A Game Theoretic Approach to Class-wise Selective Rationalization

Shiyu Chang†1,2, Yang Zhang†1,2, Mo Yu†2, Tommi S. Jaakkola3

1 MIT-IBM Watson AI Lab
2 IBM Research
3 Computer Science and Artificial Intelligence Lab, MIT

† Indicates authors with equal contribution
What Are Rationales?

Rationales as extracted text-spans should be

- Short and coherent
- Sufficient for the correct prediction

This beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. This is a real good lookin' beer.

Ratings:

Look: 5 stars

How to Generate Rationales?

Two modular components $\text{gen()}$ and $\text{enc()}$. 

Generator $\text{gen}(x)$

Encoder $\text{enc}(z)$
How to Generate Rationales?

Input $x$

Generator $\text{gen}(x)$

Encoder $\text{enc}(z)$

Distribution over possible rationales $P(z \mid x)$

- this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ...
- this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ...
- this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ...
- this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ...

0.8
0.01
0.02
0.05
0.1

Generator specifies the distribution of rationales
How to Generate Rationales?

This beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. This is a real good lookin' beer, unfortunately it gets worse from here ...

Encoder $\text{enc}(z)$

Generator $\text{gen}(x)$

Distribution over possible rationales $P(z|x)$

0.05

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. This is a real good lookin' beer, unfortunately it gets worse from here ...

0.02

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam, this is a real good lookin' beer, unfortunately it gets worse from here ...

0.8

0.1

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam, this is a real good lookin' beer, unfortunately it gets worse from here ...

0.01

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam, this is a real good lookin' beer, unfortunately it gets worse from here ...

0.05

0.02

0.1

0.8

z

Encoder makes prediction based on rationales
How to Generate Rationales?

Input $x$

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ...

Generator $\text{gen}(x)$

Encoder $\text{enc}(z)$

Distribution over possible rationales $P(z|x)$

Two components optimized jointly
How Good Are Generated Rationales?

a beer that is not sold in my neck of the woods, but managed to get while on a roadtrip. poured into an imperial pint glass with a generous head that sustained life throughout. nothing out of the ordinary here, but a good brew still. body was kind of heavy, but not thick. the hop smell was excellent and enticing. very drinkable

More examples available at: https://github.com/taolei87/rcnn/tree/master/code/rationale
Why Class-wise Rationalization?

Input

Conventional Rationalization

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.
Why Class-wise Rationalization?

Conventional Rationalization

Input

a nice and clean hotel to stay for business and leisure. **but the location is not good if you need public transport.** […] i never tried.

Label

😊
Why Class-wise Rationalization?

Conventional Rationalization

Input: a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. […] i never tried.

Label: 😞

Class-wise Rationalization

What is the possible explanation if one give a score/label y?
a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.
a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

Class-wise Rationalization Setup

Input

Target

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.
a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.
Class-wise Rationalization Setup

If ground-truth is 😞

Input

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

Target

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

If ground-truth is 😄

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.
Class-wise Rationalization Setup

If ground-truth is 😞

Input

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

Target

, 😞

Factual Rationale

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

, 😄

Counterfactual Rationale

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.
Class-wise Adversarial Rationalization

Two groups with total six components for a binary case
Class-wise Adversarial Rationalization

Two factual generators
Class-wise Adversarial Rationalization

Two factual generators
Class-wise Adversarial Rationalization

Two counterfactual generators
Class-wise Adversarial Rationalization

Two counterfactual generators
Class-wise Adversarial Rationalization

Two discriminators
Class-wise Adversarial Rationalization

Two discriminators
Class-wise Adversarial Rationalization

\[ x \sim p(x \mid y = 0) \]
\[ x \sim p(x \mid y = 1) \]

First group: class-0 group
Class-wise Adversarial Rationalization

Second group: class-1 group
**Class-wise Adversarial Rationalization**

**Discriminator:**
Identify if the rationale come from the factual generator

\[
x \sim p(x \mid y = 0)
\]

\[
x \sim p(x \mid y = 1)
\]

\[
\text{Generator } g^f_0(\cdot)
\]

\[
\text{Generator } g^c_0(\cdot)
\]

\[
\text{Discriminator } d_0(\cdot)
\]
Class-wise Adversarial Rationalization

Generators:
Try to make the discriminator believe its the factual generator
How Does It Work

Input

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

(1, 0, …, 1, 1)

Rationale

a nice and clean hotel to stay for business and leisure. but the location is not good if you need public transport. [...] i never tried.

