REDUNDANT INFORMATION NEURAL ESTIMATION

Michael Kleinman[1] **Alessandro Achille**[2,3] **Stefano Soatto**[1,2] **Jonathan C. Kao**[1]
[1]University of California, Los Angeles [2]Amazon Web Services [3]Caltech
michael.kleinman@ucla.edu aachille@caltech.edu
soatto@cs.ucla.edu kao@seas.ucla.edu

ABSTRACT

We introduce the Redundant Information Neural Estimator (RINE), a method that allows efficient estimation for the component of information about a target variable that is common to a set of sources, previously referred to as the "redundant information." We show that existing definitions of the redundant information can be recast in terms of an optimization over a family of deterministic or stochastic functions. In contrast to previous information decompositions, which can only be evaluated for discrete variables over small alphabets, by optimizing over functions we show empirically that we can recover the redundant information on simple benchmark tasks and that we can approximate the redundant information for high-dimensional predictors on image classification tasks, paving the way for application in different domains.

1 INTRODUCTION

Given a set of sources $X_1, \ldots, X_n$ and a target variable Y, we study how information about the target Y is distributed among the sources: different sources may contain information that no other source has ("unique information"), contain information that is common to other sources ("redundant information"), or contain complementary information that is only accessible when considered jointly with other sources ("synergistic information"). Such a decomposition of the information across the sources can inform design of multi-sensor systems (e.g., to reduce redundancy between sensors), or support research in neuroscience, where neural activity is recorded from two areas during a behavior. For example, a detailed understanding of the role and relationship between brain areas during a task requires understanding how much unique information about the behavior is provided by each area that is not available to the other area, how much information is redundant (or common) to both areas, and how much additional information is present when considering the brain areas jointly (i.e. information about the behavior that is not available when considering each area independently).

Standard information-theoretic quantities conflate these notions of information. Williams & Beer (2010) therefore proposed the Partial Information Decomposition (PID), which provides a principled framework for decomposing how the information about a target variable is distributed among a set of sources. For example, for two sources X_1 and X_2, the PID is given by:

$$I(X_1, X_2; Y) = UI(X_1; Y) + SI + UI(X_2; Y) + I_\cap \tag{1}$$

Here UI represents the "unique" information, SI the "synergistic" information, and $I_\cap$ represents the redundant information, shown in Fig 2. We provide details in Appendix C.1, describing how standard information-theoretic quantities like $I(X_1; Y)$ and $I(X_2; Y|X_1)$ are decomposed in terms of the PID constituents.

Despite effort and proposals for defining the constituents (Griffith et al., 2014; Bertschinger et al., 2014; Harder et al., 2013; Griffith & Ho, 2015; Banerjee et al., 2018; Kolchinsky, 2019), existing definitions involve difficult optimization problems and remain only feasible in low-dimensional spaces, limiting their practical applications. To enable optimization for high-dimensional problems, we reformulate the redundant information as a variational optimization problem over a restricted family of functions. We show that our formulation generalizes existing notions of redundant information. Additionally, we show that it correctly computes the redundant information on canonical low-dimensional examples and demonstrate that it can be used to compute the redundant information between different sources in a higher-dimensional image classification task.

2 Related Work

Central to the PID is the notion of redundant information $I_\cap$, and much of the work surrounding the PID has focused on specifying the desirable properties that a notion of redundancy should follow. Although there has been some disagreement as to which properties a notion of redundancy should follow (Williams & Beer, 2010; Harder et al., 2013; Kolchinsky, 2019), the following properties are widely accepted:

- *Symmetry*: $I_\cap(X_1; \dots ; X_n \to Y)$ is invariant to the permutation of $X_1, \dots , X_n$.
- *Self-redundancy*: $I_\cap(X_1 \to Y) = I(X_1; Y)$.
- *Monotonicity*: $I_\cap(X_1; \dots ; X_n \to Y) \leq I_\cap(X_1; \dots ; X_{n-1} \to Y)$.

