Information Bottleneck for Controlling Conciseness in Rationale Extraction

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, Luke Zettlemoyer

Code https://github.com/bhargaviparanjape/explainable_qa
Webpage https://bhargaviparanjape.github.io/
Motivation

- Complex SOTA models for Text Classification, Question Answering, Fact Verification, etc. are black boxes

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles ... It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt ....
Label: Yes

Question Answering

Context:
Beware of movies with the director's name in the title. Take John Carpenter's ghosts of mars ( please ) ... this embarrassment would surely have bypassed theaters entirely and gone straight to its proper home on the USA network ... the latest from the director of Starman, Halloween, and Escape from New York is a lousy western all gussied up to look like a futuristic horror flick. For future generations. A matriarchal society ... Well, don't get your hopes up.
Label: Negative

Text Classification
Rationales

• Complex SOTA models are black boxes

• Tasks: Text Classification, Question Answering, Fact Verification

• Humans highlight <25% of input as evidence to explain their decision

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles ... it can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt ....
Label: Yes

Question Answering

Context:
Beware of movies with the director's name in the title. Take John Carpenter's ghosts of mars ( please ) ... this embarrassment would surely have bypassed theaters entirely and gone straight to its proper home on the USA network ... the latest from the director of Starman, Halloween, and Escape from New York is a lousy western all gussied up to look like a futuristic horror flick. For future generations. A matriarchal society ... Well, don't get your hopes up.
Label: Negative

Text Classification
Rationales

- Complex SOTA models are **black boxes**
- Tasks: Text Classification, Question Answering, Fact Verification
- Humans highlight <25% of input as evidence to explain their decision

- Rationale: A subsequence of input text that is **necessary** and **sufficient** for task decision
  - Sufficient - Concise
  - Necessary - Faithful

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles ... **It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt.**
Label: Yes

**Question Answering**

Context:
Beware of movies with the director's name in the title. Take John Carpenter's ghosts of mars ( please ) ... this embarrassment would surely have bypassed theaters entirely and gone straight to its proper home on the USA network ... the latest from the director of Starman, Halloween, and Escape from New York **is a lousy western** all gussied up to look like a futuristic horror flick. For future generations. A matriarchal society ... Well, don't get your hopes up.
Label: Negative

**Text Classification**
Faithfulness

- The rationale must **actually be used** for the model’s prediction.

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles. It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt, or to bake it, and then break it apart.

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles. It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt, or to bake it, and then break it apart.

True
Problem Definition

Extract subsequence of text that is necessary and sufficient for task decision

○ Sufficient (Conciseness)
○ Necessary (Faithfulness)

Outline

• Information Bottleneck Approach
• Model Architecture
• Experiments
• Results

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles ... It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt ....
Label: Yes

Question Answering
Faithfulness

- The rationale must be **necessary** for the model’s prediction
- Faithful model design\(^1\):
  - Explainer identifies rationale
  - Predictor conditions only on explainer’s prediction

---

**Query: Can pickling salt be used as table salt?**

**Context:**
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles …. It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt, or to bake it, and then break it apart.

---

**Query: Can pickling salt be used as table salt?**

**Context:**
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles …. It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt, or to bake it, and then break it apart.

---

**Supervision**
Accuracy-Conciseness Tradeoff

- Explainer can make mistakes, leading to performance loss of predictor
- Tradeoff between predictor’s accuracy and rationale conciseness

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles .... It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt, or to bake it, and then break it apart

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles .... It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt, or to bake it, and then break it apart
Accuracy-Conciseness Tradeoff

- Explainer can make mistakes, leading to performance loss of predictor.
- Rationale must be an optimally compressed representation of input:
  1. Conciseness: Minimally informative about the original input, and
  2. Accuracy: Maximally informative about the output label.

---

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles. It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt, or to bake it, and then break it apart.

---

Query: Can pickling salt be used as table salt?
Context:
PICKLING SALT Pickling salt is a salt that is used mainly for canning and manufacturing pickles. It can be used in place of table salt, although it can cake. A solution to this would be to add a few grains of rice to the salt, or to bake it, and then break it apart.

---

True
Information Bottleneck (IB) Principle

**Information Bottleneck**: Find best tradeoff between accuracy and compression (conciseness)

Setup: A random variable $X$ that is predictive of observed variable $Y$

IB Objective: Find a compressed representation of $X$ termed **bottleneck variable $Z$** that can best predict $Y$.
Information Bottleneck Principle

If $X$ is predictive of $Y$, IB finds a compressed representation termed bottleneck variable $Z$ that best predicts $Y$.

Objective: $Z$ should be **minimally informative** about $X$ and **maximally informative** about $Y$.

Compression (Conciseness) term

$$\min_{p(z|x)} I(X; Z)$$

$I(\cdot; \cdot)$ is mutual information
Information Bottleneck Principle

If $X$ is predictive of $Y$, IB finds a compressed representation termed **bottleneck variable $Z$** that best predicts $Y$.

