; T) = I(Y; Z)$) and G does not contain any information about the task (i.e., $I(Y; G) = 0$). Let the student be a capacity-limited model as defined by $H(S) \leq \max\{H(Z), H(G)\}$ where $H(X)$ denotes the Shannon entropy of the random variable X. Then,*

(i) $I(T; S)$ is maximized when

$$S = \begin{cases} Z & ; \quad H(Z) > H(G) \\ G & ; \quad H(Z) < H(G) \end{cases}. \tag{7}$$

(ii) $Red(Y : T, S)$ is always maximized when $S = Z$.

The uninformative random variable G here can be seen as a stronger version of the nuisance defined in (Achille and Soatto, 2018, Section 2.2). In the above scenario, the task-related part of the student loss will have a tension with the distillation loss when $H(Z) < H(G)$, in which case, the distillation actually adversely affects the student. On the other hand, a distillation loss that maximizes $Red(Y : T, S)$ will always be aligned with the task-related loss.

These examples show that the frameworks based on maximizing $I(T; S)$ are not capable of selectively distilling the task-related information to the student. In an extreme case, they are not robust to being distilled from a corrupted teacher network. This is demonstrated in the experiments under Section 5.

It may appear that using information gain (conditional mutual information) $I(Y; T|S) = I(Y; T, S) - I(Y; S)$ as the measure for the amount of knowledge available to distill resolves the cases similar to Example 1. However, Example 3 below provides a counterexample.

Example 3: (Effect of synergy) Consider a scenario similar to Example 1, where the teacher is uninformative regarding the interested task. For example, let $U_1, U_2 \sim Ber(0.5)$ and $Y = U_1, T = U_1 \oplus U_2$ where $\oplus$ denotes the binary XOR operation. Suppose we were to consider conditional mutual information $I(Y; T|S)$ as the measure of the amount of knowledge available to distill information available in the teacher. Then,

$I(Y; T|S) = H(Y)$ when $S = U_2$, indicating non-zero knowledge available to distill in the teacher. This is unintuitive since in this case both $I(Y; T) = I(Y; S) = 0$ and neither the teacher nor the student can be used alone to predict Y. In contrast, the proposed measures give $Uni(Y : T \backslash S) = Red(Y : T, S) = 0$ indicating no available or transferred knowledge.

Next, we present Theorem 3.2 which highlights some important properties of the proposed metrics. These properties indicate that the proposed measures agree well with our intuition for explaining distillation.

Theorem 3.2 (Properties). *The following properties hold for knowledge to distill and already transferred knowledge defined as in Definition 3.1 and Definition 3.2 respectively.*

(i) $Uni(Y : T \backslash S)$ and $Red(Y : T, S)$ are non-negative.

(ii) When $Uni(Y : T \backslash S) = 0$, the teacher has zero knowledge available for distillation. At this point, the student has the maximum information that any one of the representations T or S has about Y; i.e.,

$$\max\{I(Y; T), I(Y; S)\} = I(Y; S). \tag{8}$$

(iii) For a given student representation S and any two teacher representations T_1 and T_2 if there exists a deterministic mapping h such that $T_1 = h(T_2)$, then $Uni(Y : T_1 \backslash S) \leq Uni(Y : T_2 \backslash S)$.

4 A FRAMEWORK FOR MAXIMIZING TRANSFERRED KNOWLEDGE

In this section, we propose a distillation framework – Redundant Information Distillation (RID) – which maximizes the transferred knowledge $Red(Y : T, S)$ with a focus on classification problems. We first show that our measure of transferred knowledge is lower-bounded by an alternative definition of redundant information (also called the intersection information – denoted by $Red_\cap(Y : T, S)$). Next, we leverage this lower-bound to develop a multi-level optimization framework to selectively distill task-relevant knowledge.

Definition 4.1 (I_α measure (Griffith and Ho, 2015)).

$$Red_\cap(Y : T, S) = \max_{P(Q|Y)} I(Y : Q)$$

$$such \ that \quad I(Y; Q|f_t(T)) = I(Y; Q|f_s(S)) = 0. \tag{9}$$

Next we show, from Theorem 4.1, that $Red_\cap(Y : T, S)$ is a lower-bound for $Red(Y : T, S)$. The significance of this result is that maximizing the lower-bound $Red_\cap(Y : T, S)$ would also increase $Red(Y : T, S)$.

(a) Phase 1: Train the teacher's filters while the student is kept frozen

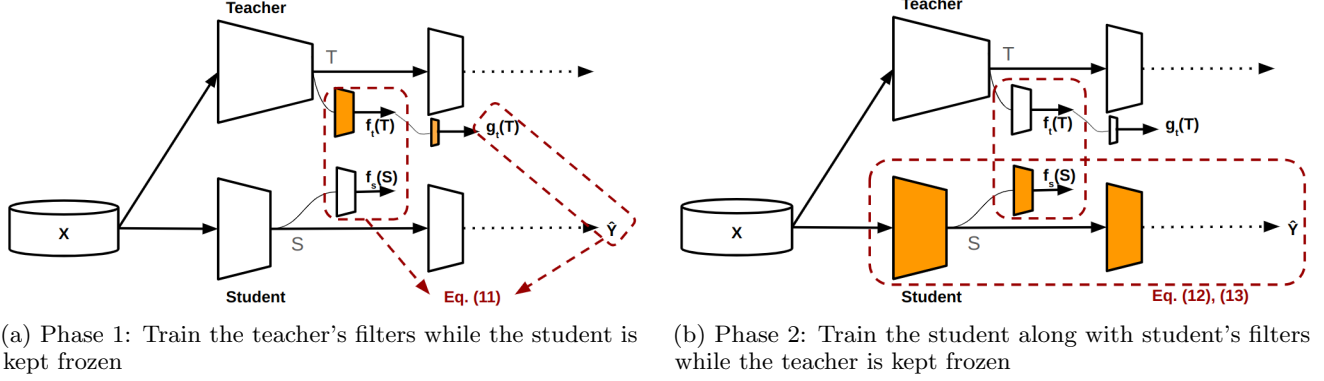(b) Phase 2: Train the student along with student's filters while the teacher is kept frozen

Figure 3: Redundant Information Distillation (RID) framework: $f_t(\cdot)$ and $f_s(\cdot)$ are the teacher and the student filter outputs, respectively. A classification head $g_t(\cdot)$ is appended to the teacher's filter during the first phase. Highlighted in amber are the components that are being updated in each phase.

Theorem 4.1 (Transferred knowledge lower bound). *For any three random variables Y, T and S,*

$$Red_{\cap}(Y:T,S) \leq Red(Y:T,S) \qquad (10)$$

where $Red_{\cap}(Y:T,S)$ is as per Definition 4.1 and $Red(Y:T,S)$ is defined in Definition 3.3.