Several notions of redundancy have been proposed that satisfy these requirements, although we emphasize that these notions were generally not defined with efficient computability in mind. Griffith et al. (2014) proposed a redundancy measure $I_\cap^\wedge$, defined through the optimization problem:

$$I_\cap^\wedge(X_1; \dots ; X_n \to Y) := \max_Q \; I(Y; Q) \quad \text{s.t.} \quad \forall i \, \exists f_i \; Q = f_i(X_i). \tag{2}$$

Here, Q is a random variable and f_i is a deterministic function. The redundant information is thus defined as the maximum information that a random variable Q, which is a deterministic function of all X_i, has about Y. This means that Q captures a component of information common to the sources X_i. A more general notion of redundant information $I_\cap^{\text{GH}}$ (Griffith & Ho, 2015; Banerjee & Griffith, 2015) is defined in terms of the following optimization problem:

$$I_\cap^{\text{GH}}(X_1; \dots ; X_n \to Y) := \max_Q I(Y; Q) \quad \text{s.t.} \; \forall i \quad I(Y; Q|X_i) = 0. \tag{3}$$

$I_\cap^{\text{GH}}$ reflects the maximum information between Y and a random variable Q such that $Y - X_i - Q$ forms a Markov chain for all X_i, relaxing that Q needs to be a deterministic function of X_i. We show in Section 3 that our definition of redundant information is a generalization of both of these notions.

2.1 Usable information in a random variable

An orthogonal line of recent work has looked at defining and computing the “usable” information $I_u(X; Y)$ that a random variable X has about Y (Xu et al., 2020; Dubois et al., 2020; Kleinman et al., 2021). This aims to capture the fact that not all information contained in a signal can be used for inference by a restricted family of functions. Given a family of decoders $\mathcal{V} \subseteq \mathcal{U} = \{f : \mathcal{X} \to \mathcal{Y}\}$, the usable information that X has about Y is defined as:

$$I_u(X; Y) = H(Y) - H_\mathcal{V}(Y|X) \tag{4}$$

where $H_\mathcal{V}(Y|X)$ is defined as:

$$H_\mathcal{V}(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}_{x,y \sim X,Y} \left[- \log f(y|x) \right]. \tag{5}$$

For example, $\mathcal{U}$ could be the the family of deterministic function and $\mathcal{V}$ could represent the family of linear deterministic functions. Thus, the “usable” information differs from Shannon’s mutual information in that it involves learning a decoder function f in a model family $\mathcal{V}$, which is a subset of all possible decoders $\mathcal{U}$. When the “usable” information is defined such that the model family corresponds to the universal model family, the definition recovers Shannon’s mutual information, $I(X; Y) = H(Y) - H_\mathcal{U}(Y|X)$. However, in many cases, the “usable information” is closer to our notion of information, reflecting the amount of information that a learned decoder, as opposed to the optimal decoder, can decode the information under computational constraints (Xu et al., 2020). We extend these ideas to compute the “usable redundant information” in the next section.

3 Redundant Information Neural Estimator

We first show that the existing definitions of redundancy can be recast in terms of an optimization over a family of functions, similar to how the “usable information” was defined above. For two sources, we can define a notion of redundancy, the Redundant Information Neural Estimator (RINE), through the following optimization over models $f_1, f_2 \in \mathcal{V}$.

$$L_\cap^\mathcal{V}(X_1; X_2 \to Y) := \min_{f_1, f_2 \in \mathcal{V}} \frac{1}{2} \big[H_{f_1}(Y|X_1) + H_{f_2}(Y|X_2) \big] \tag{6}$$

$$\text{s.t.} \;\; D(f_1, f_2) = 0 \tag{7}$$

$$I_\cap^\mathcal{V}(X_1; X_2 \to Y) := H(Y) - L_\cap^\mathcal{V}, \tag{8}$$

where $H_{f_i}(Y|X_i)$ denotes the cross-entropy when predicting Y using the decoder $f_i(y|x)$ and $D(f_1, f_2) = \mathbb{E}_{x_1,x_2}\big[\|f_1(y|x_1) - f_2(y|x_2)\|_1\big]$ denotes the expected difference of the predictions of the two decoders. Importantly, the model family $\mathcal{V}$ can be parametrized by neural networks, enabling optimization over the two model families with backpropagation. In contrast, direct optimization of eq. 2 and eq. 3 is only feasible for discrete sources with small alphabets (Kolchinsky, 2019). Our formulation can be naturally extended to n sources (Appendix C.6).

