Formalizing Limits of 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 provides an effective method for deploying complex machine learning models in resource-constrained environments. It typically involves training a smaller student model to emulate either the probabilistic outputs or the internal feature representations of a larger teacher model. By doing so, the student model often achieves substantially better performance on a downstream task compared to when it is trained independently. Nevertheless, the teacher's internal representations can also encode noise or additional information that may not be relevant to the downstream task. This observation motivates our primary question: *What are the information-theoretic limits of knowledge transfer?* To this end, we leverage a body of work in information theory called Partial Information Decomposition (PID) that unravels the joint information contained in several input random variables about another target variable, e.g., the downstream task labels. Our main contribution is to quantify the distillable and distilled knowledge of a teacher's representation for a given downstream task. Moreover, we demonstrate that this metric can be practically used in distillation to address challenges caused by the complexity gap between the teacher and the student representations.

1 Introduction

Knowledge distillation 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 during the training of the student, so that the student model performs much better than when it is trained alone. In one of its simplest forms, knowledge distillation involves the student trying to match the logits of the teacher network, in addition to the correct labels of the training examples [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] (also see Gou et al. [2021], Sucholutsky et al. [2023] for a survey). 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 *introducing a metric to quantify the distillable knowledge available in a teacher model, given a student model and a target task.* As such, we bring in an emerging body of work named Partial Information Decomposition (PID) [Williams and Beer, 2010, Griffith et al., 2014, Bertschinger et al., 2014] to define the distillable knowledge as the "unique information about the task that is available only with the teacher, but not the student." As it follows, the quantification of distillable knowledge gives rise to a quantification of already distilled knowledge, leading to a metric

*Correspondence: `pasand@umd.edu`

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

that we can optimize during the distillation process. We further provide a novel knowledge distillation framework – Redundant Information Distillation (RID)– which optimizes this quantity and filters out the task-irrelevant information from the teacher. See Appendix A for a discussion on related works.

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.

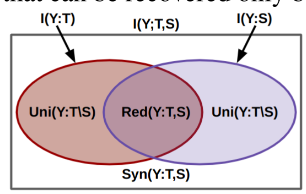

Figure 1: Partial Information Decomposition

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 a unique definition for each term does not exist, defining only one of them 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 in the context of 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: Upper-case letters denote random variables, except P and Q which represent probability distributions, C, H, W which represent the representation dimensions and K which represents the number of layers distilled. Lowercase letters are used for vectors unless specified otherwise. Lowercase Greek letters denote parameters of neural networks. 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/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.

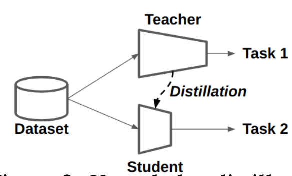

Figure 2: Knowledge distillation

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

$$\mathcal{L}(\eta_s) = \lambda_1 \mathcal{L}_{\text{ordinary}}(Y, \hat{Y}(X)) + \lambda_2 \mathcal{L}_{\text{distill}}(Y, \hat{Y}(X), S_{\eta_s}, T) \qquad (\lambda_1, \lambda_2 > 0). \quad (4)$$

Since our experiments (in Section 4) are based on a classification task, in that case Y denotes the true class label and we use the cross entropy loss $\mathcal{L}_{CE}(Y, \hat{Y}) = -\mathbb{E}_{P_X}\left[\log P_{\hat{Y}(X)}(Y)\right]$ 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.

2 Main Contribution: Quantifying Distillable and Distilled Knowledge

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 distilled to the student. Moreover, we discuss some favorable properties of the proposed metrics with examples that compare other candidate measures. Accordingly, we first define the amount of distillable information as follows:

Definition 2.1 (Distillable Knowledge)**.** *Let Y, S, and T be the target variable, student's intermediate representation, and the teacher's intermediate representation, respectively. The amount of knowledge distillable from T to S is defined as $Uni(Y : T\backslash S)$.*

With the above definition, 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 is already distilled.

Definition 2.2 (Distilled Knowledge)**.** *Let Y, S, and T be the target variable, student's intermediate representation, and the teacher's intermediate representation, respectively. The amount of distilled knowledge from T to S is defined as $Red(Y : T, S)$.*

We further propose using the unique and redundant information definitions due to Bertschinger et al. [2014] for an exact quantification.

