EasyChair Preprint

№ 9721

Analysing Information Decomposition Between Modalities

Aiden Boyd, Rishi Agrawal, Jennifer Moss and Steph Mcgory

February 15, 2023

Analysing Information Decomposition between Modalities

A Boyd, R Agarwal, J Moss, S Mcgory

Abstract

Due to the rise of multimodal deep-learning, there are now a lot of datasets and ways to represent and combine information from different signals. However, despite such advances questions on characterizing feature interactions in multimodal datasets (i.e. which are datasets that contain information from different signals or sources) is not well studied.

We propose a method based on classical information theory to measure the degree of redundancy, uniqueness, and synergy among the input features. We work with two estimators for information that work well with high-dimensional datasets. We conduct experiments with real-world datasets, to assess the quality of the proposed procedure. Next we show how these estimates can be used to measure the interactions in multimodal datasets, the kinds of interactions that multimodal models can capture, and good ways to choose a model.

1 Introduction

A core challenge in machine learning lies in capturing the interactions between multiple features or signals. Despite progress in new models that seem to better capture interactions from increasingly complex real-world multimodal datasets, several fundamental research questions remain: How can we quantify the interactions that are necessary to solve a multimodal task? Subsequently, what type of interactions are our multimodal models actually capturing? This paper aims to formalize these research questions by proposing an information-theoretic approach to quantify the nature and degree of feature interactions. Our mathematical framework and associated empirical quantification brings together 2 previously disjoint research fields: Partial Information Decomposition (PID) in information theory (Williams and Beer. 2010, Bertschinger et al., 2014 , Griffith and Koch (2014) and multimodal machine learning (Liang et al., 2022b; Baltrušaitis et al., 2018). PID provides precise definitions enabling the categorization of interactions into redundancy, uniqueness, and synergy. Redundancy describes task-relevant information shared among features, uniqueness studies the task-relevant information present in only one of the features, and synergy investigates the emergence of new information when both features are present (see Figure 11.

Leveraging insights from neural representation learning, we propose 2 new estimators for PID that are accurate and scalable to large real-world multimodal datasets and models. The first is exact, based on convex optimization, and is able to scale to features with reasonable discrete support, while the second is an approximation based on sampling, which enables us to handle features with large discrete or even continuous supports. Through extensive experiments, we demonstrate that these estimated statistics play a helpful role in:

- Dataset quantification: We apply PID to quantify distributions in the form of large-scale multimodal datasets, showing that these estimates match common intuition for interpretable modalities (e.g., language, vision, and audio) and yield new insights in relatively understudied domains (e.g, healthcare, HCI, and robotics).
- Model quantification: Across a suite of models, we apply PID to model predictions and find consistent patterns in the interactions that different models capture.
- Model selection: Given our findings that different models tend to capture different interactions, a natural question arises: given a new multimodal task, can we quantify its PID values to infer (a priori) what type of models are most suitable? We answer this question in the positive, demonstrating successful model selection for both existing benchmarks and completely new case studies engaging with domain experts in computational pathology, mood prediction, and robotics to choose the best multimodal model for their applications.

2 Background and Related Work

Let $\mathcal{X}_i$ and $\mathcal{Y}$ be sample spaces for features and labels. Define Δ to be the set of joint distributions over $(\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y})$. We are concerned with features X_1, X_2 (with support $\mathcal{X}_i$) and labels Y (with support $\mathcal{Y}$) drawn from some distribution $p \in \Delta$. We denote the probability mass (or density) function by $p(x_1, x_2, y)$, where omitted parameters imply marginalization. Key to our work is defining estimators that given p or samples $\{(x_1, x_2, y) : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}\}$ thereof (i.e., dataset or model predictions), returns estimates for the amount of redundant, unique, and synergistic interactions.