To solve the constrained minimization problem eq. 6, we can minimize the corresponding Lagrangian:

$$L_\cap^{\mathcal{V}}(X_1; X_2 \rightarrow Y, \beta) := \min_{f_1, f_2 \in \mathcal{V}} \frac{1}{2}\big[H_{f_1}(Y|X_1) + H_{f_2}(Y|X_2)\big] + \beta D(f_1, f_2). \tag{9}$$

When $\beta \rightarrow \infty$ the solution to the Lagrangian is such that $D(f_1, f_2) \rightarrow 0$, thus satisfying the constraints of the original problem. When optimizing this problem with deep networks, we found it useful to start the optimization with a low value of β, and then increase it slowly during training to some sufficiently high-value ($\beta = 50$ in our experiments).

Our definition of $\mathcal{V}$-redundant information (eq. 8) is a generalization of both $I_\cap^\wedge$ and $I_\cap^{\mathrm{GH}}$ (Sec. 2) as shown by the following proposition:

Proposition 1 (Appendix B)**.** *Let $\mathcal{V}$ be the family of* ***deterministic*** *functions, then $I_\cap^{\mathcal{V}} = I_\cap^\wedge$. If, instead, $\mathcal{V}$ is the family of* ***stochastic*** *functions, then $I_\cap^{\mathcal{V}} = I_\cap^{\mathrm{GH}}$.*

In our experiments described in the next section, we optimize over a model family $\mathcal{V}$ of deterministic neural networks using gradient descent. For the image classification tasks, we optimized over a model family $\mathcal{V}$ of ResNets (He et al., 2016), which have been successful for image classification tasks, and for other experiments we optimize over fully-connected networks. In general, the model class to optimize over should be selected such that it is not too complicated that it overfits to spurious features of the finite train set, but is of high enough capacity to learn the mapping from source to target.

4 EXPERIMENTS

We now apply our framework to estimate the redundant information on canonical examples that were previously used to study the PID, and then then demonstrate the ability to compute the redundant information for problems where the predictors are high dimensional.

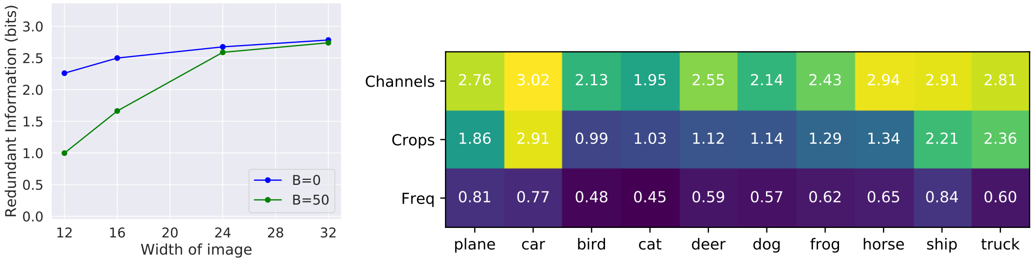
4.1 CANONICAL EXAMPLES

	True	$I_\cap^\wedge$	$I_\cap^{\mathrm{GH}}$	$I_\cap^{\mathcal{V}}$ ($\beta = 15$)
UNQ [T2]	0	0	0	**0.011**
AND [T3]	[0, 0.311]	0	0	**-0.017**
RDNXOR [T4]	1	1	1	**0.967**
IMPERFECTRDN [T5]	0.99	0	0.99	**0.989**

Table 1: Comparison of redundancy measures on canonical examples. Quantities are in bits, and $I_\cap^{\mathcal{V}}$ denotes our variational approximation (for $\beta = 15$). $I_\cap^\wedge$ denotes the redundant information in Griffith et al. (2014) and $I_\cap^{\mathrm{GH}}$ denotes the redundant information in Griffith & Ho (2015). We do this computation for different values of β in Table 6.

