南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Dataset & Example Difficulty

A Theory of Usable Information Under Computational Constraints

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, Stefano Ermon

ICLR 2020

Shannon Mutual Information

在概率论和信息论中，两个随机变量的互信息（Mutual Information，MI）度量了两个变量之间相互依赖的程度。具体来说，对于两个随机变量，MI是一个随机变量由于已知另一个随机变量而减少的“信息量”（单位通常为比特）。

离散随机变量 X 和 Y 的互信息可以计算为：

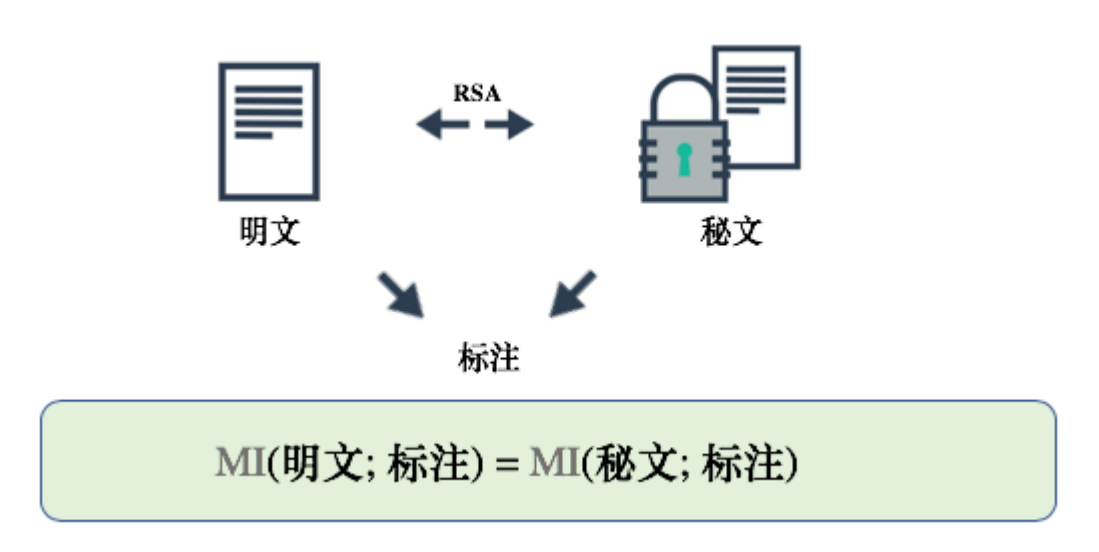
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

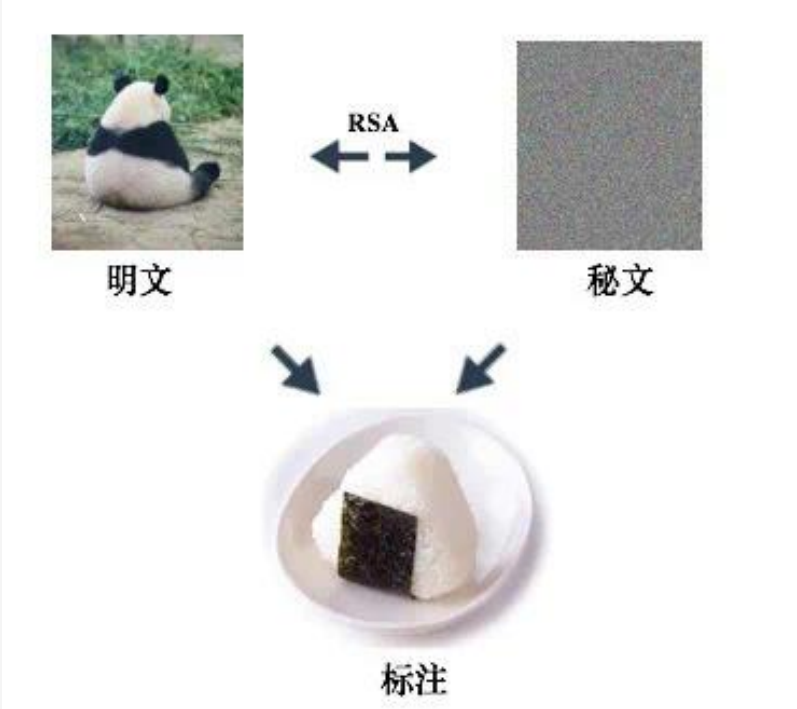
其中 $p(x,y)$ 是 X 和 Y 的联合概率质量函数，而 $p(x)$ 和 $p(y)$ 分别是 X 和 Y 的边缘概率质量函数。