2.1 Partial Information Decomposition

Information theory formalizes the amount of information that a variable (X) provides about another (Y), and is quantified by Shannon's mutual information (Shannon, 1948). However, the direct extension of information theory to 3 or more variables, such as through total correlation (Watanabe 1960: Garner, 1962) or interaction information (McGill. 1954: Te Sun 1980), both have significant shortcomings. In particular, the three-way mutual information $I(X_1; X_2; Y)$

can be both positive and negative, leading to considerable difficulty in its interpretation. Partial information decompo-

sition (PID) (Williams and Beer 2010) elegantly generalizes information theory to multiple variables, by positing a decomposition of the total information 2 variables X_1, X_2 provide about a task Y into 4 quantities (see Figure 2): redundancy R between X_1 and X_2, unique information U_1 in X_1 and U_2 in X_2, and synergy S. Williams and Beer (2010) show that PIDs should satisfy a set of consistency equations (see Appendix A for details). Since then, various valid PIDs have been proposed. In this paper, we adopt the definition used by Bertschinger et al. (2014); Griffith and Koch (2014), where the PID of a joint distribution p is defined as the solution to the optimization problems:

$$
\begin{aligned}
R &= \max_{q \in \Delta_p} I_q\left(X_1; X_2; Y\right), \\
U_1 &= \min_{q \in \Delta_p} I_q\left(X_1; Y \mid X_2\right), \\
U_2 &= \min_{q \in \Delta_p} I_q\left(X_2; Y \mid X_1\right), \\
S &= I_p\left(X_1, X_2; Y\right) - \min_{q \in \Delta_p} I_q\left(X_1, X_2; Y\right),
\end{aligned}
$$

where $\Delta_p = \{q \in \Delta : q\left(x_i, y\right) = p\left(x_i, y\right) \forall y \in \mathcal{Y}, x_i \in \mathcal{X}_i, i \in [2]\}$ and the notation $I_p(\cdot)$ and $I_q(\cdot)$ disambiguates mutual information under p and q respectively. Compared to others, this definition enjoys several useful properties in line with intuition (Bertschinger et al., 2014).

PID as a framework for multimodality: Our core insight is that PID provides a formal framework to understand both the nature and degree of interactions involved when two features X_1 and X_2 are used for task Y. However, computing PID via these optimization problems is a considerable challenge, since it involves optimization over Δ_p instead of simply estimating information-theoretic measures for the observed distribution p. Up to now, analytic approximations of these quantities were only possible for discrete and small support (Bertschinger et al. 2014, Griffith and Koch 2014. Wollstadt et al. 2019) or continuous but low-dimensional variables (Pakman et al., 2021; Wollstadt et al. 2021; Proca et al. 2022). Leveraging ideas in neural representation learning, Sections 3.1 and 3.2 are our first technical contributions enabling scalable estimation of PID values for highdimensional continuous distributions. Applying these new estimators to controllable synthetic datasets and real-world benchmarks (Section 4), PID provides a path towards understanding the nature of interactions in constructed datasets, the types of interactions learned by different models, and principled approaches for model selection.

3 Estimation

We now present two estimators for PID. The first is exact, based on convex optimization, and is able to scale to problems where $|\mathcal{X}_i|$ and $|\mathcal{Y}|$ are around

100. The second is an approximation based on sampling, which enables us to handle large or even continuous supports for X_i and Y.

3.1 CVX: Dataset-level Optimization

Our first estimator, CVX, follows the idea of Bertschinger et al. (2014) to directly compute PID from its definitions (11)-(4) using convex programming. Crucially, they show that the solution to the max-entropy optimization problem: $q^* = \arg\max_{q \in \Delta_p} H_q(Y \mid X_1, X_2)$ equivalently solves (1)(4). While Bertschinger et al. (2014) note that this is a convex objective with linear constraints, they report that directly performing optimization is numerically difficult, as routines such as Mathematica's FINDMINIMUM do not exploit convexity. We overcome this by rewriting conditional entropy as a KL-divergence (Globerson and Jaakkola, 2007), $H_q(Y \mid X_1, X_2) = \log|\mathcal{Y}| - KL(q||\tilde{q})$, where $\tilde{q}$ is an auxiliary product density of $q(x_1, x_2) \cdot \frac{1}{|\mathcal{Y}|}$. This relationship between q and $\tilde{q}$ is enforced using linear constraints, yielding the following equivalent problem:

$$\arg\min_{q, \tilde{q} \in \Delta_p} KL(q\|\tilde{q}), \quad \tilde{q}(x_1, x_2, y) = q(x_1, x_2)/|\mathcal{Y}|.$$

The KL-divergence objective is easily recognized as convex, allowing the use of conic solvers such as SCS (O'Donoghue et al. 2016), ECOS (Domahidi et al., 2013), and MOSEK (ApS 2022) without excessive parameter tuning. Plugging q^* into (1)-(4) yields the desired PID.