Next, we discuss our proposed approach for maximizing the lower-bound $Red_{\cap}(Y:T,S)$ (also see Figure 3 and Algorithm 1 for our complete strategy). Our proposed framework is based on selecting Q in Definition 4.1 to be $Q = f_t(T)$, and parameterizing $f_t(\cdot)$ and $f_s(\cdot)$ using small neural networks. To denote the parameterization, we will occasionally use the elaborated notation $f_t(\cdot; \theta_t)$ and $f_s(\cdot; \theta_s)$, where θ_t and θ_s denote the parameters of f_t and f_s, respectively. With the substitution of $Q = f_t(T)$, Definition 4.1 results in the following optimization problem:

$$\max_{\theta_t, \theta_s, \eta_s} I(Y : f_t(T; \theta_t))$$
$$\text{such that} \quad I(Y; f_t(T; \theta_t) | f_s(S_{\eta_s}; \theta_s)) = 0. \quad (\text{P1})$$

We divide the problem (P1) into two phases and employ gradient descent on two carefully designed loss functions to perform the optimization. In the first phase, we maximize the objective w.r.t. θ_t while θ_s and S are kept constant (recall that T is fixed in all cases because the teacher is not being trained during the process). For this, we append an additional classification head $g_t(\cdot; \phi_t)$ parameterized by ϕ_t to the teacher's task aware filter f_t. Then we minimize the loss function given below with respect to θ_t and ϕ_t using gradient descent.

$$\mathcal{L}_t(\theta_t, \phi_t) = \mathcal{L}_{CE}(Y, g_t(f_t(T; \theta_t); \phi_t))$$
$$+ \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbb{E}_{P_X} \left[\frac{V_{c,h,w}^2}{\sigma_c} \right] \quad (11)$$

where $V_{c,h,w}$ denotes the corresponding element of $V = f_t(T(X); \theta_t) - f_s(S(X); \theta_s) \in \mathbb{R}^{C \times H \times W}$. Here, $C, H,$ and W are the number of channels, height, and width of the outputs of f_s and f_t. $\sigma = [\sigma_1, \dots, \sigma_C]^T$ is a stand-alone vector of weights that are optimized in the second phase. Minimizing the cross-entropy term $\mathcal{L}_{CE}(Y, g_t(f_t(T; \theta_t); \phi_t))$ of $\mathcal{L}_t(\theta_t, \phi_t)$ above amounts to maximizing $I(Y; f_t(T; \theta_t))$. The second term prohibits $f_t(T)$ from diverting too far from $f_s(S)$ during the process, so that the constraint $I(Y; f_t(T; \theta_t) | f_s(S; \theta_s)) = 0$ can be ensured.

During the second phase, we freeze θ_t and maximize the objective over θ_s, S_{η_s} and σ. The loss function employed in this phase is as follows:

$$\mathcal{L}(\theta_s, \sigma, \eta_s) = \lambda_1 \mathcal{L}_{CE}(Y, \hat{Y}_{\eta_s}) + \lambda_2 \mathcal{L}_s(\theta_s, \sigma, \eta_s) \quad (12)$$

where

$$\mathcal{L}_s(\theta_s, \sigma, \eta_s) = \left(||\sigma||^2 + \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbb{E}_{P_X} \left[\frac{V_{c,h,w}^2}{\sigma_c} \right] \right) \quad (13)$$

and λ_1 and λ_2 are scalar hyperparameters which determine the prominence of ordinary learning (through $\mathcal{L}_{CE}(Y, \hat{Y}_{\eta_s})$) and distillation (through $\mathcal{L}_s(\theta_s, \sigma, \eta_s)$). V and σ are as defined earlier. $\hat{Y}_{\eta_s}$ denotes the final prediction of the student network.

The first term of the loss function is the ordinary task-related loss. The next two terms correspond to the distillation loss, which is our focus in the following explanation. Consider phase 2 as an estimation problem that minimizes the σ-weighted mean squared error, where $Q = f_t(T)$ is the estimand and $f_s(\cdot)$ is the estimator. The magnitudes of the positive weights σ are controlled using the term $||\sigma||^2$. We observe that this optimization ensures $I(Y; Q | f_s(S)) = 0$ given that the following assumption holds.

Algorithm 1: Redundant Information Distillation

Data: A dataset of samples of (X,Y), teacher model with intermediate representations $T^{(1)}, \ldots, T^{(k)}$, hyperparameters $\lambda_1, \lambda_2 > 0$, # warm-up epochs n_w, # training epochs n, # steps per cycle $q \leq n$, alternating ratio $r(0 < r < 1)$

Result: Trained student network parameterized with η_s

Initialize parameters $\theta_t^{(k)}, \theta_s^{(k)}, \phi_t^{(k)}$ and η_s;

for $i \in \{1, \ldots, n_w\}$ **do**

$\quad$ $\underset{\{\theta_t^{(k)}, \phi_t^{(k)}\}}{\text{minimize}} \sum_{k=1}^{K} \mathcal{L}_{CE}(Y, g_t^{(k)}(f_t^{(k)}(T^{(k)}; \theta_t^{(k)}); \phi_t^{(k)}))$

$\quad$;

end

for $i \in \{1, \ldots, n\}$ **do**

$\quad$ **if** $round(i/q) < q \times r$ **then**

$\quad\quad$ $\underset{\{\theta_t^{(k)}, \phi_t^{(k)}\}}{\text{minimize}} \sum_{k=1}^{K} \mathcal{L}_t(\theta_t^{(k)}, \phi_t^{(k)})$; /*Eq. (11)*/

$\quad$ **else**

$\quad\quad$ $\underset{\{\theta_s^{(k)}, \sigma^{(k)}, \eta_s\}}{\text{minimize}} \lambda_1 \mathcal{L}_{CE}(Y, \hat{Y}_{\eta_s}) +$

$\quad\quad$ $\lambda_2 \sum_{k=1}^{K} \mathcal{L}_s(\theta_s^{(k)}, \sigma^{(k)}, \eta_s)$; /*Eq. (13)*/

$\quad$ **end**

end

Assumption: *Let the estimation error be* $\epsilon = f_t(T) - f_s(S)$. *Assume* $I(\epsilon; Y|f_s(S)) = 0$, *i.e., given the estimate, the estimation error is independent of* Y.

With the above assumption, we see that

$$I(Y; Q|f_s(S)) = I(Y; f_s(S) + \epsilon|f_s(S))$$
$$= I(Y; \epsilon|f_s(S)) = 0. \quad (14)$$

Therefore, the constraint in problem P1 is satisfied by this selection of random variables. Therefore, along with the maximization of $I(Y; Q)$ during phase 1, the proposed framework can be seen as performing the optimization in Definition 4.1 in two steps.

This, along with Theorem 4.1, completes our claim that the proposed framework maximizes a lower bound for the transferred knowledge. RID loss terms can be extended to multiple layers as follows: Replace the single term $\mathcal{L}_t(\theta_t, \phi_t)$ in (11) with a sum of terms $\sum_{k=1}^{K} \mathcal{L}_t^{(k)}(\theta_t^{(k)}, \phi_t^{(k)})$ corresponding to each pair of layers $T^{(k)}$ and $S_{\eta_s}^{(k)}$ where $k = 1, \ldots, K$. Similarly, replace $\mathcal{L}_s(\theta_s, \sigma, \eta_s)$ in (13) with $\sum_{k=1}^{K} \mathcal{L}_s^{(k)}(\theta_s^{(k)}, \sigma^{(k)}, \eta_s)$. The framework is summarized in Algorithm 1. The advantage of RID over the VID framework (Ahn et al., 2019) which maximizes $I(T; S)$ can be observed in the exper-

iments that are detailed in Section 5.