(1, 0, …, 0, 1)

The dimensions of the input and rationale are independent conditional on the label
How Does It Work

Probability

Word index
How Does It Work

$p_{x|y}(1|0)$

$p_{x|y}(1|1)$
How Does It Work
The optimal strategy for the counterfactual generator given arbitrary factual selections
How Does It Work

The optimal strategy for the factual generator
Experimental Setup

Tasks:
Sentiment analysis on single/multi aspect reviews.

Data:
Amazon review, BeerAdvocate review, TripAdvisor hotel review

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Train</th>
<th>Dev</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotel Review</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>14,472</td>
<td>1,812</td>
<td>200</td>
</tr>
<tr>
<td>Service</td>
<td>101,484</td>
<td>12,688</td>
<td>199</td>
</tr>
<tr>
<td>Cleanliness</td>
<td>150,098</td>
<td>18,764</td>
<td>196</td>
</tr>
</tbody>
</table>
Objective Result

- Location:
  - RNP: 0.5
  - Post-Exp: 0.3
  - CAR: 0.4

- Service:
  - RNP: 0.4
  - Post-Exp: 0.3
  - CAR: 0.3

- Cleanliness:
  - RNP: 0.3
  - Post-Exp: 0.2
  - CAR: 0.3
Rationale Visualization

**Hotel - Service**

Label - Positive

i would definitely recommend this hotel to anyone who has plans at the staples center or nokia theater. the staff was friendly and helpful, the rooms were clean, and there was plenty of entertainment and dining options within walking distance. they provide robes, slippers, and wonderful toiletries! the shower was beautiful but the head was not working properly and we requested a non smoking room but there were n't any available. the smell of stinky cigarettes really disgusted me but if they had a non smoking room available i honestly would not have any complaints. i recommend taking a car to and from the airport. i would stay at this hotel again for sure.

**Hotel - Service**

Label - Negative

i loved the hotel and the hotel room. my daughter and i were very impressed. we had a lovely room and for being a hilton honors member we even received a free fruit basket and two bottles of water. please beware however and always check your bank account for any additional charges. we had used my daughters credit card to hold the room but asked for my visa debit card to be charged for the final amount. we were both charged for the room and it took numerous phone calls to receive a credit. we were also charged for several additional charges that we should not have been. the billing department was extremely rude as we called them numerous times to get the matter straightened out. as for the hotel i would highly recommend it and also for it's wonderful location and shuttle. i just did n't appreciate the cutomer service afterwards and the fact that it took me so long to get the matter taken care of. just be very careful looking over your final billing.
Subjective Evaluation

Please read the following partially-blocked review on a hotel and 1) guess the aspect that the revealed part is talking about; and 2) guess the score on the aspect that you selected.

Pretty good and staff was friendly / helpful

Select an option

- location - negative 1
- location - positive 2
- service - negative 3
- service - positive 4
- cleanliness - negative 5
- cleanliness - positive 6

Submit
Subjective Result

<table>
<thead>
<tr>
<th>Location</th>
<th>Service</th>
<th>Cleanliness</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNP</td>
<td>Post-Exp</td>
<td>CAR</td>
</tr>
<tr>
<td>0</td>
<td>0.15</td>
<td>0.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
<td>0.45</td>
</tr>
<tr>
<td>0.35</td>
<td>0.5</td>
<td>0.65</td>
</tr>
<tr>
<td>0.5</td>
<td>0.55</td>
<td>0.7</td>
</tr>
<tr>
<td>0.65</td>
<td>0.6</td>
<td>0.8</td>
</tr>
</tbody>
</table>

CAR-generated rationale is better for both factual and counterfactual cases.