互信息又可以等价地表示成：

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

Shannon Mutual Information





Shannon Mutual Information

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

对于香农互信息，我们有：

若我们把观测者限制为人类，我们有：

对于香农互信息，我们有：

如果我们把观测者限制为分类器，我们有：

Definitions and Notations

Let $\Omega = \{f : \mathcal{X} \cup \{\varnothing\} \to \mathcal{P}(\mathcal{Y})\}$. We say that $\mathcal{V} \subseteq \Omega$ is a **predictive family** if it satisfies

$$\forall f \in \mathcal{V}, \forall P \in \text{range}(f), \quad \exists f' \in \mathcal{V}, \quad \text{s.t.} \quad \forall x \in \mathcal{X}, f'[x] = P, f'[\varnothing] = P$$

Let X, Y be two random variables taking values in $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{V}$ be a predictive family. Then the **predictive conditional $\mathcal{V}$-entropy** is defined as

$$H_{\mathcal{V}}(Y \mid X) = \inf_{f \in \mathcal{V}} \mathbb{E}_{x,y \sim X,Y}[-\log f[x](y)]$$
$$H_{\mathcal{V}}(Y \mid \varnothing) = \inf_{f \in \mathcal{V}} \mathbb{E}_{y \sim Y}[-\log f[\varnothing](y)]$$

Let X, Y be two random variables taking values in $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{V}$ be a predictive family. The **predictive $\mathcal{V}$-information** from X to Y is defined as

$$I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y \mid \varnothing) - H_{\mathcal{V}}(Y \mid X)$$

Properties of $\mathcal{V}$-Information

Let Y and X be any random variables on $\mathcal{Y}$ and $\mathcal{X}$, and $\mathcal{V}$ and $\mathcal{U}$ be any predictive families, then we have

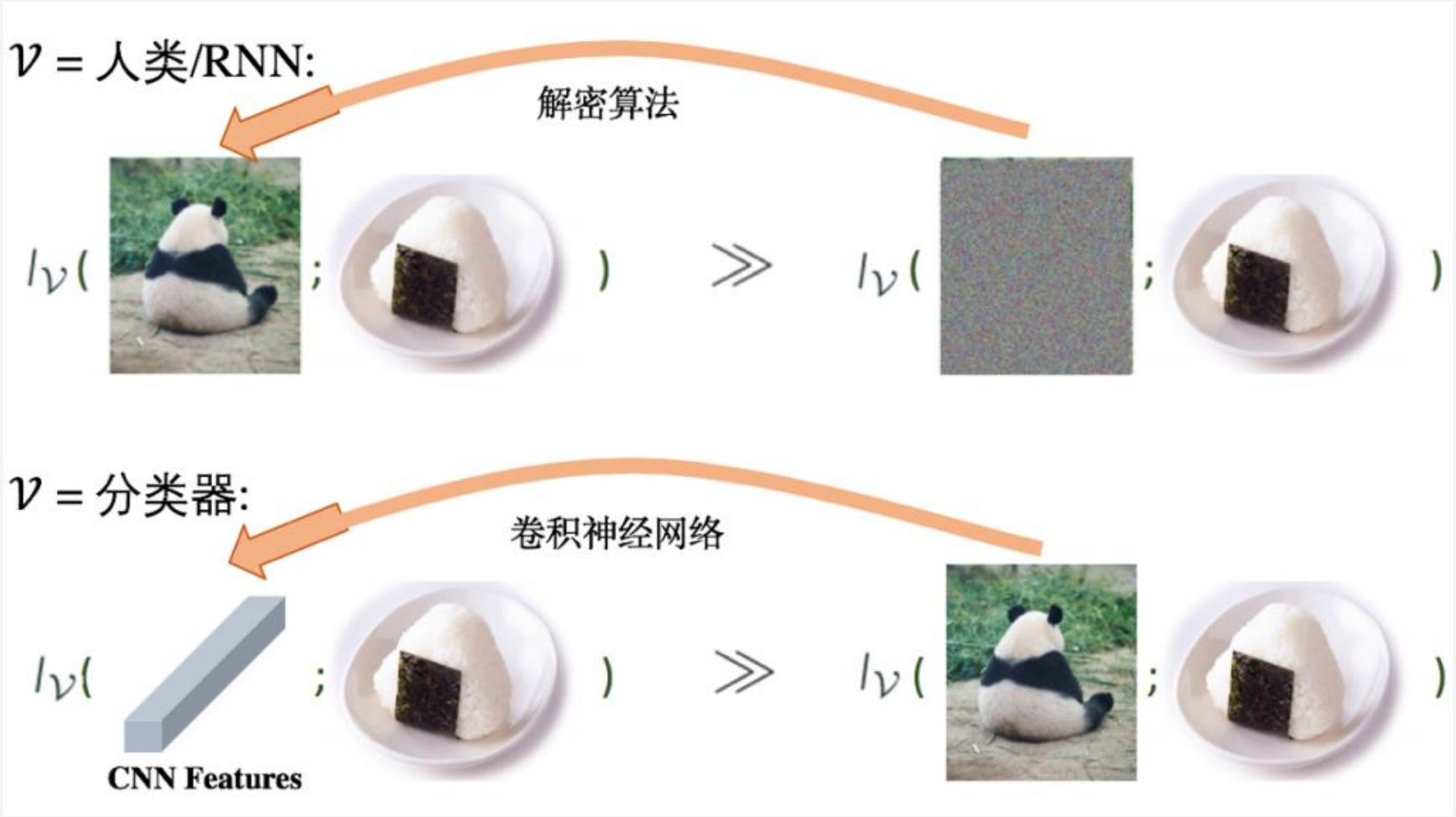
1. **Monotonicity**: If $\mathcal{V} \subseteq \mathcal{U}$, then $H_{\mathcal{V}}(Y) \geq H_{\mathcal{U}}(Y), H_{\mathcal{V}}(Y \mid X) \geq H_{\mathcal{U}}(Y \mid X)$.
2. **Non-Negativity**: $I_{\mathcal{V}}(X \to Y) \geq 0$.
3. **Independence**: If X is independent of Y, $I_{\mathcal{V}}(X \to Y) = I_{\mathcal{V}}(Y \to X) = 0$.