Remark. *We note that the Large Language Model (LLM) fine-tuning framework of Task-aware Layer-wise Distillation (TED) (Liang et al., 2023) shares intuitive similarities with RID regarding distilling task-related knowledge. However, they take a heuristic approach when designing the framework. In fact, our mathematical formulation can explain the success of TED as detailed in Appendix C. In addition to the application domain, the difference between TED and RID can mainly be attributed to the following:*

(i) During the first stage, TED trains both $f_t(\cdot)$ and $f_s(\cdot)$ whereas RID only trains $f_t(\cdot)$.

(ii) In the second stage loss, TED includes an ordinary mean squared error term whereas RID includes a weighted (using σ) mean squared error term.

To the best of our knowledge, our work is the first to information-theoretically quantify the actual task-relevant transferred knowledge and formally incorporate it into an optimization.

5 EMPIRICAL VALIDATION

Setup: We compare the performance of the proposed RID framework with (i) VID; and (ii) our adaptation of TED in this domain under two different conditions. In the first setting, the teacher network is fully trained with the complete training set, whereas in the second setting, the teacher is just randomly initialized without any training at all. Experiments are carried out on CIFAR-10 (Krizhevsky et al., 2009b), CIFAR-100 (Krizhevsky et al., 2009a), ImageNet (Deng et al., 2009) and CUB-200-2011 (Wah et al., 2011) datasets (with the last two being used in a transfer learning setup). Additionally, we train a student without any knowledge distillation, which we label as "BAS". Table 1 provides details of the model architectures used in each setting.

Table 1: Model architectures

Dataset	Teacher	Student
CIFAR-10	WRN-(40,2)	WRN-(16,1)
CIFAR-100	WRN-(28,10)	WRN-(16,8)
ImageNet to CUB-200-2011	ResNet-34	ResNet-18

We distill three layers from the teacher to the corresponding student layers. In the case of RID, each teacher layer $T^{(k)}$ has its own filter $f_t^{(k)}$ parameterized with $\theta_t^{(k)}$. Student filters are parameterized in a similar manner. Moreover, each teacher filter $f_t^{(k)}(\cdot)$

has its own classification head $g_t^{(k)}(\cdot)$ parameterized with $\phi^{(k)}$. All the student representations are parameterized by the complete weight vector η_s. In the beginning, the teacher filters are trained for n_w number of warm-up epochs with just the cross-entropy loss $\sum_{k=1}^{K} \mathcal{L}_{CE}(Y, g_t^{(k)}(f_t^{(k)}(T^{(k)}; \theta_t^{(k)}); \phi_t^{(k)}))$. Then, the optimization alternates between the first and the second phases, with each cycle taking q epochs in total. Within a cycle, phase 1 is carried out for $r \times q$ epochs followed by phase 2 for rest of the epochs (see Algorithm 1). Values of all the hyperparameters are given in Appendix D. TED is implemented similarly, except the fact that now, the student filters also have a classification head $g_s^{(k)}(\cdot)$ (see Appendix C). Since TED acts as a fine-tuning method, we use the BAS student as the base model to apply TED on.

In order to evaluate the validity of the Definitions 3.1 and 3.2, we compute the PID components (Bertschinger et al., 2014) of the joint mutual information $I(Y; T, S)$ of the innermost distilled layer using the estimation method given in Liang et al. (2024a). See Appendix D for additional details.

Results: Figure 4 shows the classification accuracy on CIFAR-10 dataset for each student model RID, VID, and TED when distilled with either a trained teacher or an untrained one. It also shows the classification accuracies of the baseline model BAS (without any distillation) and the trained teacher. Figures 5a and 5b present the PID values and the marginal mutual information estimates for the teacher and the student models for CIFAR-10 and CIFAR-100, respectively. Tables 2 and 3 in Appendix D present the classification accuracies for CIFAR-100 and the transfer learning setup ImageNet → CUB-200-2011, respectively.

Discussion: From Figure 4, we see that both RID and VID perform equally well when the teacher is trained. However, the performance degradation of VID under the untrained teacher is significantly larger w.r.t. that of the RID. This can be attributed to the fact that the VID student trying to naively mimic the teacher. RID leads to more resilient and effective distillation under nuisance teachers as it succinctly quantifies task-relevant knowledge rather than simply aligning student and teacher representations. TED, despite being rather unstable, performs very close to the baseline model in both cases. As evident from Figure 5, in the case of the trained teacher (whose $I(Y; T) > 0$), we observe that the knowledge available for distillation $Uni(Y : T \backslash S)$ in the teacher decreases with the increasing number of epochs. Consequently, the amount of knowledge transferred $Red(Y : T)S$ increases. When the teacher is not trained (i.e., $I(Y; T) = 0$), $Uni(Y : T \backslash S), Red(Y : T, S) \approx 0$ as expected. Both BAS and RID models show an increase in $I(Y; S)$ even when distilled from
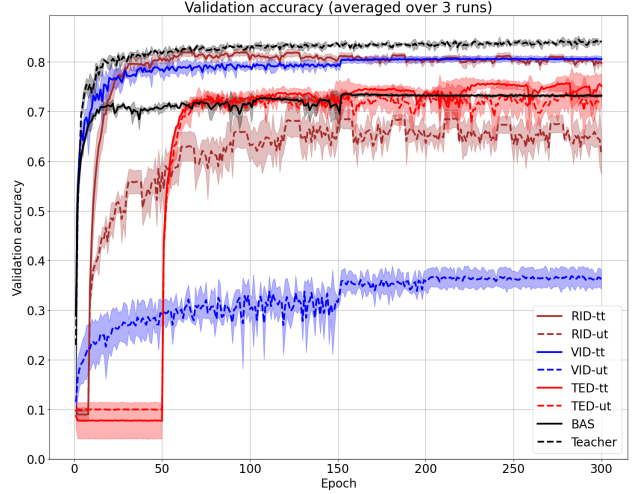

Figure 4: Classification accuracy for CIFAR-10 dataset of RID, VID, TED and BAS when distilled using a trained (abbreviated "tt"–solid lines) and an untrained (abbreviated "ut"–dashed lines) teacher: The solid and dashed lines indicate the mean over three runs. Shaded areas represent the corresponding confidence regions: mean ± standard deviation. Colors correspond to the distillation method used.

the untrained teacher. However, in this case, VID shows a very low $I(Y; S)$ as expected, caused by the distillation loss forcing to mimic the teacher.

A Python implementation of the experiments is available at `https://github.com/pasandissanayake/kd-rid`.

