Interpretable Neural Predictions with Differentiable Binary Variables

Joost Bastings    Wilker Aziz    Ivan Titov

ILLC, University of Amsterdam
ILCC, University of Edinburgh

bastings.github.io

ACL, 30 July 2019
Interpreting neural networks is difficult

Lots of documents → Classifier → Results
Interpreting neural networks is difficult

Lots of documents → Classifier → Results

Can I trust these results?
Interpreting neural networks is difficult

Lots of documents → Classifier → Results

Does it make sense?

Can I trust these results?
Interpreting neural networks is difficult

Lots of documents

Classifier

Results

What happens here?

Does it make sense?

Can I trust these results?
pours a dark amber color with decent head that does not recede much. It's a tad too dark to see the carbonation, but fairs well. Smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow.
pours a dark amber color with decent head that does not recede much. It’s a tad too dark to see the carbonation, but fairs well. Smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow.
pours a dark amber color with decent head that does not recede much. It's a tad too dark to see the carbonation, but fairs well. Smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow.
A rationale is a short and sufficient part of the input text. To be a good explanation, it must be enough to make the right prediction.
Text classification with rationales (Lei et al., 2016)

\[ Z_i \mid x \sim \text{Bernoulli}(g_i(x; \phi)) \]
Text classification with rationales (Lei et al., 2016)

\[ Y \mid x, z \sim \text{Cat}(f(x \odot z; \theta)) \]

\[ Z_i \mid x \sim \text{Bernoulli}(g_i(x; \phi)) \]

Classifier

Rationale Extractor

\( z_i \in \{0,1\} \) can erase \( x_i \)

NN that predicts a sequence \( n \) of Bernoulli parameters
Model of Lei et al. (2016)
Model of Lei et al. (2016)

Classifier

$\theta$

$z_1 x_1$

$z_1 \sim \text{Bern}(b_1)$

$z_2 x_2$

$z_2 \sim \text{Bern}(b_2)$

$z_3 x_3$

$z_3 \sim \text{Bern}(b_3)$

$z_4 x_4$

$z_4 \sim \text{Bern}(b_4)$

$z_5 x_5$

$z_5 \sim \text{Bern}(b_5)$

Rationale Extractor

$\phi$

$x_1$

$b_1$

$x_2$

$b_2$

$x_3$

$b_3$

$x_4$

$b_4$

$x_5$

$b_5$
Model of Lei et al. (2016)

\[
\begin{align*}
\theta & \quad \text{Classifier} \\
\phi & \quad \text{Rationale Extractor} \\
\end{align*}
\]

\[
\begin{align*}
& \quad \text{No gradient} \\
& \quad \text{No gradient} \\
& \quad \text{No gradient} \\
& \quad \text{No gradient} \\
& \quad \text{No gradient} \\
\end{align*}
\]

\[
\begin{align*}
& \quad b_1 \\
& \quad b_2 \\
& \quad b_3 \\
& \quad b_4 \\
& \quad b_5 \\
\end{align*}
\]

\[
\begin{align*}
& \quad x_1 \\
& \quad x_2 \\
& \quad x_3 \\
& \quad x_4 \\
& \quad x_5 \\
\end{align*}
\]

\[
\begin{align*}
z_1x_1 & \sim \text{Bern}(b_1) \\
z_2x_2 & \sim \text{Bern}(b_2) \\
z_3x_3 & \sim \text{Bern}(b_3) \\
z_4x_4 & \sim \text{Bern}(b_4) \\
z_5x_5 & \sim \text{Bern}(b_5) \\
\end{align*}
\]

\[
\phi \quad \text{trained using REINFORCE}
\]
Our Proposed Model

Rationale Extractor

\[ \phi \]

\[ \begin{align*}
    a_1 & \quad b_1 \\
    a_2 & \quad b_2 \\
    a_3 & \quad b_3 \\
    a_4 & \quad b_4 \\
    a_5 & \quad b_5 \\
\end{align*} \]
Our Proposed Model

independent uniform samples $u_i \sim U(0, 1)$

Rationale Extractor $\phi$

gradient flows through the samples
Our Proposed Model

Classifier

\[ \theta \]

Rationale Extractor

\[ \phi \]