Definition 2.3 (Unique and redundant information Bertschinger et al. [2014])**.**

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

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

where $\Delta_P = \{Q : Q(Y = y, T = t) = P(Y = y, T = t), Q(Y = y, S = s) = P(Y = y, S = s)\ \forall\ y \in \mathcal{Y}, t \in \mathcal{T}$ and $s \in \mathcal{S}\}$ and P is the joint distribution of Y, T and S.

A multitude of knowledge distillation frameworks exists 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, 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 an insight in this regard. The proposed measure $Red(Y : T, S)$ resolves these cases in an intuitive manner.

Example 1: (Uninformative teacher) An uninformative teacher representation (i.e., T with $I(Y;T) = 0$) gives $Uni(T : T\backslash S) = Red(Y : T, S) = 0$ 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. For 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 2.1 presents a more general case.

Theorem 2.1 (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., $I(Y;G) = 0$). (G can be seen as a stronger version of nuisance defined in [Achille and Soatto, 2018, Section 2.2]). Let the student be a capacity-limited model as defined by $H(S) \leq \max\{H(Z), H(G)\}$ where $H(X)$ denotes the 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 $S = Z$.

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 affects adversely on 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 4. It may appear that using $I(Y; T|S)$ as the metric for distillable knowledge resolves the cases similar to Example 1. However, Example 3 below provides a counter-example.

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 distillable information available in the teacher. Then, $I(Y; T|S) = H(Y)$ when $S = U_2$, indicating non-zero distillable information 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 $Uni(Y : T \backslash S) = Red(Y : T, S) = 0$ indicating no distillable or already distilled information available.

Next, we present Theorem 2.2 which highlights some important properties of the proposed metrics. These properties indicate that the proposed measures agree well with the intuition.

Theorem 2.2 (Properties)**.** *The following properties hold for distillable and distilled knowledge defined as in Definition 2.1 and Definition 2.2 respectively.*

1. *When $Uni(Y : T \backslash S) = 0$, the teacher has no distillable information. At this point, the student has the maximum information that any one of the representations T or S has about Y; i.e.,*
$$\max\{I(Y; T), I(Y; S)\} = I(Y; S). \tag{8}$$
2. *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 $Uni(Y : T_1 \backslash S) \leq Uni(Y : T_2 \backslash S)$.*
3. *Both $Uni(Y : T \backslash S)$ and $Red(Y : T, S)$ are non-negative.*

3 A Framework To Maximize Distilled Knowledge

In this section, we propose a distillation framework – Redundant Information Distillation (RID) – which maximizes the distilled knowledge quantified by $Red(Y : T, S)$, targeting a classification problem. Accordingly, we first show that the framework directly maps to an alternative definition of redundant information (also called the I_α measure) denoted by $Red_\cap(Y : T, S)$ [Griffith and Ho, 2015], under a certain assumption. Next, we show that $Red_\cap(Y : T, S)$ is a lower-bound for $Red(Y : T, S)$ by Bertschinger et al. [2014]. The definition of $Red_\cap(Y : T, S)$ is given below:

Definition 3.1 (I_α measure [Griffith and Ho, 2015])**.**
$$Red_\cap(Y : T, S) = \max_{P(Q|Y)} I(Y : Q) \quad \textit{subject to} \quad I(Y; Q|f_t(T)) = I(Y; Q|f_s(S)) = 0. \tag{9}$$

The proposed framework is based on selecting Q 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 3.1 results in the following optimization problem:
$$\max_{\theta_t, \theta_s, \eta_s} I(Y : f_t(T; \theta_t)) \quad \text{subject to} \quad I(Y; f_t(T; \theta_t)|f_s(S_{\eta_s}; \theta_s)) = 0. \tag{P1}$$

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 $g_t(\cdot; \phi_t)$ parametrized 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.
$$\mathcal{L}_t(\theta_t, \phi_t) = \mathcal{L}_{CE}(Y, g_t(f_t(T; \theta_t); \phi_t)) + \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbb{E}_{P_X} \left[\frac{V_{c,h,w}^2}{\sigma_c} \right] \tag{10}$$

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:

$$\mathcal{L}(\theta_s, \sigma, \eta_s) = \lambda_1 \mathcal{L}_{CE}(Y, \hat{Y}_{\eta_s}) + \lambda_2 \underbrace{\left(||\sigma||^2 + \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} \mathbb{E}_{P_X}\left[\frac{V_{c,h,w}^2}{\sigma_c}\right]\right)}_{\mathcal{L}_s(\theta_s,\sigma,\eta_s)} \tag{11}$$

where λ_1 and λ_2 are scalar hyperparameters which determine the prominence of ordinary learning and distillation. 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.