Shannon Mutual Information: Letting $t: \mathcal{X} \to \mathcal{X}$ be any function, $t(X)$ cannot have higher mutual information with Y than X: $I(t(X); Y) \leq I(X; Y)$.

$\mathcal{V}$-Information: Denoting t as the decryption algorithm and $\mathcal{V}$ as a class of natural language processing functions, we have that: $I_{\mathcal{V}}(t(X) \to Y) > I_{\mathcal{V}}(X \to Y) \approx 0$.

Asymmetry: If $\mathcal{V}$ contains all polynomial-time computable functions, then $I_{\mathcal{V}}\big(X \to h(X)\big) \gg I_{\mathcal{V}}(h(X) \to X)$, where $h: \mathcal{X} \to \mathcal{Y}$.

Properties of $\mathcal{V}$-Information



$\mathcal{V}$-Information Estimation

Let X, Y be two random variables taking values in $\mathcal{X}, \mathcal{Y}$ and $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \sim X, Y$ denotes the set of samples drawn from the joint distribution over $\mathcal{X}$ and $\mathcal{Y}$. $\mathcal{V}$ is a predictive family. The **empirical $\mathcal{V}$-information** (under $\mathcal{D}$) is the following $\mathcal{V}$-information under the empirical distribution defined via $\mathcal{D}$:

$$\hat{I}_{\mathcal{V}}(X \to Y; \mathcal{D}) = \inf_{f \in \mathcal{V}} \frac{1}{|\mathcal{D}|} \sum_{y_i \in \mathcal{D}} \log \frac{1}{f[\varnothing](y_i)} - \inf_{f \in \mathcal{V}} \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} \log \frac{1}{f[x_i](y_i)}$$

PAC bound over the empirical $\mathcal{V}$-information:

Assume $\forall f \in \mathcal{V}, x \in \mathcal{X}, y \in \mathcal{Y}, \log f[x](y) \in [-B, B]$. Then for any $\delta \in (0,0.5)$, with probability at least $1 - 2\delta$, we have:

$$\left| I_{\mathcal{V}}(X \to Y) - \hat{I}_{\mathcal{V}}(X \to Y; \mathcal{D}) \right| \leq 4\mathfrak{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}) + 2B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}$$

where we define the function family $\mathcal{G}_{\mathcal{V}} = \{g \mid g(x, y) = \log f[x](y), f \in \mathcal{V}\}$, and $\mathfrak{R}_N(\mathcal{G})$ denotes the Rademacher complexity of $\mathcal{G}$ with sample number N. Typically, the Rademacher complexity term satisfies $\mathfrak{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}) = \mathcal{O}\left(|\mathcal{D}|^{-\frac{1}{2}}\right)$.