We first describe the results of our method on standard canonical examples that have been previously used to study the PID. They are particularly appealing because for these examples it is possible to ascertain ground truth values for the decomposition. Additionally, the predictors are low dimensional and have been previously studied, allowing us to compare our variational approximation. We describe the tasks, the values of the sources X_1, X_2, and the target Y for in Section A. Briefly, in the UNQ, each input X_1 and X_2 contributes 1 bit of unique information about the output and there is no redundant information. In the AND task, it is accepted that the redundant information should be between [0, 0.311] depending on the stringency of the notion of redundancy used (Griffith & Ho, 2015). When using deterministic decoders, as we do, we expect the redundant information to be 0 bits (not 0.311 bits). The RDNXOR tasks corresponds to a redundant XOR task, where there is 1 bit of redundant and 1 bit of synergistic information. Finally the IMPERFECTRDN tasks corresponds to the case where X_1 fully specifies the output, with X_2 having a small chance of flipping one of the bits. Hence, there should be 0.99 bits of redundant information. As we show in Table 1, RINE (optimizing with a deterministic family) recovers the desired values on all these canonical examples.

4.2 CIFAR EXPERIMENTS



	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
Channels	2.76	3.02	2.13	1.95	2.55	2.14	2.43	2.94	2.91	2.81
Crops	1.86	2.91	0.99	1.03	1.12	1.14	1.29	1.34	2.21	2.36
Freq	0.81	0.77	0.48	0.45	0.59	0.57	0.62	0.65	0.84	0.60

Figure 1: **(Left)** Redundant information of different crops of CIFAR-10 images. Redundant information as a function of the width of each partition, for different values of β. A width of 16 means that both X_1 and X_2 is a 16 x 32 image. The images begin from opposing sides, so in the case of the 16 x 32 image, there is no overlap between X_1 and X_2. As the amount of overlap increases, the redundant information increases. **(Right)** Per class redundant information for different channels, crops, and frequency decompositions, with $\beta = 50$ used in the optimization.

To the best of our knowledge, computations of redundant information have been limited to predictors that were 1 dimensional (Griffith et al., 2014; Griffith & Ho, 2015; Banerjee et al., 2018; Kolchinsky, 2019). We now show the ability to compute the redundant information when the predictors are high dimensional. We focus on the ability to predict discrete target classes, corresponding to a standard classification setting. We analyze the redundant information between different views of the same CIFAR-10 image (Figure 1), by optimizing over a model family of ResNet-18's (He et al., 2016), described in Appendix C.5. In particular, we split the image in two crops, a left crop X_1 containing all pixels in the first w columns, and a right crop X_2 containing all pixels in the last w columns (Fig 4). Intuitively, we expect that as the width of the crop w increases, the two views will overlap more, and the redundant information that they have about the task will increase. Indeed, this is what we observe in Figure 1 (left).

We study the redundant information between different sensor modalities, in particular we decompose the images into different color channels (X_1 = red channel and X_2 = blue channel), and frequencies (X_1 = low-pass filter and X_2 = high-pass filter). We show example images in Fig 4.

As expected, different color channels have highly redundant information about the task (Figure 1 (right)) except when discriminating classes (like dogs and cats) where precise color information (coming from using jointly the two channels synergistically) may prove useful. On the contrary, the high-frequency and low-frequency spectrum of the image has a lower amount of redundant information, which is also expected, since the high and low-frequencies carry complementary information. We also observe that left and right crop of the image are more redundant for pictures of cars than other classes. This is consistent with the fact that many images of cars in CIFAR-10 are symmetic frontal pictures of cars, and can easily be classified using just half of the image. Overall, there is more redundant information between channels, then crops, then frequencies. Together, we show we can compute the redundant information of high dimensional sources, confirming our intuition, and providing a scalable approach to apply in other domains.