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

With the above assumption, we see that

$$I(Y;Q|f_s(S)) = I(Y; f_t(T)|f_s(S)) = I(Y; f_s(S) + \epsilon|f_s(S)) = I(Y;\epsilon|f_s(S)) = 0. \tag{12}$$

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 3.1 in two steps.

Finally, we claim through Theorem 3.1 that $Red_\cap(Y : T, S)$ is a lower bound for $Red(Y : T, S)$.

Theorem 3.1 (Distilled information lower bound)**.** *For any three random variables Y, T and S,*

$$Red_\cap(Y : T, S) \leq Red(Y : T, S) \tag{13}$$

where $Red_\cap(Y : T, S)$ is as per Definition 3.1 and $Red(Y : T, S)$ is defined in Definition 2.3.

This completes our claim that the proposed framework maximizes a lower bound for the distilled knowledge. The framework is summarized in Algorithm 1. The advantage of this framework over the VID framework [Ahn et al., 2019] (which maximizes $I(T;S)$) can be observed in the experiments in Section 4. RID losses can be extended to multiple layers by simply summing up $\mathcal{L}_t^{(k)}(\theta_t^{(k)}, \phi_t^{(k)})$ and $\mathcal{L}_s^{(k)}(\theta_s^{(k)}, \sigma^{(k)}, \eta_s)$ corresponding to the representations $T^{(k)}$ and $S_{\eta_s}^{(k)}(k = 1, \dots, K)$ in equations (10) and (11) respectively.

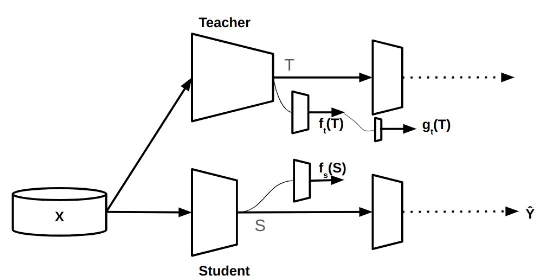

Figure 3: Redundant Information Distillation

Remark. *We observe that the Task-aware Layer-wise Distillation (TED) framework [Liang et al., 2023] shares intuitive similarities with RID, with regard to distilling task-related knowledge. However, they take a heuristic approach to the design and the focus is on large language models. In fact, our mathematical formulation can explain the success of TED as detailed in Appendix C. In addition to the domain of application, 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 distilled knowledge and formally incorporate it into an optimization.*

4 Experiments

We compare the performance of the proposed RID framework with that of the VID framework 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 the CIFAR-10 dataset [Krizhevsky et al., 2009]. Additionally, we train a student without any knowledge distillation, which we label as BAS. We distill three teacher layers to the corresponding student layers. In all cases, we compute the PID components [Bertschinger et al., 2014] of the joint information of the innermost distilled layer using the estimation method in Liang et al. [2024a]. All the teacher models are WideResNet-(40,2) and all the student models are WideResNet-(16,1). More details on the experiments are given in Appendix D.

In the case of the trained (i.e., $I(Y;T) > 0$) teacher, we observe that $Uni(Y : T\backslash S)$ decreases with the increasing number of epochs. In the case of the untrained teacher (i.e., $I(Y;T) = 0$), $Uni(Y : T\backslash S) = 0$ as expected. Both BAS and RID models show an increase in $I(Y;S)$ even under the untrained teacher. In this case, VID shows a very low $I(Y;S)$ as expected, caused by the distillation loss forcing to mimic the teacher. The results are shown in Figure 4. Figure 5 in Appendix D shows the corresponding classification accuracies.