Pre-processing via feature binning: In practice, X_1 and X_2 often take continuous rather than discrete values. We workaround this by histogramming each X_i, thereby estimating the continuous joint densities X_i 's by discrete distributions with finite suppor [1] To make our discretization as data-independent as possible, we focus on a prespecified number of fixed-width bins (except for the first and last). We discuss in Appendix B.1 how the number and width of bins on affects the quality of PID estimation.

3.2 BATCH: Batch-level Amortization

We now present our next estimator,BATCH, that is suitable for large datasets where $\mathcal{X}_i$ is high-dimensional. We wish to estimate the PID values given a sampled dataset $\mathcal{D} = \left\{\left(x_1^{(j)}, x_2^{(j)}, y^{(j)}\right)\right\}$ of size n. We propose an end-to-end model parameterizing joint distributions in Δ and a training objective whose solution allows us to approximate PID based on (1)-(4).

Simplified algorithm sketch: We first illustrate our method with the assumption that oracles for densities (or probabilities if X_i, Y are discrete), $p(y \mid x_i)$ and $p(x_i)$ are known to us. Let $\widetilde{\mathcal{X}}_i = \left\{x_i^{(j)} \mid j \in [n]\right\} \subseteq \mathcal{X}_i$ be the subsampled support of $\mathcal{D}$, with $\tilde{\mathcal{Y}}$ defined similarly. Let $\widetilde{\Delta}$ denote the set of unnormalized joint distributions over $\widetilde{\mathcal{X}_1} \times \widetilde{\mathcal{X}_2} \times \widetilde{\mathcal{Y}}$. Mimicking 11 - 4 , , we define $\widetilde{\Delta}_p = \left\{\tilde{q} \in \widetilde{\Delta} : \tilde{q}\left(x_i^{(j)}, y^{(k)}\right) = p\left(x_i^{(j)}, y^{(k)}\right) \quad \forall j, k \in [n]\right\}$. Our goal, loosely

speaking, is to optimize $\tilde{q} \in \widetilde{\Delta}_p$ for objective 1 . Instead of mathematical optimization, we apply a variant of projected gradient descent on $\tilde{q}$, where projections onto $\widetilde{\Delta}_p$ are afforded by a variant of the Sinkhorn-Knopp algorithm.

Parameterization using neural networks: In practice, the above method suffers from 2 problems, (i) $\widetilde{\Delta}_p$ is too large to explicitly specify $\tilde{q}$, and (ii) we do not have oracles for p. We overcome (ii) by approximating p using $\hat{p}$, which we decompose into $\hat{p}\left(y \mid x_i\right)$ and $\hat{p}\left(x_i\right)$, both parameterized by neural networks. These $\hat{p}$ have been trained separately. To tackle (i), we approximate $\tilde{q}$ by again parameterizing it using a neural network $g_\phi : \widetilde{\mathcal{X}}_1^n \times \widetilde{\mathcal{X}}_2^n \times \mathcal{Y}^n \rightarrow \mathbb{R}^{n \times n \times|\mathcal{Y}|}$ with parameters $\phi \in \Phi$. Given full datasets $\mathbf{X_1} \in \widetilde{\mathcal{X}}_1^n, \mathbf{X_2} \in \widetilde{\mathcal{X}}_2^n, \mathbf{Y} \in \mathcal{Y}^n, g_\phi$ learns a matrix $A \in \mathbb{R}^{n \times n \times|\mathcal{Y}|}$ to represent the unnormalized joint distribution $\tilde{q}$, i.e., we want $A[i][j][y]=\tilde{q}\left(\mathbf{X_1}[i], \mathbf{X_2}[j], y\right)$. To ensure that $\tilde{q} \in \widetilde{\Delta}_p$, we use an unrolled version of Sinkhorn's algorithm (Cuturi. 2013) which projects A onto $\widetilde{\Delta}_p$ by iteratively normalizing all rows and columns to sum to 1 and rescaling to satisfy the marginals $\hat{p}$. Overall, each gradient step involves computing, $\tilde{q}=\operatorname{SINKHORN}_{\hat{p}}(A)$, and updating ϕ to minimize (1) under $\tilde{q}$. Since Sinkhorn iterations are differentiable, gradients can be backpropagated through the projection step.