Structure Learning with $\mathcal{V}$-information

$$g^* = \underset{g \in G_{\text{tree}}}{\operatorname{argmax}} \sum_{(X_i, X_j) \in \text{edge}(g)} I\left(X_i, X_j\right)$$

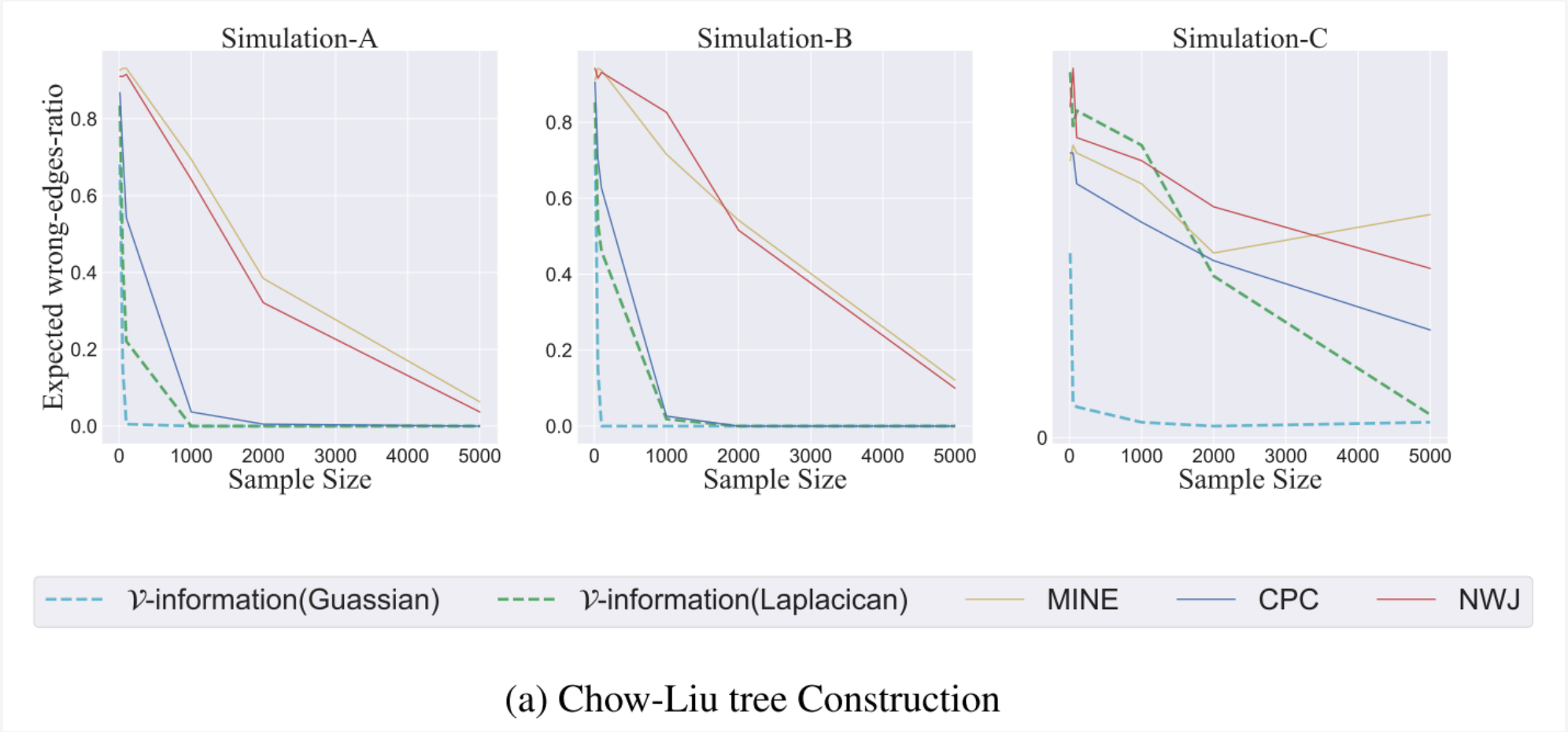
where $I\left(X_i, X_j\right)$ is the Shannon mutual information between variables X_i and X_j.

$$g^* = \underset{g \in G_{\text{d- tree}}}{\operatorname{argmax}} \sum_{i=2}^{m} I_{\mathcal{V}_{t(g)(i),i}}\left(X_{t(g)(i)} \to X_i\right)$$

where $G_{\text{d- tree}}$ is the set of directed trees, and $t(g)\colon \mathbb{N} \to \mathbb{N}$ is the function mapping each non-root node of directed tree g to its parent, and $\mathcal{V}_{i,j}$ is the predictive family for random variables X_i and X_j.