Independent uniform samples

\( u_i \sim U(0, 1) \)

Gradient flows through the samples
Our Proposed Model

Rationale Extractor $\phi$

Classifier $\theta$

independent uniform samples $u_i \sim U(0, 1)$

gradient flows through the samples
0 is **not** in the support

1 is **not** in the support

Stretch-and-rectify (Louizos et al., 2018)
0 is in the support now, but we will never sample it.
We collapse the shaded area into a point mass.
Stretch-and-rectify (Louizos et al., 2018)

Point mass at 0

Point mass at 1
In this work: Hard Kumaraswamy Distribution

- HardKuma(0.5, 0.5, -0.1, 1.1)
- Kuma(0.5, 0.5)
Kumaraswamy with various \((a, b)\)

\begin{align*}
\text{(0.5, 0.5)} \\
\text{(5, 1)} \\
\text{(2, 2)} \\
\text{(2, 5)} \\
\text{(0.1, 0.1)} \\
\text{(1.0, 1.0)}
\end{align*}
Getting short rationales

- We want **short** rationales without breaking backpropagation
- Solution: relax $L_0$ (Louizos et al., 2018)

$$L_0(z) \quad \text{Lei et al.} \quad \text{Compute } L_0 \text{ for one specific assignment of } z$$

$$\mathbb{E}_{p(z|x)}[L_0(z)] \quad \text{Proposed model:} \quad \text{L}_0 \text{ computed for all assignments on expectation}$$
Getting coherent rationales

- **Baseline:** penalty for transitions using fused lasso
  \[
  \sum_{i=1}^{n-1} |z_i - z_{i+1}|
  \]

- Proposed model: compute a relaxation of fused lasso by computing the expected number of zero-to-nonzero and nonzero-to-zero changes:

\[
\mathbb{E}_{p(z|x)} \left[ \sum_{i=1}^{n-1} \mathbb{I}[z_i = 0, z_{i+1} \neq 0] \right] + \mathbb{E}_{p(z|x)} \left[ \sum_{i=1}^{n-1} \mathbb{I}[z_i \neq 0, z_{i+1} = 0] \right]
\]
Specify target selection rate

- We want a maximum selection rate e.g. 10% of the text
- We propose a **constrained optimization** problem:

\[
\min_{\phi, \theta} L(\phi, \theta) \quad \text{s.t.} \quad \mathbb{E}[L_0] < r
\]

- We use Lagrangian relaxation
Experiments

1. Multi-aspect sentiment analysis (BeerAdvocate, Lei et al. 2016)
   - Regression, sentiment score in [0,1]

2. Stanford Sentiment (SST)
   - Classification {very negative, ..., very positive}

3. Stanford Natural Language Inference (SNLI)
   - Classification {entailment, contradiction, neutral}
## Beer Precision per Aspect

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Precision</th>
<th>% Selected</th>
<th>Precision</th>
<th>% Selected</th>
<th>Precision</th>
<th>% Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Look</td>
<td>80.6</td>
<td>13</td>
<td>88.4</td>
<td>7</td>
<td>65.3</td>
<td>7</td>
</tr>
</tbody>
</table>

### Multiple aspects

Attention (Lei et al.)

| Threshold | 80.6 | 13 | 88.4 | 7 | 65.3 | 7 |
# Beer Precision per Aspect

<table>
<thead>
<tr>
<th></th>
<th>Look</th>
<th></th>
<th>Smell</th>
<th></th>
<th>Taste</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>% Selected</td>
<td>Precision</td>
<td>% Selected</td>
<td>Precision</td>
</tr>
<tr>
<td>Attention (Lei et al.)</td>
<td>Threshold</td>
<td>80.6</td>
<td>13</td>
<td>88.4</td>
<td>7</td>
<td>65.3</td>
</tr>
<tr>
<td>Bernoulli / REINFORCE</td>
<td>Tuned $\lambda$</td>
<td>96.3</td>
<td>14</td>
<td>95.1</td>
<td>7</td>
<td>80.2</td>
</tr>
</tbody>
</table>