6 CONCLUSION

With the growing interest in knowledge distillation, our work provides critical insights into the explainability of knowledge distillation. We propose using $Uni(Y : T \backslash S)$ to quantify the knowledge available in a teacher model for distillation w.r.t. a student and a downstream task. This, in turn, leads to the definition of the amount of knowledge that has been distilled to a student as $Red(Y : T, S)$. We show that knowledge distillation frameworks which use mutual information between the teacher and the student representations to achieve knowledge distillation have a fundamental problem: These frameworks force the student to mimic the teacher regardless of the usefulness of the teacher's information to perform the task at hand. In contrast, through many examples we demonstrate that the proposed metrics can correctly characterize the amounts of knowledge available for distillation and the already transferred knowledge. Moreover, we show the advantage of the proposed metric by implementing a new distillation framework – Redundant Information Dis-
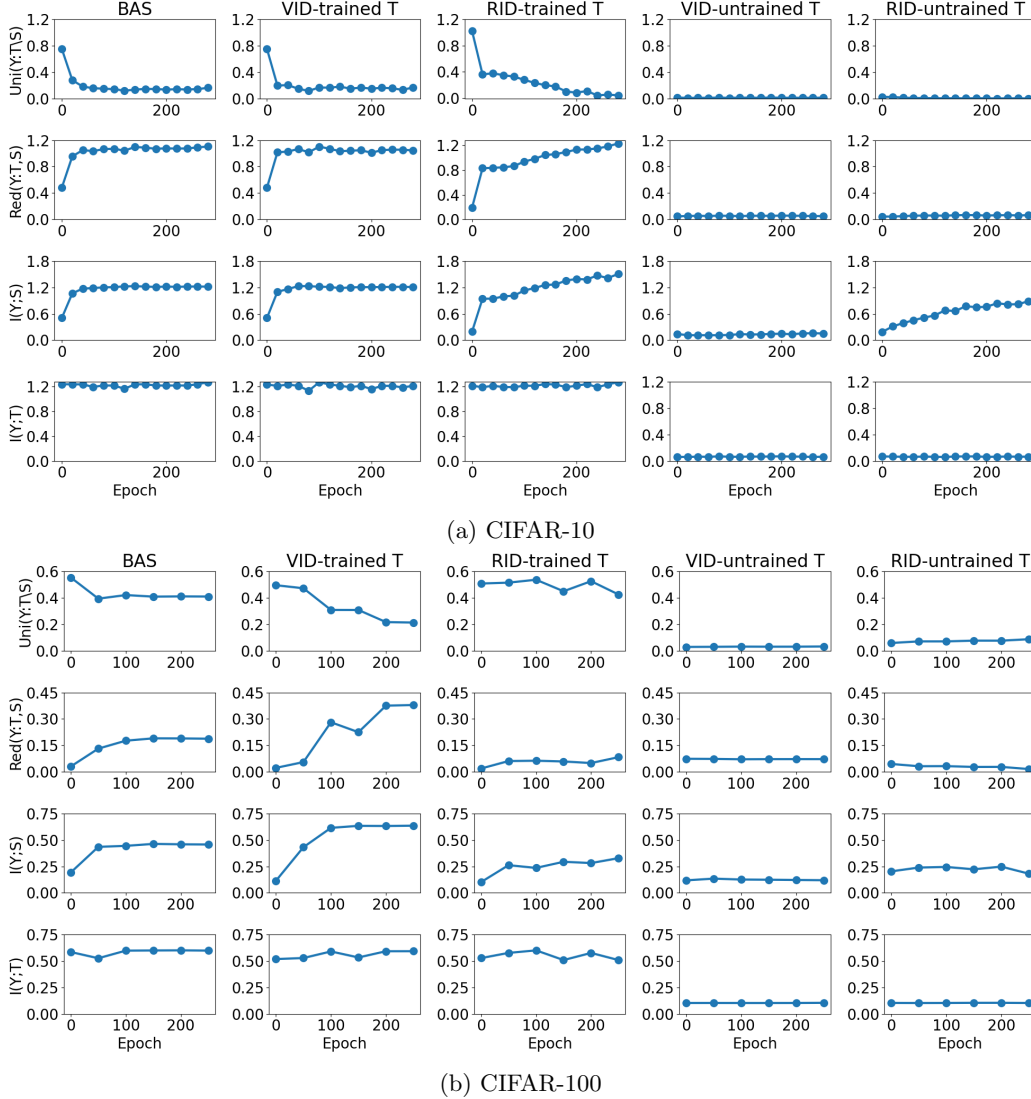
(a) CIFAR-10

(b) CIFAR-100

Figure 5: Information atoms of $I(Y; T, S)$ for BAS, VID and RID when distilled using a trained and an untrained teacher: Values are shown for the innermost distilled layer. The first two rows show that when distilled from a trained teacher, the remaining amount of knowledge available in the teacher for distillation $Uni(Y : T\backslash S)$ decreases, whereas the already transferred knowledge $Red(Y : T, S)$ increases. Observe from the third row how VID performs worse than both RID and BAS when the teacher is not trained.

tillation (RID) – and comparing its performance with the existing technique VID (Ahn et al., 2019). While VID and RID perform similarly when the teacher is well-trained for the downstream task, VID performance degrades largely when the teacher is not trained. However, RID performs close to a student model that is trained without knowledge distillation.

Limitations and future work: While the RID framework uses an alternative definition for redundant information, computation of exact $Red(Y : T, S)$ during training can be computationally prohibitive due to the optimization over Δ_P. Moreover, characterizing the extent to which the assumption in Section 4 holds is not explored in this work. Extending the mathematical formulation in Section 4 to analyze other knowledge distillation frameworks is an interesting path for future research. Other potential research directions include: (i) distilling from an ensemble of teachers (Malinin et al., 2020) in a way that the adverse effects of corrupted teachers are mitigated; (ii) dataset distillation (Sucholutsky and Schonlau, 2021); or (iii) distillation for model reconstruction from counterfactual explanations (Dissanayake and Dutta, 2024). Incorporating fundamentally different definitions for PID components, such as Lyu et al. (2024) which provides explicit formulae, as regularizers can also be interesting.

ACKNOWLEDGMENTS

REFERENCES

A. Achille and S. Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171, 2019.

P. K. Banerjee, E. Olbrich, J. Jost, and J. Rauh. Unique informations and deficiencies. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 32–38. IEEE, 2018.

E. Ben-Baruch, M. Karklinsky, Y. Biton, A. Ben-Cohen, H. Lawen, and N. Zamir. It's all in the head: Representation knowledge distillation through classifier sharing. *arXiv preprint arXiv:2201.06945*, 2022.

N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. Quantifying unique information. *Entropy*, 16 (4):2161–2183, 2014.

L. Chen, D. Wang, Z. Gan, J. Liu, R. Henao, and L. Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16296–16305, 2021.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

S. Dewan, R. Zawar, P. Saxena, Y. Chang, A. Luo, and Y. Bisk. Diffusion pid: Interpreting diffusion via partial information decomposition. *Advances in Neural Information Processing Systems*, 37:2045–2079, 2024.