Approximation with small subsampled batches: Problem (i) is not fully resolved, since the intermediate $\tilde{q}$ is too large to store. Hence, for each gradient iteration t, we bootstrap $m \ll n$ datapoints $\mathcal{D}_t=\left\{\left(x_1^{(j)}, x_2^{(j)}, y^{(j)}\right)\right\} \subseteq \mathcal{D}$. The network $g_\phi : \widetilde{\mathcal{X}}_1^m \times \widetilde{\mathcal{X}}_2^m \times \mathcal{Y}^m \rightarrow \mathbb{R}^{m \times m \times|\mathcal{Y}|}$ now takes in a batch of m datapoints and returns the unnormalized joint distribution $A \in \mathbb{R}^{m \times m \times|\mathcal{Y}|}$ for the subsampled points. Using A, we perform Sinkhorn's projection and a gradient step and ϕ as before, as if $\mathcal{D}_t$ was the full dataset. This use of a mini-batch of size m can be seen as an approximation of full-batch gradient descent. While it is challenging to obtain an unbiased estimator of the full-batch gradient since computing the full A is intractable, we found our approach to work in practice for large m. Our approach can also be informally viewed as performing amortized optimization (Amos. 2022) by using ϕ to implicitly share information about the full-batch using subsampled batches. Upon convergence, we extract PID by approximating $I_{\tilde{q}}\left(\left\{X_1, X_2\right\} ; Y\right)$ by sampling and plugging into (1)-44.

4 Experiments

4.1 Quantifying Real-world Multimodal Benchmarks

We now apply these estimators to quantify the interactions in real-world multimodal benchmarks.

Real-world multimodal data setup: We use a large collection of real-world datasets in MultiBench (Liang et al. 2021b) which test multimodal fusion of different input signals (including images, video, audio, text, time-series, sets, and tables) and require representation learning of complex real-world interactions for different tasks (predicting humor, sentiment, emotions, mortality rate,

Model	EF	ADDITIVE	AGREE	AlIGN	ELEM	TENSOR	MI	MULT	LoWER	REC	AVERAGE
R	0.35	**0.48**	**0.44**	**0.47**	0.27	**0.55**	0.20	0.40	**0.47**	**0.53**	**0.41 ± 0.11**
Acc ($\mathcal{D}_R$)	0.71	**0.74**	**0.73**	**0.74**	0.70	**0.75**	0.67	0.73	**0.74**	**0.75**	0.73 ± 0.02
U	0.29	0.31	0.19	0.44	0.20	0.52	0.18	0.45	**0.55**	**0.55**	**0.37 ± 0.14**
Acc ($\mathcal{D}_U$)	0.66	0.55	0.60	0.73	0.66	0.73	0.66	0.72	**0.73**	**0.73**	0.68 ± 0.06
S	0.13	0.09	0.08	0.29	0.14	**0.33**	0.12	**0.29**	**0.31**	**0.32**	**0.21 ± 0.10**
Acc ($\mathcal{D}_S$)	0.56	0.66	0.63	0.72	0.66	**0.74**	0.65	**0.72**	**0.73**	**0.74**	0.68 ± 0.06

Table 1: Average interactions ($R/U/S$) learned by models alongside their average performance on interaction-specialized datasets ($\mathcal{D}_R/\mathcal{D}_U/\mathcal{D}_S$). Synergy is the hardest to capture and redundancy relatively easier to capture by existing models.

Task	VQA 2.0				CLEVR			
Measure	R	U_1	U_2	S	R	U_1	U_2	S
Value	0.79	0.87	0	**4.92**	0.55	0.48	0	**5.16**

Table 2: Estimating PID on QA (Antol et al. 2015) datasets ($\times 10^{-3}$ scale). Synergy is consistently the highest.

ICD-9 codes, imagecaptions, human activities, digits, and design interfaces). We also include experiments on question-answering (Visual Question Answering 2.0(Antol et al. 2015. Goyal et al. 2017) and CLEVR (Johnson et al. 2017)) which test grounding of language into the visual domain (see Appendix C.4 for full dataset details).

Results on multimodal fusion: From Table 2 we find that different datasets do require different interactions. Some interesting observations: (1) all pairs of modalities on MUS- Table 2:

TARD sarcasm detection show high synergy values, which aligns with intuition on sarcasm in human communication, (2) uniqueness values are strongly correlated with unimodal performance (e.g., modality 1 in AV-MNIST and MIMIC), (3) datasets with high synergy do indeed benefit from interaction modeling as also seen in prior work (Liang et al. 2021b) (e.g., MUSTARD, UR-FUNNY), and (4) conversely datasets with low synergy are those where modeling higher-order interactions do not help (e.g., MIMIC).