5 CONCLUSION

Central to the PID, the notion of redundant information offers promise for characterizing the component of task-related information present across a set of sources. Despite its appeal for providing a more fine-grained depiction of the information content of multiple sources, it has proven difficult to compute in high-dimensions, limiting widespread adoption. Here, we show that existing definitions of redundancy can be recast in terms of optimization over a family of deterministic or stochastic functions. By optimizing over a subset of these functions, we show empirically that can recover the redundant information on simple benchmark tasks and that we can indeed approximate the redundant information for high-dimensional predictors, paving the way for application in different domains.

ACKNOWLEDGMENTS

MK was supported by the National Sciences and Engineering Research Council (NSERC). JCK was supported by an NSF CAREER Award (#1943467). This research was supported by a UCLA Computational Medicine Amazon Web Services Award.

REFERENCES

P. K. Banerjee, J. Rauh, and G. Montúfar. Computing the unique information. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 141–145, 2018. doi: 10.1109/ISIT.2018.8437757.

Pradeep Kr. Banerjee and Virgil Griffith. Synergy, redundancy and common information. *CoRR*, abs/1509.03706, 2015. URL http://arxiv.org/abs/1509.03706.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, Apr 2014. ISSN 1099-4300. doi: 10.3390/e16042161. URL http://dx.doi.org/10.3390/e16042161.

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

Yann Dubois, Douwe Kiela, David J Schwab, and Ramakrishna Vedantam. Learning optimal representations with the decodable information bottleneck. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18674–18690. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d8ea5f53c1b1eb087ac2e356253395d8-Paper.pdf.

Virgil Griffith and Tracey Ho. Quantifying redundant information in predicting a target random variable. *Entropy*, 17(12):4644–4653, Jul 2015. ISSN 1099-4300. doi: 10.3390/e17074644. URL http://dx.doi.org/10.3390/e17074644.

Virgil Griffith, Edwin Chong, Ryan James, Christopher Ellison, and James Crutchfield. Intersection information based on common randomness. *Entropy*, 16(4):1985–2000, Apr 2014. ISSN 1099-4300. doi: 10.3390/e16041985. URL http://dx.doi.org/10.3390/e16041985.

Malte Harder, Christoph Salge, and Daniel Polani. Bivariate measure of redundant information. *Phys. Rev. E*, 87:012130, Jan 2013. doi: 10.1103/PhysRevE.87.012130. URL https://link.aps.org/doi/10.1103/PhysRevE.87.012130.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Michael Kleinman, Alessandro Achille, Daksh Idnani, and Jonathan Kao. Usable information and evolution of optimal representations during training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=p8agn6bmTbr.

Artemy Kolchinsky. A novel approach to multivariate redundancy and synergy. *CoRR*, abs/1908.08642, 2019. URL http://arxiv.org/abs/1908.08642.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010. URL http://arxiv.org/abs/1004.2515.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

A Canonical tasks

The probabilities on the right hand side of the table denote the probability $p(x_1, x_2, y)$.

X_1	X_2	Y	
a	b	ab	1/4
a	B	aB	1/4
A	b	Ab	1/4
A	B	AB	1/4

Table 2: UNQ. X_1 and X_2 contribute uniquely 1 bit of Y. Hence, there is no redundant and synergistic information.

X_1	X_2	Y	
0	0	0	1/4
0	1	0	1/4
1	0	0	1/4
1	1	1	1/4

Table 3: AND. X_1 and X_2 combine nonlinearly to produce the output Y. It is generally accepted that the redundant information is between [0,0.311] bits (Griffith & Ho, 2015), where $I(X_1;Y) = I(X_2;Y) = 0.311$ bits.

X_1	X_2	Y	
r0	r0	r0	1/8
r0	r1	r1	1/8
r1	r0	r1	1/8
r1	r1	r0	1/8
R0	R0	R0	1/8
R0	R1	R1	1/8
R1	R0	R1	1/8
R1	R1	R0	1/8

Table 4: RDNXOR. A combination of redundant a synergistic information where X_1 and X_2 contributes 1 bit of redundant information, and 1 bit of synergistic information.