Experimental Results

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY



(a) Chow-Liu tree Construction

Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information

Kawin Ethayarajh, Yejin Choi, Swabha Swayamdipta

ICML 2022

$\mathcal{V}$-Usable Information

Let X, Y denote random variables with sample spaces $\mathcal{X}, \mathcal{Y}$ respectively. Let $\varnothing$ denote a null input that provides no information about Y. Given predictive family $\mathcal{V} \subseteq \Omega = \{f: \mathcal{X} \cup \varnothing \to P(\mathcal{Y})\}$

- predictive $\mathcal{V}$-entropy $$H_{\mathcal{V}}(Y) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)]$$
- conditional $\mathcal{V}$-entropy $$H_{\mathcal{V}}(Y \mid X) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)]$$
- $\mathcal{V}$-information $$I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y \mid X)$$

Properties of $\mathcal{V}$-Usable Information

- Non-Negativity: $I_{\mathcal{V}}(X \to Y) \geq 0$
- Independence: If X is independent of Y, $I_{\mathcal{V}}(X \to Y) = 0$.
- Montonicity: If $\mathcal{U} \subseteq \mathcal{V}$, then $H_{\mathcal{U}}(Y) \geq H_{\mathcal{V}}(Y)$ and $H_{\mathcal{U}}(Y \mid X) \geq H_{\mathcal{V}}(Y \mid X)$

$\mathcal{V}$-Usable Information in Practice

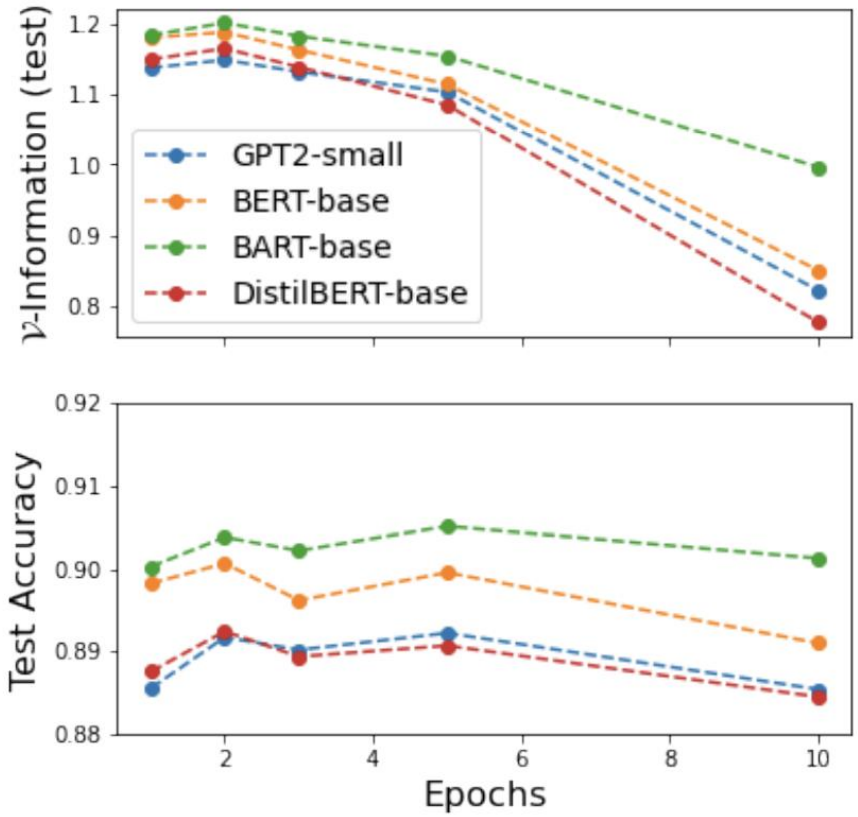


Figure 2. Comparing the $\mathcal{V}$-usable information estimate to accuracy in SNLI. In the first three epochs, estimates on the test set are similar across all models (top), but due to over-fitting, the estimates diverge and decline. The test accuracy (bottom) for each model loosely tracks the $\mathcal{V}$-information estimate for that model, since extracting information makes prediction easier.

- 大模型 accuracy 与 $\mathcal{V}$-Usable Information 都更高，因为提取更多的信息让识别更容易
- $\mathcal{V}$ Information 相比 accuracy 对过拟合更加敏感