Objective: $Z$ should be *minimally informative* about $X$ and *maximally informative* about $Y$.

\[
\min_{p(z|x)} I(X; Z) - \beta I(Z; Y)
\]

$I(;;)$ is mutual information
Information Bottleneck Principle

If $X$ is predictive of $Y$, IB aims to find a compressed representation termed bottleneck variable $Z$ of that best predicts $Y$.

Objective: $Z$ should be minimally informative about $X$ and maximally informative about $Y$.

$$\min_{p(z|x)} I(X; Z) - \beta I(Z; Y)$$

$\beta$ is the tradeoff parameter.

$I(\cdot; \cdot)$ is mutual information.
Variational Information Bottleneck

- Variational lower bound on mutual information for gradient-based optimization\textsuperscript{[2]}
Variational Information Bottleneck

• Variational lower bound on mutual information for gradient-based optimization\(^2\)

• Objective to minimize:
  
  ◦ Task Loss: Likelihood of predicting \(y\) from \(z\)

\[
L_{VIB} = \mathbb{E}_{z \sim p(z|x)} \left[ - \log q_\phi(y|z) \right]
\]

Task Loss
Relevance (Accuracy) term
Variational Information Bottleneck

- Variational lower bound on mutual information for parametric optimization\textsuperscript{[2]}

- Objective to minimize:
  - Task Loss: Likelihood of predicting $y$ from $z$
  - Information Loss: Divergence between posterior $p(z|x)$ and a prior $r(z)$ that contains no information about $x$.

\[
L_{VIB} = \mathbb{E}_{z \sim p_{\theta}(z|x)}[-\log q_{\phi}(y|z)] + \beta KL[p_{\theta}(z|x), r(z)],
\]

- Relevance (Accuracy) term
- Compression (Conciseness) term
Variational IB for Interpretability

• Shortcoming of the previous formulation: Bottleneck representation $z$ is not interpretable as the rationale!

• Our formulation: $X$ is a sequence of words or sentences and $Z$ is constrained to be human readable subsequence in $X$

Masked version of the input
Variational IB for Interpretability

- **Interpretable Information Bottleneck Formulation:**
  - Input $x$ is a sequence of words/sentences
  - Binary mask vector $m$ of same size as $x$
  - Bottleneck $z$ is obtained by masking $x$ with a binary vector $m$
Variational IB for Interpretability

- **Interpretable Information Bottleneck Formulation:**
  - Input $x$ is a sequence of words/sentences
  - Binary mask vector $m$ of same size as $x$
  - Bottleneck $z$ is obtained by masking $x$ with a binary vector $m$

\[
L_{IVIB} = \mathbb{E}_{m \sim p_\theta(m|x)}[- \log q_\phi(y|m \odot x)] + \beta \sum_j KL[p_\theta(m_j|x)||r(m_j)],
\]

- **Task Loss**
  - Relevance (Accuracy) term
- **Information Loss**
  - Compression (Conciseness) term
Our Approach: Sparse IB

\[ L_{IVIB} = \mathbb{E}_{m \sim p_\theta(m|x)}[- \log q_\phi(y|m \odot x)] + \beta \sum_j KL[p_\theta(m_j|x) || r(m_j)], \]

Apply knowledge of how sparse the mask should be to assign the prior over mask, \( r(m) \) a fixed value \( \pi \)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>% Input as Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEVER</td>
<td>20.0</td>
</tr>
<tr>
<td>MultiRC</td>
<td>17.4</td>
</tr>
<tr>
<td>Movies</td>
<td>19.1</td>
</tr>
<tr>
<td>BoolQ</td>
<td>6.6</td>
</tr>
<tr>
<td>Evidence</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Table: % of input masked as rationale by humans can be used as \( \pi \)
Model Architecture

- Two independent transformer-based explainer and predictor models
Model Architecture - Explainer

- Input consists of a sequence of sentences $x_1, x_2, \ldots x_i, \ldots, x_n$

- Explainer predicts posterior probability $p(m_i \mid x)$ that $i^{th}$ sentence is in rationale.
Model Architecture - Sampling

- **Bernoulli** distribution with $p(m_i | x)$ used to sample binary mask value $m_i$
- Gumbel softmax trick to reparameterize $m_i$ for end-to-end differentiability

![Diagram](image)
Model Architecture - Predictor

- Predictor/Classifier applies the sampled sentence mask over its input representation while predicting task label
Experiments

Five text classification tasks from the ERASER benchmark (DeYoung et al., 2019):

- **Movies** sentiment analysis
- **FEVER** fact verification
- **MultiRC** and **BoolQ** reading comprehension datasets
- **Evidence inference** over scientific text.

All these datasets have *sentence-level rationale annotations* for validation and test sets.
Experiments

• Five text classification tasks from the ERASER benchmark (DeYoung et al., 2019). All these datasets have sentence-level rationale annotations for validation and test sets.