X_1	X_2	Y	
0	0	0	0.499
0	1	0	0.001
1	1	1	0.500

Table 5: IMPERFECTRDN. X_1 fully specifies the output, with X_2 having a small chance of flipping one of the bits. There should be 0.99 bits of redundant information.

B Proofs

Proposition 1: Let $\mathcal{V}$ be the family of **deterministic** functions, then $I_\cap^{\mathcal{V}} = I_\cap^{\wedge}$. If, instead, $\mathcal{V}$ is the family of **stochastic** functions, then $I_\cap^{\mathcal{V}} = I_\cap^{\mathrm{GH}}$.

Proof. We first note that if $\mathcal{V}$ represents the entire family of deterministic functions, than $\mathcal{V} = \mathcal{U}_d$. If instead, $\mathcal{V}$ represents the entire family of stochastic functions, than $\mathcal{V} = \mathcal{U}_s$.

Additionally, $D(f_1, .., f_n) = 0 \iff Q = f_i(X_i)\ \forall i$.

In both $I_\cap^{\wedge}$ and $I_\cap^{\mathrm{GH}}$, the objective is $\max_Q\ I(Y; Q)$. Since $I(Y; Q) = H(Y) - H(Y|Q)$, we can rewrite the objective as:

$$\max_Q\ I(Y; Q) = \min_Q\ H(Y|Q) \tag{10}$$

Now we know that $Q = f_i(X_i)\ \forall i$. Therefore we can perform the minimization over f_i:

$$\min_Q\ H(Y|Q) = \inf_{f_i \in \mathcal{V}} H_{f_i}(Y|X_i) \tag{11}$$

In our objective in Eqn 6, we technically minimize the average $\mathcal{V}$-cross entropy loss across sources X_i, but when $D(f_1, ..., f_n) = 0$, all the terms are equal since $Q = f_i(X_i)\ \forall i$, hence considering the average is equivalent to considering any particular source.

When f_i are **deterministic** functions, the constraint corresponds to that of $I_\cap^{\wedge}$ (i.e. that $Q = f_i(X_i)$, with f_i being a deterministic function). When f_i are more general **stochastic** functions, the constraint corresponds to a Markov chain (i.e $Y - X_i - Q$), as in $I_\cap^{\mathrm{GH}}$, or in other words that $Q = f_i(X_i)$, with f_i being a stochastic function.

By performing the optimization over a restricted set $\mathcal{V} \subseteq \mathcal{U}$ of either deterministic or stochastic functions, we recover the $\mathcal{V}$-redundant information analogues of $I_\cap^{\wedge}$ and $I_\cap^{\mathrm{GH}}$.

□

C Additional details

C.1 Partial Information Decomposition

Information theory provides a powerful framework for understanding the dependencies of random variables through the notion of mutual information (Cover & Thomas, 2006). However, information theory does not naturally describe how the information is distributed. For example, while we could compute the mutual information $I(X_1, X_2; Y)$, we would not understand how much information that X_1 contained about Y was also contained in X_2, how much information about Y was unique to X_1 (or X_2), as well as how much information about Y was only present when knowing both X_1 and X_2 together. These ideas were presented in Williams & Beer (2010) in the Partial Information Decomposition.

Standard information-theoretic quantities $I(X_1; Y)$, $I(X_1; Y|X_2)$, and $I(X_1, X_2; Y)$ can be formed with components of the decomposition:

$$I(X_1; Y) = UI(X_1; Y) + I_\cap \tag{12}$$
$$I(X_2; Y|X_1) = UI(X_2; Y) + SI \tag{13}$$
$$I(X_1, X_2; Y) = UI(X_1; Y) + SI + UI(X_2; Y) + I_\cap \tag{14}$$

Here UI represents the "unique" information and SI represents the "synergistic" information. Equation 14 comes form the chain rule of mutual information, and by combining equation 12 and equation 13. These quantities are shown in the PID diagram (Fig. 2). Computing any of these quantities allows us to compute all of them (Bertschinger et al., 2014). In Banerjee et al. (2018), they described an approach to compute the unique information, which was only feasible in low dimensions. Here, we primarily focus on computing the "redundant" information.