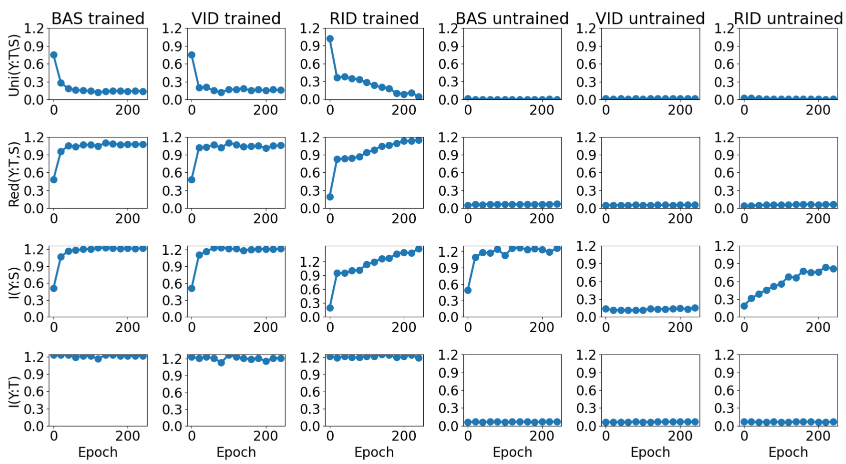

Figure 4: Information atoms of $I(Y;T,S)$ for BAS, VID and RID when distilled using a trained and an untrained teacher. Values are shown for the innermost distilled layer. Notice how VID performs worse than BAS when the teacher is not trained.

5 Conclusion

We propose using $Uni(Y : T\backslash S)$ and $Red(Y : T, S)$ to quantify distillable and distilled knowledge, corresponding to a given teacher-student pair regarding a given task. We show that knowledge distillation frameworks which use mutual information between the teacher and the student representations to quantify distillation have a fundamental problem. In contrast, through many examples we demonstrate that the proposed metric can correctly characterize the distillable and distilled knowledge. Moreover, we show the advantage of the proposed metric by implementing a new distillation framework – Redundant Information Distillation (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 independently, without knowledge distillation.

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. Extending the mathematical formulation in Section 3 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].

References

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

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *ICLR*, 2015.

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2020.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pages 20852–20867. PMLR, 2023.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Quanshi Zhang, Xu Cheng, Yilan Chen, and Zhefan 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.

Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. *Advances in Neural Information Processing Systems*, 35:607–619, 2022.

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

Virgil Griffith, Edwin K. P. Chong, Ryan G. James, Christopher J. Ellison, and James P. Crutchfield. Intersection information based on common randomness. *Entropy*, 16(4):1985–2000, 2014. ISSN 1099-4300. doi: 10.3390/e16041985. URL `https://www.mdpi.com/1099-4300/16/4/1985`.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

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

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

Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16296–16305, 2021.

Roy Miles, Adrian Lopez Rodriguez, and Krystian Mikolajczyk. Information theoretic representation distillation. *arXiv preprint arXiv:2112.00459*, 2021.

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

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

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36, 2024a.

Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BygSP6Vtvr`.

Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021. doi: 10.1109/IJCNN52387.2021.9533769.

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

Yichen Zhu, Ning Liu, Zhiyuan Xu, Xin Liu, Weibin Meng, Louis Wang, Zhicai Ou, and Jian Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. *Advances in Neural Information Processing Systems*, 35:32011–32024, 2022.

Souvik Kundu, Qirui Sun, Yao Fu, Massoud Pedram, and Peter Beerel. Analyzing the confidentiality of undistillable teachers in knowledge distillation. *Advances in Neural Information Processing Systems*, 34:9181–9192, 2021.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

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

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

Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. Fairness under feature exemptions: Counterfactual and observational measures. *IEEE Transactions on Information Theory*, 67(10):6675–6710, 2021.

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

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

Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alexander Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Russ Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. In *The Twelfth International Conference on Learning Representations*, 2024b.

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.

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

David A Ehrlich, Andreas C Schneider, Michael Wibral, Viola Priesemann, and Abdullah Makkeh. Partial information decomposition reveals the structure of neural representations. *arXiv preprint arXiv:2209.10438*, 2022.

Patricia Wollstadt, Sebastian Schmitt, and Michael 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.

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

Praveen Venkatesh, Corbett Bennett, Sam Gale, Tamina Ramirez, Greggory Heller, Severine Durand, Shawn Olsen, and Stefan Mihalas. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Michael Kleinman, Alessandro Achille, Stefano Soatto, and Jonathan C Kao. Redundant information neural estimation. *Entropy*, 23(7):922, 2021.

Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables. *Advances in neural information processing systems*, 34:20295–20307, 2021.

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

Pradeep Kr Banerjee, Eckehard Olbrich, Jürgen Jost, and Johannes Rauh. Unique informations and deficiencies. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 32–38. IEEE, 2018.

A Related works

Multi-layer knowledge distillation was introduced in FitNets [Romero et al., 2015]. There onwards 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 perpective to arrive at a solution. We refer the reader to [Gou et al., 2021, Sucholutsky et al., 2023] for a thorough survey of different techniques used in knowledge distillation. In this article, 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 undistillable classes due to the unmatched capacity of the student model. Kundu et al. [2021] presents a distillation scheme much similar to TED, with the difference in the teacher's representation used for the distillation. While TED uses intermediate representations of the teacher, Kundu et al. [2021] uses the penultimate layer.

Information theory has been instrumental in the attempts to explain 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 to learn relevant features quickly. They reveal that a partially trained checkpoint of the teacher can help the student more than the fully converged teacher. [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 there exists a gap in characterizing the fundamental limits of knowledge distillation which we seek to address using 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]. 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 shows 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.

B Proofs

B.1 Proof of Theorem 2.1

Theorem 2.1 (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., $I(Y;G) = 0$). (G can be seen as a stronger version of nuisance defined in [Achille and Soatto, 2018, Section 2.2]). Let the student be a capacity-limited model as defined by $H(S) \leq \max\{H(Z), H(G)\}$ where $H(X)$ denotes the 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 $S = Z$.

Proof. To prove claim 1, observe that

$$I(T;S) = H(T) - H(T|S) \tag{14}$$
$$= H(Z,G) - H(Z,G|S) \tag{15}$$
$$= H(Z) + H(G) - H(Z,G|S). \tag{16}$$

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

$$I(T;S) = \begin{cases} H(Z) & ; \quad S = Z \\ H(G) & ; \quad S = G \end{cases}. \tag{17}$$

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 $I(Y;T|S)$:

$$I(Y;T|S) = I(Y;Z,G|S) \tag{18}$$
$$= I(Y;G|S) + I(Y;Z|G,S) \tag{19}$$

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

$$\mathit{Red}(Y : T, S) = I(Y;T) - \underbrace{\min_{Q \in \Delta_P} I_Q(Y;T|S)}_{=0 \text{ with } Q=P \text{ when } S=Z} \tag{20}$$

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

B.2 Proof of Theorem 2.2

Theorem 2.2 (Properties)**.** *The following properties hold for distillable and distilled knowledge defined as in Definition 2.1 and Definition 2.2 respectively.*

1. *When $Uni(Y : T \backslash S) = 0$, the teacher has no distillable information. At this point, the student has the maximum information that any one of the representations T or S has about Y; i.e.,*

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

2. *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 $Uni(Y : T_1 \backslash S) \leq Uni(Y : T_2 \backslash S)$.*
3. *Both $Uni(Y : T \backslash S)$ and $\mathit{Red}(Y : T, S)$ are non-negative.*

Proof of the first property is given below:

Proof.

$$\max\{I(Y;T), I(Y;S)\} = \max\{\mathit{Red}(Y : T, S) + \underbrace{Uni(Y : T \backslash S)}_{=0}, \tag{21}$$
$$\mathit{Red}(Y : T, S) + Uni(Y : S \backslash T)\} \tag{22}$$
$$= \mathit{Red}(Y : T, S) + Uni(Y : S \backslash T) \tag{23}$$
$$= I(Y;S). \tag{24}$$

□

The second and third properties directly follow from Banerjee et al. [2018, Lemma 31] and Bertschinger et al. [2014, Lemma 5].