P. Dissanayake and S. Dutta. Model reconstruction using counterfactual explanations: A perspective from polytope theory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

S. Dutta and F. Hamman. A review of partial information decomposition in algorithmic fairness and explainability. *Entropy*, 25(5):795, 2023.

S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover. An information-theoretic quantification of discrimination with exempt features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3825–3833, 2020.

S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover. Fairness under feature exemptions: Counterfactual and observational measures. *IEEE Transactions on Information Theory*, 67(10):6675–6710, 2021.

D. A. Ehrlich, A. C. Schneider, M. Wibral, V. Priesemann, and A. Makkeh. Partial information decomposition reveals the structure of neural representations. *arXiv preprint arXiv:2209.10438*, 2022.

C. Goswami and A. Merkley. Analytically deriving partial information decomposition for affine systems of stable and convolution-closed distributions. *Advances in Neural Information Processing Systems*, 37:86749–86835, 2024.

J. Gou, B. Yu, S. J. Maybank, and D. Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

V. Griffith and T. Ho. Quantifying redundant information in predicting a target random variable. *Entropy*, 17(7):4644–4653, 2015.

V. Griffith, E. K. P. Chong, R. G. James, C. J. Ellison, and J. P. Crutchfield. Intersection information based on common randomness. *Entropy*, 16(4):1985–2000, 2014. ISSN 1099-4300.

B. Halder, F. Hamman, P. Dissanayake, Q. Zhang, I. Sucholutsky, and S. Dutta. Quantifying spuriousness of biased datasets using partial information decomposition. *ICML Workshop on Data-centric Machine Learning Research (DMLR): Datasets for Foundation Models*, 2024.

F. Hamman and S. Dutta. Demystifying local and global fairness trade-offs in federated learning using partial information decomposition. In *International Conference on Learning Representations*, 2024a.

F. Hamman and S. Dutta. A unified view of group fairness tradeoffs using partial information decomposition. In *IEEE International Symposium on Information Theory (ISIT)*, pages 214–219, 2024b.

G. Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

W. Huang, Z. Peng, L. Dong, F. Wei, J. Jiao, and Q. Ye. Generic-to-specific distillation of masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15996–16005, 2023.

X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. Tinybert: Distilling bert

for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

M. Kleinman, A. Achille, S. Soatto, and J. C. Kao. Redundant information neural estimation. *Entropy*, 23(7):922, 2021.

A. Krizhevsky, V. Nair, and G. Hinton. Cifar-100 (canadian institute for advanced research), 2009a. URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research), 2009b. URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

S. Kundu, Q. Sun, Y. Fu, M. Pedram, and P. Beerel. Analyzing the confidentiality of undistillable teachers in knowledge distillation. *Advances in Neural Information Processing Systems*, 34:9181–9192, 2021.

C. Liang, S. Zuo, Q. Zhang, P. He, W. Chen, and T. Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pages 20852–20867. PMLR, 2023.

P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36, 2024a.

P. P. Liang, C. K. Ling, Y. Cheng, A. Obolenskiy, Y. Liu, R. Pandey, A. Wilf, L.-P. Morency, and R. Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. In *The Twelfth International Conference on Learning Representations*, 2024b.

A. Lyu, A. Clark, and N. Raviv. Explicit formula for partial information decomposition. *arXiv preprint arXiv:2402.03554*, 2024.

A. Malinin, B. Mlodozeniec, and M. Gales. Ensemble distribution distillation. In *International Conference on Learning Representations*, 2020.

R. Miles, A. L. Rodriguez, and K. Mikolajczyk. Information theoretic representation distillation. *arXiv preprint arXiv:2112.00459*, 2021.

S. Mohamadi, G. Doretto, and D. A. Adjeroh. More synergy, less redundancy: Exploiting joint mutual information for self-supervised learning. *arXiv preprint arXiv:2307.00651*, 2023.

A. Pakman, A. Nejatbakhsh, D. Gilboa, A. Makkeh, L. Mazzucato, M. Wibral, and E. Schneidman. Estimating the unique information of continuous variables. *Advances in neural information processing systems*, 34:20295–20307, 2021.

D. Y. Park, M.-H. Cha, D. Kim, B. Han, et al. Learning student-friendly teacher networks for knowledge distillation. *Advances in neural information processing systems*, 34:13292–13303, 2021.

A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *ICLR*, 2015.

I. Sucholutsky and M. Schonlau. Soft-label dataset distillation and text dataset distillation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

I. Sucholutsky, L. Muttenthaler, A. Weller, A. Peng, A. Bobu, B. Kim, B. C. Love, E. Grant, I. Groen, J. Achterberg, J. B. Tenenbaum, K. M. Collins, K. L. Hermann, K. Oktar, K. Greff, M. N. Hebart, N. Jacoby, Q. Zhang, R. Marjieh, R. Geirhos, S. Chen, S. Kornblith, S. Rane, T. Konkle, T. P. O'Connell, T. Unterthiner, A. K. Lampinen, K.-R. Müller, M. Toneva, and T. L. Griffiths. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

T. Tax, P. Mediano, and M. Shanahan. The partial information decomposition of generative neural network models. *Entropy*, 19(9):474, 2017.

Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. *ICLR*, 2020.

N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

P. Venkatesh and G. Schamberg. Partial information decomposition via deficiency for multivariate gaussians. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2892–2897. IEEE, 2022.

P. Venkatesh, C. Bennett, S. Gale, T. Ramirez, G. Heller, S. Durand, S. Olsen, and S. Mihalas. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. *Advances in Neural Information Processing Systems*, 36, 2024.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset, 2011.

C. Wang, Q. Yang, R. Huang, S. Song, and G. Huang. Efficient knowledge distillation from model checkpoints. *Advances in Neural Information Processing Systems*, 35:607–619, 2022.

P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

P. Wollstadt, S. Schmitt, and M. Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *J. Mach. Learn. Res.*, 24: 131–1, 2023.

Q. Zhang, X. Cheng, Y. Chen, and Z. Rao. Quantifying the knowledge in a dnn to explain knowledge distillation for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5099–5113, 2022.

Y. Zhu, N. Liu, Z. Xu, X. Liu, W. Meng, L. Wang, Z. Ou, and J. Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. *Advances in Neural Information Processing Systems*, 35:32011–32024, 2022.