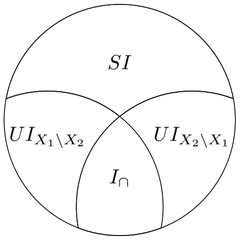


Figure 2: Decomposition of the mutual information of a sources X_1, X_2 and target Y into the synergistic information SI, the unique information UI of X_1 with respect to Y and X_2 with respect to Y, and the redundant information $I_\cap$.

C.2 Additional notion of redundancy

Recently Kolchinsky (2019) proposed to quantify redundancy via the following optimization problem:

$$I_\cap^K(X_1;\ldots;X_n \to Y) := \max_{s_{Q|Y}} I(Q;Y) \quad \text{s.t.} \quad \forall i \; s_{Q|Y} \preceq p_{X_i|Y} \tag{15}$$

The notation $p_{Q|Y} \preceq p_{X_i|Y}$ was used to indicate that there exists a channel $p_{Q|X_i}$ such that Equation 16 holds for all q and y.

$$p_{Q|Y}(q|y) = \sum_{x_i} p_{Q|X_i}(q|x_i) p_{X_i|Y}(x_i|y). \tag{16}$$

In some sense, Equation 16 indicates that Q is a “statistic” of X_i, or that the information about each X_i is contained in Q.

It would be interesting to apply a similar approximation towards this more general notion of redundancy.

C.3 Setting value of β

When optimizing the equation in practice, it is more difficult to optimize initially using very large values of β, since the network could easily learn a trivial solution. We therefore adaptively set β depending on the epoch of training, so that the final solution could be as redundant as possible, but the network would not immediately settle in a trivial solution. In this manner, we find that the network settles in a redundant solution that performs well on the task, as opposed to a redundant solution that is trivial. We smoothly increase β_i during training following the formula, so that the value of β at epoch i (γ = 0.97):

$$\beta_i = \beta(1 - \gamma^i) \tag{17}$$

When we perform an ablation study, where we fix $\beta_i = \beta$, we find that the network settles at a trivial solution (Fig 3).

C.4 Training details for canonical examples

We trained a small fully-connected network with hidden layers of size $[25 - 15 - 10]$, using batcnhorm and ReLU activations, with an initial learning rate of 0.01 decaying smoothly by 0.97 per epoch, for 15 epochs. We generated a dataset consisting of $10,000$ samples, of which 80% corresponded to training data, and the remaining 20% corresponding to test data. We trained with different values of β. $\beta = 0$ corresponds to the usable information of $I(X_1;Y)$ and $I(X_2;Y)$ (more precisely, it is $(I(X_1;Y) + I(X_2;Y))/2$, but in most cases $I(X_1;Y) = I(X_2;Y)$). As β increases, the quantity $I_\cap^\mathcal{V}$ more strongly reflects redundant information. RINE produces values close to the ground truth for these canonical examples. The tasks, with their corresponding inputs, outputs and associated probabilities are shown in the Section A. Our comparison is shown in Table 1. Note, that there is some randomness that occurs due to different initialization optimizing the neural networks, hence the values may differ slightly.

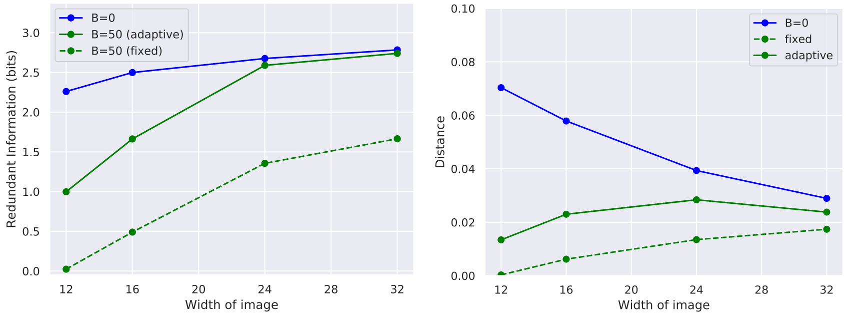


Figure 3: **(Left)** If $B = 50$ for all epochs of training, the networks is stuck in a trivial solution in learning. Setting β adaptively leads to an improved solution. **(Right)** The final distance terms are comparable.