Evaluation Metrics

• Task Performance: Macro F1 for classification tasks
• Rationale Performance: Token-level macro F1 of predicted rationale to gold annotations
Baseline Approaches - Sparse Norm

- Previous work\textsuperscript{[1]} minimize norm of the mask vector for conciseness.

- Value of norm is no smaller than the value of the prior $\pi$

$$L_{IVIB} = \mathbb{E}_{m \sim p_\theta(m|x)}[-\log q_\phi(y|m \odot x)] + \lambda \max (0, ||m|| - \pi)$$

Same as the fixed prior used in information loss!
Baseline Approaches - No Conciseness Loss

- Baseline that does not use the information loss term for optimization

\[ L_{IVIB} = \mathbb{E}_{m \sim p_{\theta}(m|x)} \left[ -\log q_{\phi}(y|m \odot x) \right] + \lambda \max(0, ||m|| < \pi) \]
Results - Task Performance

- No Conciseness Loss
- Sparse Norm
- Sparse IB (Us)

Task F1

- Evidence Inf
- MultiRC
- BoolQ
- Movies
- FEVER
Results - Task Performance

- No Conciseness Loss
- Sparse Norm-Controlled
- Sparse IB (Us)

Task F1

- Evidence Inf.
- MultiRC
- BoolQ
- Movies
- FEVER
Discussion: Controlled Sparsity

Sparse IB (Ours)

• Achieves desired prior sparsity in expectation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$\pi$</th>
<th>Sparse Norm-C</th>
<th>Sparse IB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Var</td>
<td>Mean</td>
</tr>
<tr>
<td>FEVER</td>
<td>0.20</td>
<td>0.17</td>
<td>0.94</td>
</tr>
<tr>
<td>MultiRC</td>
<td>0.25</td>
<td>0.11</td>
<td>1.14</td>
</tr>
<tr>
<td>Movies</td>
<td>0.40</td>
<td>0.38</td>
<td>2.90</td>
</tr>
<tr>
<td>BoolQ</td>
<td>0.20</td>
<td>0.04</td>
<td>0.84</td>
</tr>
<tr>
<td>Evidence</td>
<td>0.20</td>
<td>0.10</td>
<td>1.17</td>
</tr>
</tbody>
</table>

Table 2: Average mask length (sparsity) attained by Sparse IB and the Sparse Norm-C baseline for a given prior $\pi$ for different tasks, averaged over 100 runs.
Discussion: Controlled Sparsity

Sparse IB (Ours)
- Achieves desired prior sparsity in expectation
- Is able to adapt to different examples

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$\pi$</th>
<th>Sparse Mean</th>
<th>Norm-C Mean</th>
<th>Sparse IB Mean</th>
<th>Norm-C Var</th>
<th>Sparse IB Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEVER</td>
<td>0.20</td>
<td>0.17</td>
<td>0.94</td>
<td>0.21</td>
<td>1.24</td>
<td></td>
</tr>
<tr>
<td>MultiRC</td>
<td>0.25</td>
<td>0.11</td>
<td>1.14</td>
<td>0.26</td>
<td>1.67</td>
<td></td>
</tr>
<tr>
<td>Movies</td>
<td>0.40</td>
<td>0.38</td>
<td>2.90</td>
<td>0.42</td>
<td>3.02</td>
<td></td>
</tr>
<tr>
<td>BoolQ</td>
<td>0.20</td>
<td>0.04</td>
<td>0.84</td>
<td>0.22</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>Evidence</td>
<td>0.20</td>
<td>0.10</td>
<td>1.17</td>
<td>0.20</td>
<td>1.61</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Average mask length (sparsity) attained by Sparse IB and the Sparse Norm-C baseline for a given prior $\pi$ for different tasks, averaged over 100 runs.
Results - Task Performance

- Evidence Inf.
- MultiRC
- BoolQ
- Movies
- FEVER

Task F1

- No Conciseness Loss
- Sparse Norm-Controlled
- Sparse IB (Us)
- Full Input
Results - Semisupervised

• Use limited rationale supervision to close gap with a model that uses full input

• Replacing information loss term with cross-entropy term between predicted mask and human-annotated gold mask

\[
L_{IVIB} = \mathbb{E}_{m \sim p_\theta(m|x)} \left[ -\log q_\phi(y|m \odot x) \right] + \gamma \sum_j -\hat{m}_j \log p(m_j|x)
\]

- Task Loss
- Rationale Loss
Task accuracy gap can be bridged with <50% annotations for rationales with diminishing returns as more annotated data is used.

Task Performance vs. % of Rationale Annotations
FEVER (left), MultiRC (right)
Conclusion

• Faithful and interpretable model using information bottleneck that jointly optimizes for **conciseness of rationale** and **accuracy of task**.

• Improvement in task and rationale performance over prior work

• Nears performance of full-input model with <50% annotation for rationales
Thank you!