$\mathcal{V}$-Usable Information in Practice

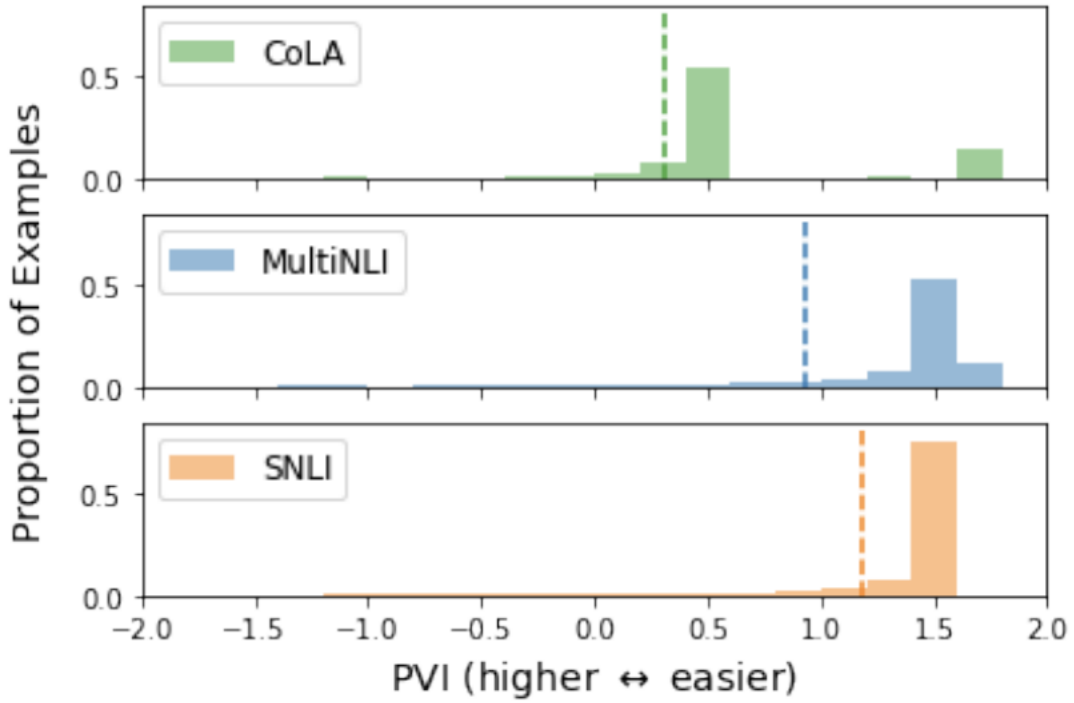


Figure 1. The Stanford NLI dataset contains more BERT-usable information than the MultiNLI and CoLA datasets, making it easier for BERT-base. Above, the distribution of instance difficulty (PVI) in the held-out sets for each; dotted lines denote the average PVI.

- 提供了衡量不同数据集难度的方法

Pointwise $\mathcal{V}$-Information

Given random variables X, Y and a predictive family $\mathcal{V}$, the pointwise $\mathcal{V}$-information (PVI) of an instance (x, y) is

$$\mathrm{PVI}(x \to y) = -\log_2 g[\varnothing](y) + \log_2 g'[x](y)$$

where $g \in \mathcal{V}$ s.t. $\mathbb{E}[-\log g[\varnothing](Y)] = H_{\mathcal{V}}(Y)$ and $g' \in \mathcal{V}$ s.t. $\mathbb{E}[-\log g'[X](Y)] = H_{\mathcal{V}}(Y \mid X)$

PVI is to $\mathcal{V}$-information what PMI is to Shannon information:

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{x,y\sim P(X,Y)}[\mathrm{PMI}(x,y)] \\ I_{\mathcal{V}}(X \to Y) &= \mathbb{E}_{x,y\sim P(X,Y)}[\mathrm{PVI}(x \to y)] \end{aligned}$$

Algorithm

Algorithm 1 After finetuning on a dataset of size n, the $\mathcal{V}$-information and PVI can be calculated in $O(n)$ time.

Input: training data $\mathcal{D}_{\text{train}} = \{(\text{input } x_i, \text{gold label } y_i)\}_{i=1}^{m}$, held-out data $\mathcal{D}_{\text{test}} = \{(\text{input } x_i, \text{gold label } y_i)\}_{i=1}^{n}$, model $\mathcal{V}$

do

$g' \leftarrow$ Finetune $\mathcal{V}$ on $\mathcal{D}_{\text{train}}$

$\varnothing \leftarrow$ empty string (null input)

$g \leftarrow$ Finetune $\mathcal{V}$ on $\{(\varnothing, y_i) \mid (x_i, y_i) \in \mathcal{D}_{\text{train}}\}$

$H_\mathcal{V}(Y), H_\mathcal{V}(Y|X) \leftarrow 0, 0$

for $(x_i, y_i) \in \mathcal{D}_{\text{test}}$ **do**

$H_\mathcal{V}(Y) \leftarrow H_\mathcal{V}(Y) - \frac{1}{n} \log_2 g[\varnothing](y_i)$

$H_\mathcal{V}(Y|X) \leftarrow H_\mathcal{V}(Y|X) - \frac{1}{n} \log_2 g'[x_i](y_i)$