C.5 Training details for Cifar-10

To compute the redundant information for Cifar-10, we optimized over the weights in Equation 6 using ResNet-18's He et al. (2016). We trained the network for 40 epochs, with an initial learning rate of 0.075, decreasing smoothly by 0.97 per epoch, with weight decay of 0.005. We show example images that represent the inputs x_1 and x_2 in Fig 4. We jointly train two networks that process inputs x_1 and x_2 respectively, constrained to have similar predictions through including $D(f_1, f_2)$ in the loss. To compute $D(f_1, f_2)$, we quantified the L_1 norm of the distance between the softmax scores of the predictions. We evaluate the cross-entropy loss on the test set.

C.6 Generalization to n sources

Our formulation naturally generalizes to n sources $X_1, ..., X_n$. In particular, Equation 9 can be generalized as:

$$L_\cap^{\mathcal{V}}(X_1; ...; X_N \to Y, \beta) := \min_{f_1,...,f_n \in \mathcal{V}} \frac{1}{n} \Big[\sum_{i=1}^{n} H_{f_i}(Y|X_i) \Big] + \beta D(f_1, ..., f_n). \tag{18}$$

We note that when computing the redundant information, we compute the loss without the distance term $D(f_1, ..., f_n)$. A naive extension of the distance term to n sources is computing the sum of all the pairwise distance terms. If the number of sources is large, however, it may be beneficial to consider efficient approximations of this distance term.

C.7 Details on canonical examples

	True	$I_\cap^\wedge$	$I_\cap^{\mathrm{GH}}$	$I_\cap^{\mathcal{V}}$ ($\beta = 0$)	$I_\cap^{\mathcal{V}}$ ($\beta = 5$)	$I_\cap^{\mathcal{V}}$ ($\beta = 15$)
UNQ [T2]	0	0	0	0.987	0.613	0.011
AND [T3]	[0, 0.311]	0	0	0.310	-0.001	-0.017
RDNXOR [T4]	1	1	1	0.991	0.988	0.967
IMPERFECTRDN [T5]	0.99	0	0.99	0.989	0.988	0.989

Table 6: Comparison of redundancy measures on canonical examples for additional values of β than Table 1. Quantities are in bits. $I_\cap^{\mathcal{V}}$ denotes our variational approximation, for different values of β. $I_\cap^\wedge$ denotes the redundant information in Griffith et al. (2014) and $I_\cap^{\mathrm{GH}}$ corresponds to the redundant information in Griffith & Ho (2015).

C.8 Example images in the decomposition

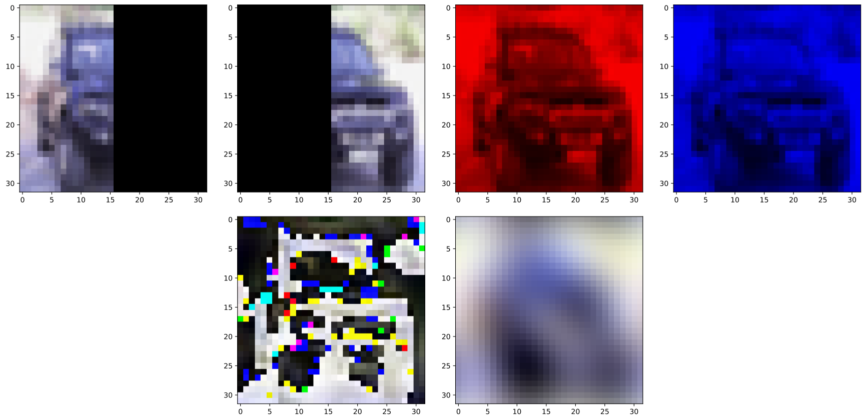


Figure 4: Example decompositions of an image (car) from CIFAR-10. This is an example of x_1 and x_2 in our CIFAR experiments. **(Top left)**: different crops, **(top right)** colors of channels, and **(bottom)**: frequencies.