A ADDITIONAL RELATED WORKS

The nature of the teacher model plays an important role in distillation. Kundu et al. (2021) presents a distillation scheme to distill from a teacher which has been intentionally made difficult-to-distill. They distill the penultimate layer of the teacher to several intermediate representations of the student through auxiliary student branches. Park et al. (2021) prepares the teacher during its training phase so that it can later be used for better distillation when training a student. This is done through training an augmented version of the teacher formed by adding several student branches, all of which are trained together. Ben-Baruch et al. (2022) propose two novel frameworks: Teacher-Head Knowledge Distillation (TH-KD) and Student-Head Knowledge Distillation (SH-KD). In TH-KD, the teacher's classification head is attached to the student and the distillation loss is computed as a weighted sum of the discrepancies of predictions from the two heads of the student with respect to the teacher's predictions. SH-KD involves three steps: (i) a student is trained in the conventional way; (ii) the classification head of the student is fixed to the teacher and the teacher is trained while keeping the classification head frozen; and (iii) this teacher model is used to train a new student using conventional distillation. The assumption is that the teacher trained with the student classification head will adapt the teacher to transfer the knowledge better suited to the capacity-limited student. Jiao et al. (2019) propose an interesting knowledge distillation framework for compressing the well-known BERT models into a much smaller version called TinyBERT. The process consists of two steps. During the first step, a generic TinyBERT is distilled from a pre-trained BERT model using a large corpus. In the second step, this TinyBERT model is further fine-tuned by distilling a fine-tuned BERT on a task-specific dataset. While the broad goal of Jiao et al. (2019) seems to align with ours, the approach is quite different: we intend to filter out task-specific information from a generalized teacher by defining a measure that precisely captures this whereas Jiao et al. (2019) focus on distilling from a task-specific (i.e., fine-tuned) teacher in an efficient manner. The work on distilling vision transformers by Huang et al. (2023) also proceeds in two steps. Similar to Jiao et al. (2019), the first step focuses on distilling task-agnostic information using a pre-trained teacher within an encoder-decoder setup. In the second step, the decoder is abandoned and the task-specific information is transferred using a fine-tuned teacher. As pointed out in Huang et al. (2023, Section 4.3), this framework can be seen as maximizing the mutual information (conditioned on the dataset being used) between the teacher and the student in each step (rather than quantifying

task-specific information), and hence, can be conceptually categorized together with VID. In contrast, our RID framework focuses on quantifying and extracting task-specific information from a generalized teacher.

B PROOFS

B.1 Proof of Theorem 3.1

Theorem 3.1 (Teacher with nuisance). *Let $T = (Z, G)$ where Z contains all the task-related information (i.e., $I(Y;T) = I(Y;Z)$) and G does not contain any information about the task (i.e., $I(Y;G) = 0$). Let the student be a capacity-limited model as defined by $H(S) \leq \max\{H(Z), H(G)\}$ where $H(X)$ denotes the Shannon entropy of the random variable X. Then,*

(i) $I(T;S)$ is maximized when

$$S = \begin{cases} Z & ; & H(Z) > H(G) \\ G & ; & H(Z) < H(G) \end{cases}. \tag{7}$$

(ii) $Red(Y : T, S)$ is always maximized when $S = Z$.

Proof. To prove claim 1, observe that

$$I(T;S) = H(T) - H(T|S) \tag{15}$$
$$= H(Z, G) - H(Z, G|S) \tag{16}$$
$$= H(Z) + H(G) - H(Z, G|S). \tag{17}$$

Now, $S = Z \implies H(Z, G|S) = H(G)$ and $S = G \implies H(Z, G|S) = H(Z)$. Therefore,

$$I(T;S) = \begin{cases} H(Z) & ; & S = Z \\ H(G) & ; & S = G \end{cases}. \tag{18}$$

Claim 1 follows from the above since $I(T;S) \leq H(S) \leq \max\{H(Z), H(G)\}$.

To prove claim 2, first observe that $I(Y;T) = I(Y;Z) \implies I(Y;G|Z) = 0$. Now consider the conditional mutual information $I(Y;T|S)$:

$$I(Y;T|S) = I(Y;Z, G|S) \tag{19}$$
$$= I(Y;G|S) + I(Y;Z|G, S) \tag{20}$$

Note that the right-hand side above vanishes when $S = Z$. Therefore, $S = Z \implies I(Y;T|S) = 0$. Now since

$$Red(Y : T, S) = I(Y;T) - \underbrace{\min_{Q \in \Delta_P} I_Q(Y;T|S)}_{=0 \text{ with } Q=P \text{ when } S=Z} \tag{21}$$

and $Red(Y : T, S) \leq I(Y;T)$, setting $S = Z$ achieves the maximum $Red(Y : T, S)$. □

B.2 Proof of Theorem 3.2

Theorem 3.2 (Properties). *The following properties hold for knowledge to distill and already transferred knowledge defined as in Definition 3.1 and Definition 3.2 respectively.*

(i) $Uni(Y : T\backslash S)$ and $Red(Y : T, S)$ are non-negative.

(ii) When $Uni(Y : T\backslash S) = 0$, the teacher has zero knowledge available for distillation. At this point, the student has the maximum information that any one of the representations T or S has about Y; i.e.,

$$\max\{I(Y;T), I(Y;S)\} = I(Y;S). \quad (8)$$

(iii) For a given student representation S and any two teacher representations T_1 and T_2 if there exists a deterministic mapping h such that $T_1 = h(T_2)$, then $Uni(Y : T_1\backslash S) \le Uni(Y : T_2\backslash S)$.

Proof of the second property is given below:

Proof.

$$\max\{I(Y;T), I(Y;S)\}$$
$$= \max\Big\{ Red(Y : T, S) + \underbrace{Uni(Y : T\backslash S)}_{=0}, \quad (22)$$
$$Red(Y : T, S) + Uni(Y : S\backslash T)\Big\} \quad (23)$$
$$= Red(Y : T, S) + Uni(Y : S\backslash T) \quad (24)$$
$$= I(Y;S). \quad (25)$$
$$\square$$

The first and the third properties directly follow from Bertschinger et al. (2014, Lemma 5) and Banerjee et al. (2018, Lemma 31).

B.3 Proof of Lemma B.1

Lemma B.1. *Let Y, T and S be any three random variables with supports $\mathcal{Y}, \mathcal{T}$ and $\mathcal{S}$ respectively and $g(\cdot)$ be a deterministic function with domain $\mathcal{S}$. Then*

$$I(Y;T|g(S), S) = I(Y;T|S). \quad (26)$$

Proof. By applying the mutual information chain rule to $I(Y;T, S, g(S))$ we get

$$I(Y;T, S, g(S))$$
$$= I(Y;S) + I(Y;T|S) + I(Y;g(S)|T, S) \quad (27)$$
$$= I(Y;S) + I(Y;T|S)$$
$$+ \underbrace{H(g(S)|T, S)}_{=0} - \underbrace{H(g(S)|Y, T, S)}_{=0} \quad (28)$$
$$= I(Y;S) + I(Y;T|S). \quad (29)$$

Also, from a different decomposition, we get

$$I(Y;T, S, g(S)) = I(Y;S) + \underbrace{I(Y;g(S)|S)}_{=0}$$
$$+ I(Y;T|g(S), S) \quad (30)$$
$$= I(Y;S) + I(Y;T|g(S), S). \quad (31)$$

Combining the two right-hand sides yields the final result. $\square$

B.4 Proof of Theorem 4.1

Theorem 4.1 (Transferred knowledge lower bound). *For any three random variables Y, T and S,*

$$Red_\cap(Y : T, S) \le Red(Y : T, S) \quad (10)$$

where $Red_\cap(Y : T, S)$ is as per Definition 4.1 and $Red(Y : T, S)$ is defined in Definition 3.3.