Results on QA: We observe consistently high synergy values as shown in Table 4 This is consistent with prior work studying how these datasets were balanced (e.g., VQA 2.0 having different images for the same question such that the answer can only be obtained through synergy) (Goyal et al. 2017) and that models trained on these datasets require non-additive interactions (Hessel and Lee, 2020).

4.2 Quantifying Multimodal Model Predictions

We now shift our focus to quantifying multimodal models. Do different multimodal models learn different interactions? Better understanding the types of interactions where our current models struggle to capture can provide new insights on improving these models.

Setup: For each dataset, we train a suite of models on the train set $\mathcal{D}_{\text{train}}$ and apply it to the validation set $\mathcal{D}_{\text{val}}$, yielding a predicted dataset $\mathcal{D}_{\text{pred}} = \{(x_1, x_2, \hat{y}) \in \mathcal{D}_{\text{val}}\}$. Running PID on $\mathcal{D}_{\text{pred}}$ summarizes the interactions that the model captures. We categorize and implement a comprehensive suite of models (spanning representation fusion at different feature levels, types of interaction inductive biases, and training objectives) that have been previously motivated to capture redundant, unique, and synergistic interactions (see Appendix C.5 for full model descriptions).

Results: We show results in Table 3 and highlight the following observations:

General observations: We first observe that model PID values are consistently higher than dataset PID. The sum of model PID is also a good indicator of test performance, which agrees with their formal definition since their sum is equal to $I(\{X_1, X_2\}; Y)$, the total explained mutual information between multimodal data and Y.

On redundancy: Several methods succeed in capturing redundancy, with an overall average of $R = 0.41 \pm 0.11$ and accuracy of $73.0 \pm 2.0\%$ on redundancy-specialized datasets. Additive, agreement, and alignment-based methods are particularly strong which align with their motivation (Ding et al. 2022; Radford et al. 2021), but other methods based on tensor fusion (synergy-based), including lower-order interactions, and adding reconstruction objectives (uniquebased) also capture redundancy well.

On uniqueness: Uniqueness is harder to capture than redundancy, with an average of $U = 0.37 \pm 0.14$. Redundancybased methods like additive and agreement do poorly on uniqueness, while those designed for uniqueness (lowerorder interactions (Zadeh et al. 2017) and modality reconstruction objectives (Tsai et al. 2019b) do well, with $U = 0.55$ and 73.0% accuracy on uniqueness datasets.

On synergy: On average, synergy is the hardest to capture, with an average score of only $S = 0.21 \pm 0.10$. Some of the strong methods are tensor fusion (Fukui et al. 2016), tensors with lower-order interactions (Zadeh et al., 2017), modality reconstruction (Tsai et al. 2019b), and multimodal transformer (Xu et al. 2022), which achieve around $S = 0.30$, acc = 73.0%. Additive, agreement, and element-wise interactions do not seem to capture synergy well.

On robustness: Finally, we also show empirical connections between estimated PID values with model performance in the presence of noisy or missing modalities. Specifically, we find high correlation ($\rho = 0.62$) between the performance drop when X_i is missing and the model's U_i value. Inspecting the graph closely in Figure 4 (left), we find that the correlation is not perfect because the implication only holds in one direction: high U_i coincides with large performance drops, but low U_i can also lead to performance drops. The latter can be further explained by the presence of large R and S values: when X_i is missing, R and S interactions can no longer be discovered by the model which affects performance. For the subset of points when $U_i \leq 0.05$, the correlation between R, S and performance drop is $\rho = 0.41$, $\rho = 0.25$ respectively, and $\rho = 0.48$ for $R + S$.

4.3 PID Agreement and Model Selection

Now that we have quantified datasets and models individually, the natural next question unifies both: what does the agreement between dataset and model PID measures tell us about model performance? We hypothesize that when a model is able to capture the interactions necessary in a given dataset (i.e., high agreement), the model should also achieve high performance.