$\text{PVI}(x_i \to y_i) \leftarrow -\log_2 g[\varnothing](y_i) + \log_2 g'[x_i](y_i)$

end for

$\hat{I}_\mathcal{V}(X \to Y) = \frac{1}{n} \sum_i \text{PVI}(x_i \to y_i) = H_\mathcal{V}(Y) - H_\mathcal{V}(Y|X)$

end do

PVI in Practice

Sentence	Label	PVI
Wash you!	No	-4.616
Who achieved the best result was Angela.	No	-4.584
Sue gave to Bill a book.	No	-3.649
Only Churchill remembered Churchill giving the Blood, Sweat and Tears speech.	No	-3.571
Cynthia chewed.	No	-3.510
It is a golden hair.	Yes	-3.251
I won't have some money.	No	-3.097
You may pick every flower, but leave a few for Mary.	No	-2.875
I know which book Mag read, and which book Bob said that you hadn't.	Yes	-2.782
John promise Mary to shave himself.	Yes	-2.609

Table 1. The 10 hardest (lowest PVI) instances in the CoLA in-domain test set for grammaticality detection (label indicates grammaticality), according to BERT-base. Examples in red are assessed to be mislabelled by authors of this work. For e.g., 'Cynthia chewed.' might be grammatical because the verb 'chew' could be intransitive in this usage. This suggests that PVI could be used to identify mislabelled examples. All of these examples were predicted incorrectly by BERT-base.

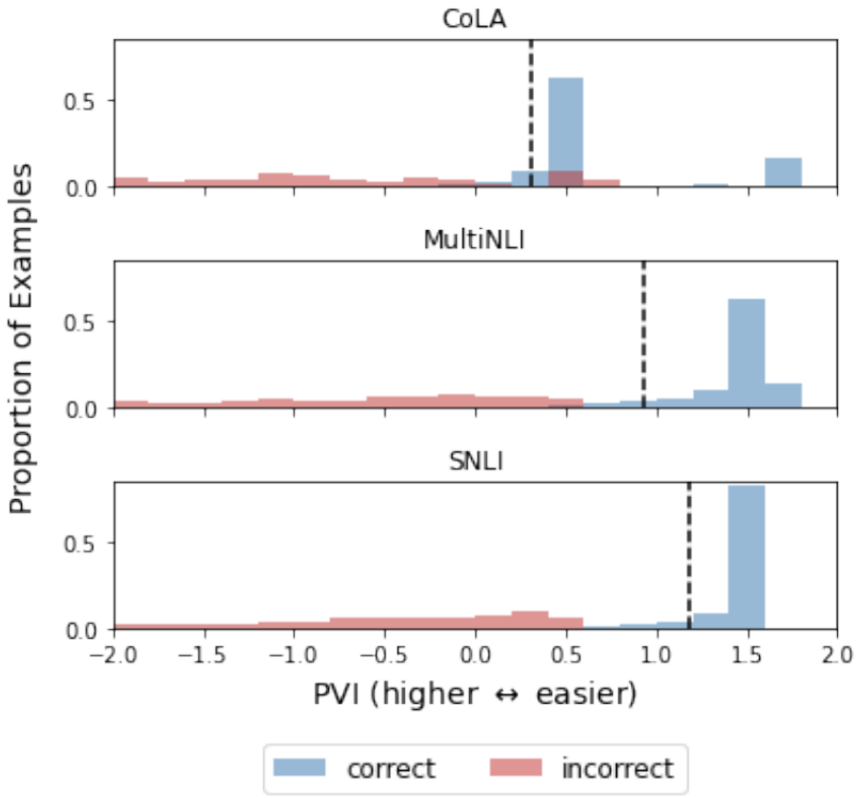
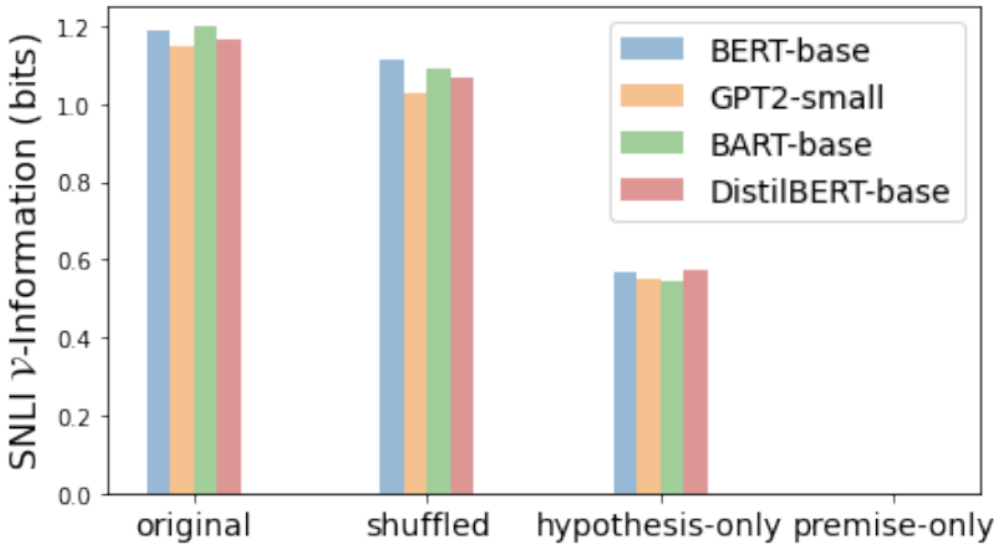