Proof. For a given set of random variables Y, T and S, let $f_t^*(T)$ and $f_s^*(S)$ achieve the maximum $I(Y;Q)$ in Definition 4.1, i.e., $Red_\cap(Y : T, S) = I(Y;f_t^*(T))$ while $I(Y;f_t^*(T)|f_s^*(S)) = 0$. We first observe that $Red(Y : f_t^*(T), f_s^*(S)) = Red_\cap(Y : T, S) = I(Y;f_t^*(T))$ as shown below:

$$Red(Y : f_t^*(T), f_s^*(S))$$
$$= I(Y;f_t^*(T)) - \min_{Q\in\Delta_P} I_Q(Y;f_t^*|f_s^*(S)) \quad (32)$$
$$= I(Y;f_t^*(T)) \quad (\because I(Y;f_t^*(T)|f_s^*(S)) = 0) \quad (33)$$
$$= Red_\cap(Y : T, S). \quad (34)$$

Next, we show that $Red(Y : f_t^*(T), f_s^*(S)) < Red(Y : T, S)$. In this regard, we use the following lemma due to Bertschinger et al. (2014).

Lemma B.2 (Lemma 25, Bertschinger et al. (2014)). *Let $X, Y, Z_1, Z_2\ldots, Z_k$ and Z_{k+1} be a set of random variables. Then,*

$$Uni(X : Y\backslash Z_1, Z_2\ldots, Z_k)$$
$$\ge Uni(X : Y\backslash Z_1, Z_2\ldots, Z_k, Z_{k+1}). \quad (35)$$

Consider the set of random variable $Y, f_t^*(T), f_s^*(S)$ and S. From the above lemma we get

$$Uni(Y : f_t^*(T)\backslash f_s^*(S))$$
$$\ge Uni(Y : f_t^*(T)\backslash f_s^*(S), S) \quad (36)$$
$$= I(Y;f_t^*(T)) - I_{Q^*}(Y;f_t^*(T)|f_s^*(S), S) \quad (37)$$

where $Q^* = \arg\min_{Q\in\Delta_P} I_{Q^*}(Y;f_t^*(T)|f_s^*(S), S)$. Now, by applying Lemma B.1 to the right-hand side

we arrive at

$$Uni(Y : f_t^*(T) \backslash f_s^*(S))$$
$$\geq I(Y; f_t^*(T)) - I_{Q^*}(Y; f_t^*(T)|S) \quad (38)$$
$$= Uni(Y : f_t^*(T) \backslash S). \quad (39)$$

Next, observe that the following line arguments holds from Definition 3.3:

$$Uni(Y : f_t^*(T) \backslash f_s^*(S)) \geq Uni(Y : f_t^*(T) \backslash S) \quad (40)$$

$$\iff I(Y; f_t^*(T)) - Uni(Y : f_t^*(T) \backslash f_s^*(S))$$
$$\leq I(Y; f_t^*(T)) - Uni(Y : f_t^*(T) \backslash S) \quad (41)$$
$$\iff Red(Y : f_t^*(T), f_s^*(S)) \leq Red(Y : f_t^*(T), S). \quad (42)$$

Noting that $Red(Y : A, B)$ is symmetric w.r.t. A and B, we may apply the previous argument to the pair $Red(Y : f_t^*(T), S)$ and $Red(Y : T, S)$ to obtain

$$Red(Y : f_t^*(T), f_s^*(S)) \leq Red(Y : f_t^*(T), S)$$
$$\leq Red(Y : T, S), \quad (43)$$

concluding the proof. □

C   VID AND TED FRAMEWORKS

C.1   Variational Information Distillation (VID)

The VID framework (Ahn et al., 2019) is based on maximizing a variational lower bound to the mutual information $I(T; S)$. It finds a student representation S which minimizes the following loss function:

$$\mathcal{L}_{VID}(\eta_s, \mu)$$
$$= \mathcal{L}_{CE}(Y, \hat{Y}_{\eta_s}) + \lambda \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\log \sigma_c \right.$$
$$\left. + \mathbb{E}_{P_X} \left[\frac{(T_{c,h,w} - \mu_{c,h,w}(S_{\eta_s}))^2}{2\sigma_c^2} \right] \right).$$
$$(44)$$

Here, C, H and W are the number of channels, height and width of the representation T respectively (i.e., $T \in \mathbb{R}^{C \times H \times W}$). μ is a deterministic function parameterized using a neural network and learned during the training process. $\sigma = [\sigma_1, \ldots, \sigma_c]^T$ is a vector of independent positive parameters, which is also learned during the training process. $\hat{Y}_{\eta_s}$ is the final prediction of the student model of the target label Y.

C.2   Task-aware Layer-wise Distillation (TED)

The TED framework Liang et al. (2023) fine-tunes a student in two stages. During the first stage, task-aware filters appended to the teacher and the student

are trained with task-related heads while the student and the teacher parameters are kept constant. In the next step, the task-related heads are removed from the filters and the student is trained along with its task-aware filter while the teacher and its task-aware filter is kept unchanged. We observe that each of these steps implicitly maximizes the redundant information under Definition 4.1. To see the relationship between the TED framework and the above definition of redundant information, let Q be parameterized using the teacher's task-aware filter as $Q = f_t(T)$. Now consider the first stage loss corresponding to the teacher's task-aware filter which is given below:

$$\mathcal{L}_t(T, \theta_t) = \mathbb{E}_{x \sim \mathcal{X}} [\ell(f_t(T; \theta_t))]. \quad (45)$$

Here, $\ell(\cdot)$ is the task specific loss, f_t is the task-aware filter parameterized by θ_t. During the first stage, this loss is minimized over θ_t. A similar loss corresponding to the student (i.e., $\mathbb{E}_{x \sim \mathcal{X}} [\ell(f_s(S; \theta_t))]$) is minimized in order to train the student's task aware filter. Note that during this process, both $I(Y; f_t(T))$ and $I(Y; f_s(S))$ are increased.

During stage 2, the distillation loss which is given below is minimized over θ_s and S while θ_t and T being held constant.

$$\mathcal{D}_{TED}(T, S) = \mathbb{E}_{x \sim \mathcal{X}} \left[||f_t(T; \theta_t) - f_s(S; \theta_s)||^2 \right]. \quad (46)$$

Consider stage 2 as an estimation problem which minimizes the mean square error, where $Q = f_t(T)$ is the estimand and $f_s(\cdot)$ is the estimator. We observe that this optimization ensures $I(Y; Q|f_s(S)) = 0$ given that the same assumption as in Section 4 holds. Following similar steps as in Section 4, we see that TED framework maximizes a lower bound for the transferred knowledge, quantified as in Definition 3.2.

The main difference of this scheme w.r.t. the RID framework is two-fold. First, in RID we optimize $f_t(\cdot)$ in addition to $f_s(\cdot)$ and S during stage 2. In contrast, TED does not modify the teacher's filter during the second stage. Second, RID distillation loss employs a weighting parameter similar to that of VID.