Setup: Given $\{R, U_1, U_2, S\}_{\mathcal{D}}$ on a dataset $\mathcal{D}$ and $\{R, U_1, U_2, S\}_{f(\mathcal{D})}$ on a model f trained on $\mathcal{D}$, define the agreement for each interaction $I \in \{R, U_1, U_2, S\}$ as

$$\alpha_I(f, \mathcal{D}) = \hat{I}_{\mathcal{D}} I_{f(\mathcal{D})}, \quad \hat{I}_{\mathcal{D}} = \frac{I_{\mathcal{D}}}{\sum_{I' \in \{R,U_1,U_2,S\}} I'_{\mathcal{D}}},$$

which summarizes the quantity of an interaction captured by a model $\left(I_{f(\mathcal{D})}\right)$ weighted by its normalized importance in the dataset $\left(\hat{I}_{\mathcal{D}}\right)$. The total agreement sums over all interactions $\alpha(f, \mathcal{D}) = \sum_I \alpha_I(f, \mathcal{D})$.

Results: Our key finding is that PID agreement scores $\alpha(f, \mathcal{D})$ correlate ($\rho = 0.81$) with model accuracy across all 10 synthetic datasets, as illustrated in Figure 4 (right). This shows that PID agreement can be a useful proxy for model performance. For the specialized datasets, we find that α_R is $\rho = 0.98$ correlated with model performance on $\mathcal{D}_R, \alpha_U$ is $\rho = 0.87$ correlated with performance on $\mathcal{D}_U$, and α_S is $\rho = 0.86$ correlated with performance on $\mathcal{D}_S$, and negatively correlated with other specialized datasets. For mixed datasets with roughly equal ratios of each interaction, the measures that correlate most with performance are α_R ($\rho = 0.81$) and α_S ($\rho = 0.72$); for datasets with relatively higher redundancy, the correlation of α_R increases to $\rho = 0.90$; those with higher uniqueness increases the correlation of α_{U_1} and α_{U_2} to $\rho = 0.83$ and $\rho = 0.87$; those with higher synergy increases the correlation of α_S to $\rho = 0.79$.

Using these observations, our final experiment is model selection: can we choose the most appropriate model to tackle the interactions required for a given dataset?

Setup: Given a new dataset $\mathcal{D}$, we first compute its similarity via difference in normalized PID values with respect to $\mathcal{D}'$ among our suite of 10 synthetic datasets,

$$s\left(\mathcal{D}, \mathcal{D}'\right) = \sum_{I \in \{R,U_1,U_2,S\}} \left| \hat{I}_{\mathcal{D}} - \hat{I}_{\mathcal{D}'} \right|,$$

to rank the dataset $\mathcal{D}^*$ with the most similar interactions, and return the top- 3 performing models on $\mathcal{D}^*$. In other words, we select models that best capture interactions that are of similar nature and degree as those in $\mathcal{D}$. We emphasize that even though we restrict dataset and model search to those only on synthetic datasets, our model selection procedure generalizes to real-world datasets.

Results: We test our selected models on 5 new synthetic datasets with different PID ratios and 6 real-world datasets, summarizing results in Table 5 We

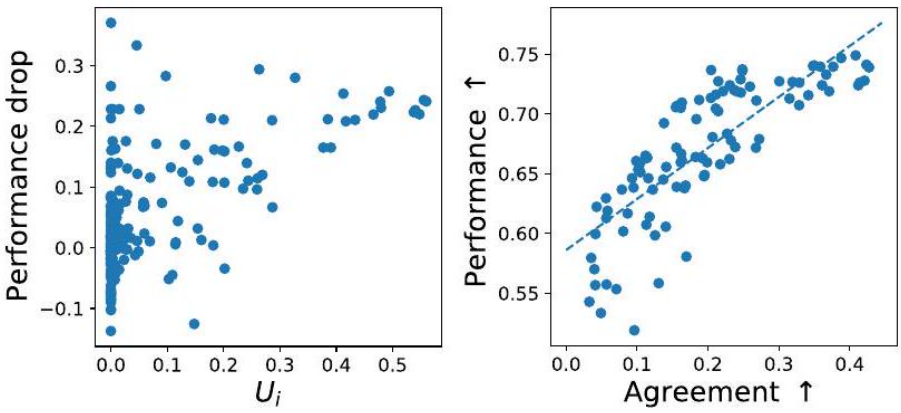

Figure 1: We find high correlation ($\rho = 0.62$) between the performance drop when X_i is missing and the model's U_i value: high U_i coincides with large performance drops, but low U_i can also lead to performance drops. The latter can be further explained by the presence of large R and S values which also cause performance drops. Right: PID agreement scores $\alpha(f, \mathcal{D})$ between datasets and models strongly correlate ($\rho = 0.81$) with model accuracy across all 10 datasets with varying PID values.