Figure 3. The distribution of PVI for correctly and incorrectly predicted instances in each dataset. Note that the point at which instances start being incorrectly predicted is similar across datasets (~ 0.5 bits). In contrast, because the label space is different across CoLA and the other two datasets, such a comparison could not be made with a performance-based metric.

Input Transformations

Attribute	Transformation	Transformed Input
Original		PREMISE: Two girls kissing a man with a black shirt and brown hair on the cheeks. HYPOTHESIS: Two girls kiss.
Shuffled	shuffle tokens randomly	PREMISE: girls two a kissing man with a black cheeks shirt and hair brown on the . HYPOTHESIS: kiss two . girls
Hypothesis-only	only include hypothesis	HYPOTHESIS: Two girls kiss.
Premise-only	only include premise	PREMISE: Two girls kissing a man with a black shirt and brown hair on the cheeks.
Overlap	hypothesis-premise overlap	PREMISE: Two girls [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] . HYPOTHESIS: Two girls [MASK] .

Table 7. Given an NLI instance (see 'Original'), each transformation isolates some attribute from the input. The headers 'PREMISE' and 'HYPOTHESIS' were added by us to transform the two sentence inputs into a single text input for all models that were evaluated.



Slicing Datasets

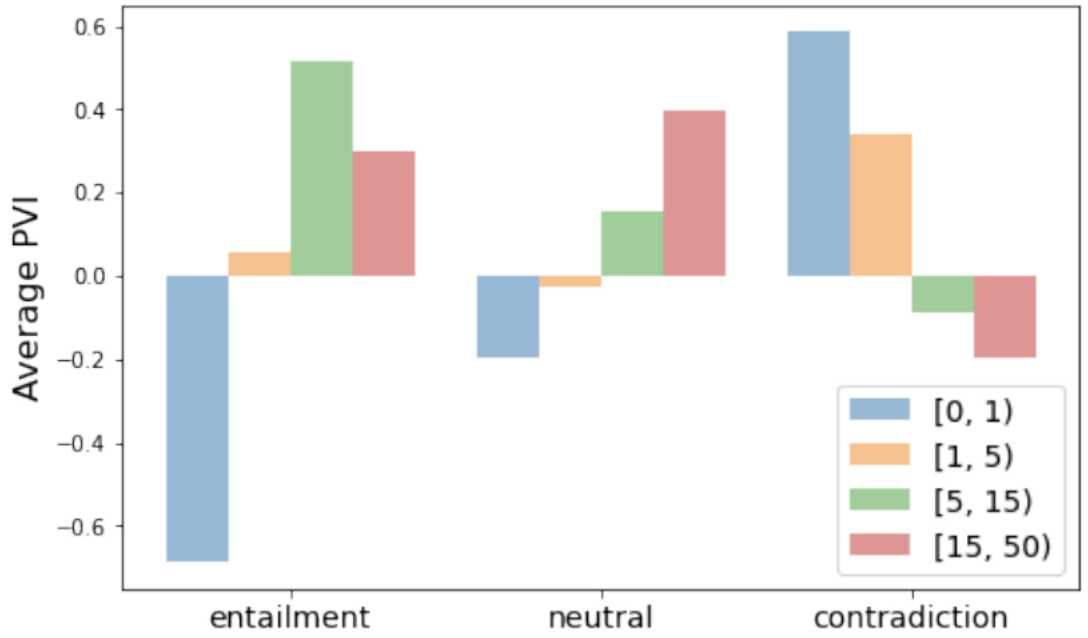


Figure 5. The mean PVI of SNLI instances according to BERT-base, broken down by the overlap length (i.e., the number of tokens shared by the hypothesis and premise). Entailment examples with no overlap are the most difficult (i.e., lowest mean PVI).

Thanks for watching!

2023-03-08