D   EMPIRICAL VALIDATION

D.1   Datasets

We use the CIFAR-10 dataset (Krizhevsky et al., 2009b) with 60,000 32x32 colour images belonging to 10 classes, with 6,000 images per class. The training set consists of 50,000 images (5,000 per class) and the test set is 10,000 images (1,000 per class). The PID values are evaluated over the same test set. Similarly, CIFAR-100 dataset (Krizhevsky et al., 2009a) contains 32x32 color

images belonging to 100 classes with 600 images from each class. The dataset is split into training and test sets with 500 and 100 images per class in each split, respectively. While this split was used to train the teacher, the students were trained only on a subset of training data with only 100 samples per class. For the transfer learning setup, a teacher model initialized with pre-trained weights for the ImageNet dataset (Deng et al., 2009) was used. Students were distilled using the CUB-200-2011 dataset (Wah et al., 2011). The ImageNet dataset contains 1,281,167 training images, 50,000 validation images and 100,000 test images belonging to 1,000 general object classes. The CUB-200-2011 dataset contains 11,768 images of 200 bird species.

D.2 Models and hyperparameters

CIFAR-10: Teacher models are WideResNet-(40,2) and the student models are WideResNet-(16,1). For the VID distillation, the value for λ was set to 100. Learning rate was 0.05 at the beginning and was reduced to 0.01 and 0.002 at 150th and 200th epochs respectively. Stochastic Gradient Descent with a weight decay=0.0005 and momentum=0.9 with Nesterov momentum enabled was used as the optimizer. We choose three intermediate layers; the outputs of the second, third and the fourth convolutional blocks of both the teacher and student models. The function $\mu(\cdot)$ for each layer is parameterized using a sequential model with three convolutional layers, ReLU activations and batch normalization in between the layers. A similar architecture and a training setup was used for the baseline (BAS, no distillation) and the RID models. In case of the RID models, the filters $f_s(\cdot)$ and $f_t(\cdot)$ were parameterized using 2-layer convolutional network with a batch normalization layer in the middle. The classification head $g_t(\cdot)$ is a linear layer. We set $n_w = 30, q = 30, r = 1/4$ and the total number of epochs $n + n_w = 300$. Teacher, Baseline and VID models are trained for 300 epochs. In both cases of VID and RID, the independent parameter vector σ has a dimension equal to the number of channels in the outputs of functions μ, f_s or f_t. All the training was carried out on a computer with an AMD Ryzen Threadripper PRO 5975WX processor and an Nvidia RTX A4500 graphic card. The average training time per model is around 1.5 hours.

CIFAR-100: Teacher models are WideResNet-(28,10) and the student models are WideResNet-(16,8). For the VID distillation, the value for λ was set to 100. Learning rate was 0.1 at the beginning and was multiplied by a factor of 0.2 at 60th, 120th and 160th epochs respectively. Stochastic Gradient Descent with a weight decay=0.0005 and momentum=0.9 with Nes-

terov momentum enabled was used as the optimizer. We choose three intermediate layers; the outputs of the second, third and the fourth convolutional blocks of both the teacher and student models. The function $\mu(\cdot)$ for each layer is parameterized using a sequential model with three convolutional layers, ReLU activations and batch normalization in between the layers. A similar architecture and a training setup was used for the baseline (BAS, no distillation) and the RID models. In case of the RID models, the filters $f_s(\cdot)$ and $f_t(\cdot)$ were parameterized using 2-layer convolutional network with a batch normalization layer in the middle. The classification head $g_t(\cdot)$ is a linear layer. We set $n_w = 5, q = 5, r = 1/5$ and the total number of epochs $n + n_w = 250$. Teacher, Baseline and VID models are trained for 250 epochs. In both cases of VID and RID, the independent parameter vector σ has a dimension equal to the number of channels in the outputs of functions μ, f_s or f_t. All the training was carried out on a computer with an AMD EPYC 7763 64-Core processor and an Nvidia RTX A6000 Ada graphic card. The average training time per model is around 3.75 hours.

ImageNet $\rightarrow$ CUB-200-2011: The teacher is a ResNet-34 model initialized with weights pre-trained on ImageNet (from Torchvision). The students are ResNet-18 models. For the VID distillation, the value for λ was set to 100. Learning rate was 0.1 at the beginning and was multiplied by a factor of 0.2 at 150th and 200th epochs respectively. Stochastic Gradient Descent with a weight decay=0.0005 and momentum=0.9 with Nesterov momentum enabled was used as the optimizer. We choose two intermediate layers; the outputs of the third and the fourth convolutional blocks of both the teacher and student models. The function $\mu(\cdot)$ for each layer is parameterized using a sequential model with two convolutional layers, ReLU activations and batch normalization in between the layers. A similar architecture and a training setup was used for the baseline (BAS, no distillation) and the RID models. In case of the RID models, the filters $f_s(\cdot)$ and $f_t(\cdot)$ were parameterized using 2-layer convolutional network with a batch normalization layer in the middle. The classification head $g_t(\cdot)$ is a linear layer. We set $n_w = 20, q = 20, r = 1/5$ and the total number of epochs $n + n_w = 250$. Teacher, Baseline and VID models are trained for 250 epochs. In both cases of VID and RID, the independent parameter vector σ has a dimension equal to the number of channels in the outputs of functions μ, f_s or f_t. All the training was carried out on a computer with an AMD Ryzen Threadripper PRO 5975WX processor and an Nvidia RTX A4500 graphic card. The average training time per model is around 7 hours.

D.3 PID computation

We compute the PID components of the joint information of innermost distilled layers $I(Y;T,S)$, using the framework proposed in Liang et al. (2024a) as follows:

1. Representations are individually flattened
2. Compute n_{PCA}-component PCA on each set of representations
3. Cluster representations into n_C clusters to discretize
4. Compute the joint distribution $p(Y,T,S)$
5. Compute PID components using the joint distribution

For CIFAR-10, we use $n_{\mathrm{PCA}} = n_C = 10$. For CIFAR-100, we use $n_{\mathrm{PCA}} = 5$ and $n_C = 3$. For the transfer learning setting with ImageNet and CUB-200-2011 the computation was infeasible due to the intermediate representation dimensionality being extremely large.

D.4 Results

Table 2 presents the accuracies of each framework over CIFAR-100 dataset. Table 3 presents the results of the transfer learning experiment conducted using a teacher trained on ImageNet and students distilled and evaluated over the CUB-200-2011 dataset. In each scenario, RID exhibits more resilience to the untrained teacher compared to VID.

Table 2: Accuracy of each framework on CIFAR-100 dataset when the teacher is trained vs. not trained. A subset of the training data (100 samples per class) was used for the distillation.

Framework	Trained	Untrained
RID	50%	41%
VID	67%	20%
BAS	43%	43%

Table 3: Accuracy of each framework in the transfer learning setting (ImageNet teacher $\rightarrow$ CUB-200-2011 student) when the teacher is trained vs. not trained.

Framework	Trained	Untrained
RID	65%	33%
VID	70%	11%
BAS	36%	36%