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*

$$I(Y;T|g(S), S) = I(Y;T|S). \tag{25}$$

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

$$I(Y;T,S,g(S)) = I(Y;S) + I(Y;T|S) + I(Y;g(S)|T,S) \tag{26}$$
$$= I(Y;S) + I(Y;T|S) + \underbrace{H(g(S)|T,S)}_{=0} - \underbrace{H(g(S)|Y,T,S)}_{=0} \tag{27}$$
$$= I(Y;S) + I(Y;T|S). \tag{28}$$

Also, from a different decomposition, we get

$$I(Y;T,S,g(S)) = I(Y;S) + \underbrace{I(Y;g(S)|S)}_{=0} + I(Y;T|g(S),S) \tag{29}$$
$$= I(Y;S) + I(Y;T|g(S),S). \tag{30}$$

Combining the two right-hand sides yields the final result. □

B.4 Proof of Theorem 3.1

Theorem 3.1 (Distilled information lower bound)**.** *For any three random variables Y, T and S,*

$$Red_{\cap}(Y:T,S) \leq Red(Y:T,S) \tag{13}$$

where $Red_{\cap}(Y:T,S)$ is as per Definition 3.1 and $Red(Y:T,S)$ is defined in Definition 2.3.

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 3.1, 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:

$$Red(Y:f_t^*(T),f_s^*(S)) = I(Y;f_t^*(T)) - \min_{Q\in\Delta_P} I_Q(Y;f_t^*|f_s^*(S)) \tag{31}$$
$$= I(Y;f_t^*(T)) \quad (\because I(Y;f_t^*(T)|f_s^*(S)) = 0) \tag{32}$$
$$= Red_{\cap}(Y:T,S). \tag{33}$$

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 $X,Y,Z_1,Z_2\dots,Z_k$ and Z_{k+1} be a set of random variables. Then,*

$$Uni(X:Y\backslash Z_1,Z_2\dots,Z_k) \geq Uni(X:Y\backslash Z_1,Z_2\dots,Z_k,Z_{k+1}). \tag{34}$$

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

$$Uni(Y:f_t^*(T)\backslash f_s^*(S)) \geq Uni(Y:f_t^*(T)\backslash f_s^*(S),S) \tag{35}$$
$$= I(Y;f_t^*(T)) - I_{Q^*}(Y;f_t^*(T)|f_s^*(S),S) \tag{36}$$

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

$$Uni(Y:f_t^*(T)\backslash f_s^*(S)) \geq I(Y;f_t^*(T)) - I_{Q^*}(Y;f_t^*(T)|S) \tag{37}$$
$$= Uni(Y:f_t^*(T)\backslash S). \tag{38}$$

Next, observe that the following line arguments hold from Definition 2.3:

$$Uni(Y:f_t^*(T)\backslash f_s^*(S)) \geq Uni(Y:f_t^*(T)\backslash S) \tag{39}$$
$$\iff I(Y;f_t^*(T)) - Uni(Y:f_t^*(T)\backslash f_s^*(S)) \leq I(Y;f_t^*(T)) - Uni(Y:f_t^*(T)\backslash S) \tag{40}$$
$$\iff Red(Y:f_t^*(T),f_s^*(S)) \leq Red(Y:f_t^*(T),S). \tag{41}$$

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

$$Red(Y:f_t^*(T),f_s^*(S)) \leq Red(Y:f_t^*(T),S) \leq Red(Y:T,S), \tag{42}$$

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:

$$\mathcal{L}_{VID}(\eta_s,\mu) = \mathcal{L}_{CE}(Y,\hat{Y}_{\eta_s}) + \lambda\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(\log\sigma_c + \mathbb{E}_{P_X}\left[\frac{(T_{c,h,w}-\mu_{c,h,w}(S_{\eta_s}))^2}{2\sigma_c^2}\right]\right). \quad (43)$$

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 3.1. 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:

$$\mathcal{L}_t\left(T,\theta_t\right) = \mathbb{E}_{x\sim\mathcal{X}}\left[\ell(f_t(T;\theta_t))\right]. \quad (44)$$

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., $\mathbb{E}_{x\sim\mathcal{X}}\left[\ell(f_s(S;\theta_t))\right]$) 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.

$$\mathcal{D}_{TED}\left(T,S\right) = \mathbb{E}_{x\sim\mathcal{X}}\left[||f_t(T;\theta_t)-f_s(S;\theta_s)||^2\right]. \quad (45)$$

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 3 holds. Following similar steps as in Section 3, we see that TED framework maximizes a lower bound for the distilled knowledge, quantified as in Definition 2.2.

The main difference of this scheme w.r.t. the RED framework is two-fold. First, in RED 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, RED distillation loss employs a weighting parameter similar to that of VID.