find that the top 3 chosen models achieve $95\% - 100\%$ of the best-performing model accuracy, and above 98.5% for all datasets except MUSTARD which gets 95.2%. For example, UR-FUNNY and MUStARD have the highest synergy ($S = 0.13, S = 0.3$) and indeed transformers and higher-order interactions are helpful (MULT: 0.65%, MI: 0.61%, TENSOR: 0.6%). ENRICO has the highest $R = 0.73$ and $U_2 = 0.53$, and indeed methods for redundant and unique interactions perform best (LOWER: 0.52%, Align: 0.52%, AgreE: 0.51%). MIMIC has the highest $U_1 = 0.17$, and indeed unimodal models are mostly sufficient (Liang et al. 2021b).

4.4 Real-world Applications

Finally, we apply PID to 3 real-world case studies: pathology, mental health, and robotic perception

Case Study 1: Computational pathology. Cancer prognostication is a challenging, multimodal survival task in Table 5: Model selection results on unseen synthetic and real-world datasets. Given a new dataset $\mathcal{D}$, finding the closest synthetic dataset $\mathcal{D}'$ with similar PID values and recommending the best models on $\mathcal{D}'$ consistently achieves $95\% - 100\%$ of the best-performing model on $\mathcal{D}$.

anatomic pathology that requires integration of both wholeslide imaging (WSI) and molecular features for patient stratification (Mobadersany et al. 2018: Chen et al., 2021; Lipkova et al. 2022). We quantify the interactions of these modalities on The Cancer Genome Atlas (TCGA), a large public data

Dataset	MIMIC	UR-FUNNY	MOSEI	MUSTARD	MAPS
% Performance	99.78%	98.58%	99.35%	95.15%	100%

Table 3: Model selection results on unseen synthetic and real-world datasets. Given a new dataset $\mathcal{D}'$, finding the closest synthetic dataset $\mathcal{D}$ with similar PID values and recommending the best models on $\mathcal{D}'$ consistently achieves 95%-100% of the best-performing model on $\mathcal{D}$

consortium of paired WSI, molecular, and survival information (Weinstein et al., 2013; Tomczak et al. 2015). The modalities include: (1) a sequence of pre-extracted histology image features from diagnostic WSIs and (2) feature vector of bulk gene mutation status, copy number variation, and RNA-Seq abundance values. We evaluate these interactions on two cancer datasets in TCGA, lower-grade glioma (TCGA-LGG (Network, 2015), $n = 479$) and pancreatic adenocarcinoma (TCGAPAAD (Raphael et al. 2017), $n = 209$).

Results: In TCGA-LGG, most PID measures were nearzero except $U = 0.06$ for genomic features, which indicates that genomics is the only modality containing task-relevant information. This conclusion corroborates with the high performance of unimodal-genomic and multimodal models in Chen et al. (2022), while unimodal-pathology performance was low. In TCGA-PAAD, the uniqueness values (for pathology and genomic features) and synergy value were $U_1 = 0.06$, and $U_2 = 0.08$ and $S = 0.15$ respectively, which also match the improvement of using multimodal models that capture synergistic interactions.

Case Study 2: Mental health. Suicide is the second leading cause of death among adolescents (CDC, 2020). Intensive monitoring of behaviors via adolescents' frequent use of smartphones may shed new light on the early risk of suicidal ideations (Glenn and Nock, 2014, Nahum-Shani et al. 2018), since smartphones provide a valuable and natural data source with rich behavioral markers spanning online communication, keystroke patterns, and application usage (Liang et al. 2021a). We used a dataset, MAPS, of mobile behaviors from high-risk adolescent populations with consent from participating groups (approved by NIH IRB for central institution and secondary sites). Passive sensing data is collected from each participant's smartphone across 6 months. The modalities include (1) text entered by the user represented as a bag of top 1000 words, (2) keystrokes that record the exact timing and duration of each typed character, and (3) mobile applications used per day as a bag of 137 apps. Every morning, users self-report their daily mood, which we discretized into $-1, 0, +1$. In total, MAPS dataset has 844 samples from 17 participants. Results: We first experiment with MAPS $_{T,A}$ using text and application usage features. PID measures show that MAPS$_{T,A}$ has high synergy (0.26) and low redundancy and uniqueness (0.08). The synthetic dataset $y = (z_2^*, z_c^*)$ has the most similar interactions, which enables us to select models REC, MULT, and EF, which turned out to achieve 100%, 86%, and 76% of the best-performing model respectively. We also experiment with MAPS $_{T,K}$ using text and keystroke features. MAPS $_{T,K}$ has high synergy (0.40), some redundancy (0.12), and low uniqueness (0.04). We

found $\mathcal{D}_S$ has the most similar interactions and our suggested models LOWER, REC, and TENSOR were indeed found to be the top 3 best-performing models on MAPS$_{T,K}$. This shows that model selection is quite effective on MAPS.