*Multiple aspects*
# Beer Precision per Aspect

<table>
<thead>
<tr>
<th></th>
<th>Look</th>
<th>Smell</th>
<th>Taste</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>% Selected</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention (Lei et al.)</td>
<td>Threshold</td>
<td>80.6</td>
<td>13</td>
</tr>
<tr>
<td>Bernoulli / REINFORCE (Lei et al.)</td>
<td>Tuned $\lambda$</td>
<td>96.3</td>
<td>14</td>
</tr>
<tr>
<td>HardKuma</td>
<td>Lagrange</td>
<td>98.1</td>
<td>13</td>
</tr>
</tbody>
</table>

*Multiple aspects*
Beer MSE/selection tradeoff (all aspects)
SST: accuracy/selection tradeoff

![Graph showing the accuracy of selected text]

- **HardKuma**
- **Bernoulli / REINFORCE**
Analysis: Word Count per Sentiment

- Total
- HardKuma
- Bernoulli

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Total</th>
<th>HardKuma</th>
<th>Bernoulli</th>
</tr>
</thead>
<tbody>
<tr>
<td>very negative</td>
<td>146</td>
<td>119</td>
<td>112</td>
</tr>
<tr>
<td>negative</td>
<td>992</td>
<td>603</td>
<td>489</td>
</tr>
<tr>
<td>neutral</td>
<td>18231</td>
<td>3378</td>
<td>3806</td>
</tr>
<tr>
<td>positive</td>
<td>1511</td>
<td>803</td>
<td>795</td>
</tr>
<tr>
<td>very positive</td>
<td>394</td>
<td>264</td>
<td>299</td>
</tr>
</tbody>
</table>
NLI: predict \{entailment, contradiction, neutral\} given premise & hypothesis

Baseline: Decomposable Attention model (Parikh et al., 2016)

We replace Hypothesis-Premise attention with HardKuma attention
### SNLI Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM (Bowman et al.)</td>
<td>–</td>
<td>80.6</td>
</tr>
<tr>
<td>DA (Parikh et al.)</td>
<td>–</td>
<td>86.3</td>
</tr>
<tr>
<td>DA (reimpl.)</td>
<td>86.9</td>
<td>86.5</td>
</tr>
<tr>
<td>DA HardKuma</td>
<td>86.0</td>
<td>85.5</td>
</tr>
</tbody>
</table>

Only drop 1% with 8.6% non-zero attention cells
Related work

- Rationales
- Sparsity
Related work

- Rationales
  - Zaidan et al. (NAACL’07)
  - Zaidan and Eisner (EMNLP’08)
  - Zhang et al. (EMNLP’16)

- Sparsity

Learn from rationales
Related work

Rationales

Sparsity

Local explanation / surrogate

Learn from rationales

Zaidan et al. (NAACL’07),
Zaidan and Eisner (EMNLP’08),
Zhang et al. (EMNLP’16)

LIME
Ribeiro et al. (2016)
Related work

Lei et al. (EMNLP’16)
Rationalizing Neural Predictions

Jointly train & learn rationales

Rationales

Learn from rationales

Zaidan et al. (NAACL’07), Zaidan and Eisner (EMNLP’08), Zhang et al. (EMNLP’16)

Sparsity

LIME
Ribeiro et al. (2016)

Local explanation / surrogate

Rationalizing Neural Predictions

Learn from rationales

Zaidan et al. (NAACL’07), Zaidan and Eisner (EMNLP’08), Zhang et al. (EMNLP’16)
Related work

Rationales

Jointly train & learn rationales

Learn from rationales

Sparsity

Lei et al. (EMNLP’16)
Rationalizing Neural Predictions

Local explanation / surrogate

LIME
Ribeiro et al. (2016)

Rationalizing Neural Predictions

Learn from rationales

Zaidan et al. (NAACL’07),
Zaidan and Eisner (EMNLP’08),
Zhang et al. (EMNLP’16)

Sparsity

Stretch & rectify

Louizos et al. (ICLR’18)
Sparsifying parameters
Related work

- **Rationales**
  - Local explanation / surrogate
    - LIME (Ribeiro et al., 2016)
  - Sparsity
    - Sparsifying parameters (Louizos et al., ICLR’18)
  - Learn from rationales
    - Zaidan et al. (NAACL’07), Zaidan and Eisner (EMNLP’08), Zhang et al. (EMNLP’16)
  - This work
    - Jointly train & learn rationales
    - Rationalizing Neural Predictions (Lei et al., EMNLP’16)
    - Learn from rationales (Zaidan et al.)
Related work