D Experiments

Dataset: We use the CIFAR-10 dataset [Krizhevsky et al., 2009] with 60000 32x32 colour images belonging to 10 classes, with 6000 images per class. The training set consists of 50000 images (5000 per class) and the test set is 10000 images (1000 per class). The PID values are evaluated over the same test set.

Redundant Information Distillation algorithm: We distill from multiple teacher layers $T^{(1)},\ldots,T^{(K)}$ to corresponding student layers $S^{(1)},\ldots,S^{(K)}$. 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 $g^{(k)}(\cdot)$ 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 second stages, 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).

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**

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

end

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

  if $\mathit{round}(i/q) < q \times r$ **then**

$$\underset{\{\theta_t^{(k)}, \phi_t^{(k)}\}}{\text{minimize}} \sum_{k=1}^{K} \mathcal{L}_t(\theta_t^{(k)}, \phi_t^{(k)}) \ ;$$ /* See equation (10) */

  else

$$\underset{\{\theta_s^{(k)}, \sigma^{(k)}, \eta_s\}}{\text{minimize}} \lambda_1 \mathcal{L}_{CE}(Y, \hat{Y}_{\eta_s}) + \lambda_2 \sum_{k=1}^{K} \mathcal{L}_s(\theta_s^{(k)}, \sigma^{(k)}, \eta_s) \ ;$$ /* See equation (11) */

  end

end

Models and hyperparameters: 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 150^{th} and 200^{th} epochs respectively. Stochastic Gradient Descent with a weight decay=0.0005 and momentum=0.9 with Nesterov momentum enabled was used as the optimiser. We choose three intermediate layers for distillation from the last three 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 basline (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.

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 PCA on each set of representations
3. Cluster representations to discretize
4. Compute the joint distribution $p(Y, T, S)$
5. Compute PID components using the joint distribution

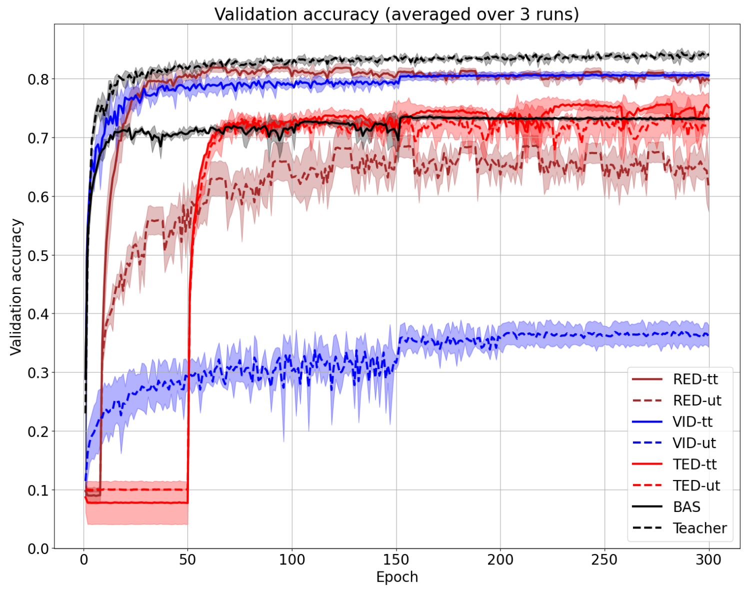

Figure 5: Classification accuracy for CIFAR10 dataset of BAS, VID, TED and RID when distilled using a trained and an untrained teacher. The suffixes "tt" stands for a trained teacher and "ut" stands for an untrained teacher. Graphs show the average over 3 runs and the shaded areas indicate mean $\pm$ standard deviation regions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Main claims made in the abstract and the introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 includes a "Limitations and future work" section where we discuss the limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are mentioned in the main text where they are used and all the proofs are provided in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the hyperparameters, details about the dataset and the algorithm in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Dataset used in the experiments is publicly available. Code will be released soon.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: These details are provided in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Accuracy of the models are evaluated over 3 runs and the one-standard deviation is depicted in pale colors. However, the PID values are computed over a single run due to the high computational cost.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Computational resources used are listed in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work does not involve human subjects or sensitive data.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss some of the potential applications in Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Paper does not poses any risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The publicly available dataset that is being used in the paper has been cited properly.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No assets are released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.