Case Study 3: Robotic Perception. MuJoCo PuSH (Lee et al. 2020) is a contact-rich planar pushing task in MuJoCo (Todorov et al. 2012), where a 7-DoF Panda Franka robot is pushing a circular puck with its end-effector in simulation. The pushing actions are generated by a heuristic controller that tries to move the end-effector to the center of the object. The dataset consists of 1000 trajectories with 250 steps sampled at 10Hertz. The multimodal inputs are gray-scaled images from an RGB camera, force and binary contact information from a force/torque sensor, and the 3D position of the robot end-effector. We estimate the 2D position of the unknown object on a table surface while the robot intermittently interacts with it.

Results: CVX predicts $R = 0.24, U_1 = 0.03, U_2 = 0.06, S = 0.04$ and BATCH predicts $R = 0.75, U_1 = 1.79, U_2 = 0.03, S = 0.08$. We find that BATCH predicts U_1 as the highest PID value, which aligns with our observation that image is the best unimodal predictor. Comparing both estimators, CVX underestimates U_1 and R since the high-dimensional time-series modalities cannot be easily described by clusters without losing information. In addition, both estimators predict a low U_2 value but attributes relatively high R, implying that a multimodal model with higher-order interactions would not be much better than unimodal models. Indeed, we observe no difference in performance between these two models in our experiment.

References

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Paul D Allison. Testing for interaction in multiple regression. *American Journal of Sociology*, 83(1):144–153, 1977.

Brandon Amos. Tutorial on amortized optimization for learning to optimize over continuous domains. *arXiv preprint arXiv:2202.00665*, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. *Advances in neural information processing systems*, 17, 2004.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

Jan Beirlant, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Anthony J Bell. The co-information lattice. In *Proceedings of the fifth international workshop on independent component analysis and blind signal separation: ICA*, volume 2003, 2003.

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

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

Cameron W Brennan, Roel GW Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477, 2013.

Nancy Ann Oberheim Bush, Susan M Chang, and Mitchel S Berger. Current and future strategies for treatment of glioma. *Neurosurgical Review*, 40(1): 1–14, 2017.

Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 2020.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021*, 2021.

C Mario Christoudias, Raquel Urtasun, and Trevor Darrell. Multiview learning in the presence of view disagreement. *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 88–96, 2008.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, volume 26, 2013.

Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17 (83):1–5, 2016.

Daisy Yi Ding, Shuangning Li, Balasubramanian Narasimhan, and Robert Tibshirani. Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38):e2202113119, 2022.

Alexander Domahidi, Eric Chu, and Stephen Boyd. Ecos: An socp solver for embedded systems. In *2013 European Control Conference (ECC)*, pages 3071–3076. IEEE, 2013.

Ross Flom and Lorraine E Bahrick. The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental psychology*, 43(1):238, 2007.

Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological bulletin*, 2017.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The annals of applied statistics*, 2(3):916–954, 2008.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*, pages 457–468. ACL, 2016.

Wendell R. Garner. Uncertainty and structure as psychological concepts. *Journal of Verbal Learning and Verbal Behavior*, 1962a.

Wendell R Garner. Uncertainty and structure as psychological concepts. 1962b.

Timothy J. Gawne and Barry J. Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, 13(7):2758–2771, 1993a.

Timothy J Gawne and Barry J Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, 13(7):2758–2771, 1993b.

AmirEmad Ghassami and Negar Kiyavash. Interaction information for causal inference: The case of directed triangle. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1326–1330. IEEE, 2017a.

AmirEmad Ghassami and Negar Kiyavash. Interaction information for causal inference: The case of directed triangle. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1326–1330. IEEE, 2017b.

Catherine R. Glenn and Matthew K. Nock. Improving the short-term prediction of suicidal behavior. *American Journal of Preventive Medicine*, 2014.

Amir Globerson and Tommi Jaakkola. Approximate inference using conditional entropy decompositions. In *Artificial Intelligence and Statistics*, pages 131–138. PMLR, 2007.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.