This work

Rationales

Jointly train & learn rationales

Rationalizing Neural Predictions

Lei et al. (EMNLP’16)

Local explanation / surrogate

Learn from rationales

LIME

Ribeiro et al. (2016)

Sparsity

Sparsifying parameters

Stretch & rectify

Analyzing Multi-Head Self-Attention

Voita et al. (ACL’19)

This work

Learn from rationales

Zaidan et al. (NAACL’07),
Zaidan and Eisner (EMNLP’08),
Zhang et al. (EMNLP’16)

Rationales

24

Rationales

Sparsity

Voita et al. (ACL’19)

Lei et al. (EMNLP’16)

Rationalizing Neural Predictions

Zaidan et al. (NAACL’07),
Zaidan and Eisner (EMNLP’08),
Zhang et al. (EMNLP’16)

LIME

Ribeiro et al. (2016)

Local explanation / surrogate

Sparsifying parameters

Stretch & rectify

Analyzing Multi-Head Self-Attention

Voita et al. (ACL’19)

Lei et al. (EMNLP’16)

Rationalizing Neural Predictions

Zaidan et al. (NAACL’07),
Zaidan and Eisner (EMNLP’08),
Zhang et al. (EMNLP’16)

LIME

Ribeiro et al. (2016)

Local explanation / surrogate

Sparsifying parameters

Stretch & rectify

Analyzing Multi-Head Self-Attention

Voita et al. (ACL’19)

Lei et al. (EMNLP’16)

Rationalizing Neural Predictions

Zaidan et al. (NAACL’07),
Zaidan and Eisner (EMNLP’08),
Zhang et al. (EMNLP’16)

LIME

Ribeiro et al. (2016)

Local explanation / surrogate

Sparsifying parameters

Stretch & rectify

Analyzing Multi-Head Self-Attention

Voita et al. (ACL’19)

Lei et al. (EMNLP’16)

Rationalizing Neural Predictions

Zaidan et al. (NAACL’07),
Zaidan and Eisner (EMNLP’08),
Zhang et al. (EMNLP’16)

LIME

Ribeiro et al. (2016)

Local explanation / surrogate

Sparsifying parameters

Stretch & rectify

Analyzing Multi-Head Self-Attention

Voita et al. (ACL’19)

Lei et al. (EMNLP’16)

Rationalizing Neural Predictions

Zaidan et al. (NAACL’07),
Zaidan and Eisner (EMNLP’08),
Zhang et al. (EMNLP’16)

LIME

Ribeiro et al. (2016)

Local explanation / surrogate

Sparsifying parameters

Stretch & rectify

Analyzing Multi-Head Self-Attention

Voita et al. (ACL’19)
Summary

- Differentiable approach to extractive rationales
  - Stretch and rectify using HardKuma
  - Support for binary outcomes
- Objective to specify the percentage of selected text
- Future work: interpretable QA / fact checking
- Code online: [github.com/bastings](https://github.com/bastings)
  - DIY: add a HardKuma layer to your classifier!
Thank you!

Code online @ github.com/bastings

Check out our new NMT toolkit for novices Joey NMT at github.com/joeynmt
Example: Contradiction

Prediction: entailment
Example: Entailment

Correct

Incorrect

Prediction: neutral
Example: Neutral

Incorrect Prediction: entailment

<table>
<thead>
<tr>
<th>&lt;s&gt;</th>
<th>They</th>
<th>are</th>
<th>in</th>
<th>the</th>
<th>desert</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>People</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>walking</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>through</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dirt</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>81</td>
</tr>
<tr>
<td>.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Correct

Incorrect

<table>
<thead>
<tr>
<th>&lt;s&gt;</th>
<th>A</th>
<th>dog</th>
<th>found</th>
<th>a</th>
<th>bone</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dog</td>
<td>0</td>
<td>0</td>
<td>89</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>gnawing</td>
<td>0</td>
<td>0</td>
<td>89</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>on</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bone</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Parameter estimation

\[
\log P(y | x) = \log \mathbb{E}_{P(z|x,\phi)} \left[ P(y | x, z, \theta) \right]
\]

\[
J_I \geq \mathbb{E}_{P(z|x,\phi)} \left[ \log P(y | x, z, \theta) \right]
\]

\[
= \mathcal{C}(\phi, \theta)
\]

We maximise this lower bound on the log-likelihood
Baseline Regularizers

Loss for **sufficient** rationales

\[
\min_{\phi, \theta} L(\phi, \theta) + \lambda_0 \sum_{i=1}^{n} z_i + \lambda_1 \sum_{i=1}^{n-1} |z_i - z_{i+1}|
\]

\(L_0\) for **short** rationales

Fused lasso for **coherent** rationales
How to get a HardKuma sample from a uniform variable

$F^{-1}(U \sim \mathcal{U}(0,1))$
From uniform source to Kuma samples

\[ F^{-1}_K(u; a, b) = \left( 1 - (1 - u)^{\frac{1}{b}} \right)^{\frac{1}{a}} \quad u \in [0,1] \]

\[ F^{-1}_Z(U; a, b) \sim \text{Kuma}(a, b) \quad U \sim \mathcal{U}(0,1) \]
We **stretch** the support of the Kuma to \((l, r)\):

\[
F_T(t; a, b, l, r) = F_K\left(\frac{(t - l)}{(r - l)}; a, b\right)
\]

And define a **rectified** random variable:

\[
H \sim \text{HardKuma}(a, b, l, r)
\]

by passing a Kuma sample \(t\) through a hard sigmoid:

\[
T \sim \text{Kuma}(a, b, l, r) \quad h = \min(1, \max(0, t))
\]
Rectified Kumaraswamy (Formal) (2)

- Sampling $h=0$ means sampling any $t \in (l,0]$
- with mass under Kuma:
  \[ \mathbb{P}(H = 0) = F_K \left( \frac{-l}{r-l}; a, b \right) \]
- Sampling $h=1$ means sampling any $t \in [1,r)$
- with mass under Kuma:
  \[ \mathbb{P}(H = 1) = 1 - F_K \left( \frac{1-l}{r-l}; a, b \right) \]
\[ \mathbb{E}_{p(z|x)} \left[ L_0(z) \right] \overset{\text{ind}}{=} \sum_{i=1}^{n} \mathbb{E}_{p(z_i|x)} \left[ \mathbb{I}[z_i \neq 0] \right] = \sum_{i=1}^{n} 1 - \mathbb{P}(Z_i = 0) , \]
We can also compute a relaxation of fused lasso by computing the expected number of zero-to-nonzero and nonzero-to-zero changes:

\[
\mathbb{E}_{p(z|x)} \left[ \sum_{i=1}^{n-1} \mathbb{I}[z_i = 0, z_{i+1} \neq 0] \right] + \mathbb{E}_{p(z|x)} \left[ \sum_{i=1}^{n-1} \mathbb{I}[z_i \neq 0, z_{i+1} = 0] \right]
\]

\[
\text{ind} = \sum_{i=1}^{n-1} \mathbb{P}(Z_i = 0)(1 - \mathbb{P}(Z_{i+1} = 0)) + (1 - \mathbb{P}(Z_i = 0))\mathbb{P}(Z_{i+1} = 0)
\]
Reparameterization Trick

\[ U \sim \mathcal{U}(0,1) \]

\[ F_X^{-1}(u) \sim X \]
Why are gradients possible?

- We consider the case where we need derivatives of a function \( L(u) \) of the underlying uniform variable \( u \), as when we compute reparameterized gradients in variational inference. By chain rule:

\[
\frac{\partial L}{\partial u} = \frac{\partial L}{\partial h} \times \frac{\partial h}{\partial t} \times \frac{\partial t}{\partial k} \times \frac{\partial k}{\partial u}
\]

- depends on differentiable observation model
- Derivative for hard sigmoid:
  - 0 for \( t<0 \),
  - 1 for \( 0<t<1 \),
  - 0 for \( t>1 \),
  - undef for \( t={0,1} \)
- Depends on Kuma inverse CDF, no challenge
- r-l