Rationalizing Neural Predictions

Tao Lei
Regina Barzilay  Tommi Jaakkola

EMNLP 2016
Motivation

• Complex (neural) models come at the cost of interpretability
• Applications often need interpretable justifications — rations.
Motivation

• Complex (neural) models come at the cost of interpretability
• Applications often need interpretable justifications — *rationales*.

There is no evidence of extranodal extension.
BREAST (RIGHT), EXCISIONAL BIOPSY:
**INVASIVE DUCTAL CARCINOMA** (SEE TABLE #1). DUCTAL CARCINOMA IN-SITU, GRADE 1. ATYPICAL DUCTAL HYPERPLASIA. LOBULAR NEOPLASIA (ATYPICAL LOBULAR HYPERPLASIA). TABLE OF PATHOLOGICAL FINDINGS #1 INVASIVE CARCINOMA

prediction: high risk of recurring cancer

*Doctors won’t trust machines, unless evidence is provided*
Motivation

• Complex (neural) models come at the cost of interpretability
• Applications often need interpretable justifications — *rationales*.

**Our goal:** make powerful models more interpretable by learning rationales behind the prediction
Problem Setup

Interpretability via providing concise evidence from input

Rationales (evidence) should be:
- short and coherent pieces
- sufficient for correct prediction

Rationales are not provided during training
in contrast to (Zaidan et al., 2007; Marshall et al., 2015; Zhang et al., 2016)

Use powerful neural nets to avoid accuracy loss
in contrast to (Thrun, 1995; Craven and Shavlik, 1996; Ribeiro et al., 2016)
Model Architecture

two modular components \( \text{gen}(\cdot) \) and \( \text{enc}(\cdot) \)
**Model Architecture**

**Encoder** $\texttt{enc}(z)$

**Generator** $\texttt{gen}(x)$

```
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 ...
```

```
distribution over possible rationales $P(z|x)$
```

- 0.8
- 0.02
- 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 ...
```

```
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 ...
```

```
0.05
0.01
```

```
generator specifies the distribution of rationales
```

7
Model Architecture

**Generator** $\text{gen}(x)$

**Encoder** $\text{enc}(z)$

**distribution over possible rationales** $P(z|\ x)$

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 ...

The encoder makes prediction given rationale $z$. The distribution over possible rationales is shown with probabilities $0.8$, $0.02$, $0.1$, $0.05$, and $0.01$. The prediction $y$ is made based on the encoded rationale.
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...

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

0.8

0.02

0.1

0.05

0.01

two components optimized jointly
Generator Implementations

binary selection $z$: 0 1 0 1 1

$P(z)$: 0 0 0 0 0

hidden states: 0 0 0 0 0

input words $x$: 0 0 1 1 1

independent selection, feedforward net
Generator Implementations

binary selection $z$: 0 1 0 1 1

$P(z)$:

hidden states:

input words $x$:

independent selection, bi-directional RNNs
Generator Implementations

binary selection \( z \): 0 1 0 1 1

\( P(z) \): 0 1 0 1 1

hidden states: 0 1 0 1 1

input words \( x \): 0 1 0 1 1

dependent selection, bi-directional RNNs

choose networks based on the data/application
Training Objective

$$\text{cost}(\mathbf{z}, \mathbf{y}) = \text{loss}(\mathbf{z}, \mathbf{y}) + \lambda_1 |\mathbf{z}|_1 + \lambda_2 \sum_i |\mathbf{z}_i - \mathbf{z}_{i-1}|$$

- **sufficiency**: correct prediction
- **sparsity**: rationale is short
- **coherency**: continuous selection

- receive this training signal after $\mathbf{z}$ is produced

**Minimizing expected cost:**

$$\min_\theta \sum_{(\mathbf{x}, \mathbf{y}) \in D} \mathbb{E}_{\mathbf{z} \sim \text{gen}(\mathbf{x})} \left[ \text{cost}(\mathbf{z}, \mathbf{y}) \right]$$

- intractable because summation over $\mathbf{z}$ is exponential
Learning Method

• Possible to sample the gradient, e.g.:

$$\mathbb{E}_{z \sim \text{gen}(x)} \left[ \text{cost}(z, y) \frac{\partial \log P(z|x)}{\partial \theta_g} \right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \text{cost}(z_i, y_i) \frac{\partial \log P(z_i|x_i)}{\partial \theta_g}$$

where \(z_i\) are sampled rationales

• Stochastic gradient decent on sampled gradients
Learning as Policy Gradient Method

---

**Input State**

---

**Policy Function** $P(z \mid x)$

---

**Set of Actions** $z$

---

**Cost (Reward)**

---

**Generator** $\text{gen}(x)$

---

**Encoder** $\text{enc}(z)$

---

**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, unfortunately it gets worse from here...

---

A type of REINFORCE learning (Williams, 1992)
Experiments

Three real-world datasets and applications for evaluation:

- Predicting sentiment for product reviews
- Parsing medical pathology reports
- Finding similar posts on QA forum
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 ... first, **the aroma is kind of bubblegum-like and grainy.** next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter.

**Ratings**

- **Look:** 5 stars
- **Aroma:** 2 stars
Evaluation: Product Review

Set-up: ratings are fractional; treat the task as regression following (McAuley et al, 2012)

use recurrent networks for \textit{gen()} and \textit{enc()}

Metrics: precision:
percentage of selected words in correct sentences

\textbf{mean squared error} on sentiment prediction

Baselines: SVM classifier
attention-based RNN
Sentiment Prediction

Various runs by changing sparsity & coherency

Full text
Sentiment Prediction

rationales getting close performance to full text
Sentiment Prediction

advantage of neural models over linear classifiers still clear
Precision of Rationales

Examples and precisions of 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

poured into a snifter. produces a small coffee head that reduces quickly. black as night. pretty typical imp. roasted malts hit on the nose. a little sweet chocolate follows. big toasty character on the taste. in between i'm getting plenty of dark chocolate and some bitter espresso. it finishes with hop bitterness. nice smooth mouthfeel with perfect carbonation for the style. overall a nice stout i would love to have again, maybe with some age on it.

more examples available at https://github.com/taolei87/rcnn/tree/master/code/rationale
Precision of Rationales

proper modeling leads to better rationale.
Learning Curves

Learning curves of \textit{cost(z)} on dev and precision on test

find good rationales after epochs of exploration
# Evaluation: Parsing Pathology Report

**Dataset:** patients’ pathology reports from hospitals such as MGH

**Task:** check if a disease/symptom is positive in text binary classification for each category

**Statistics:** several thousand report for each category pathology report is long (>1000 words) but structured

**Model:** use CNNs for $\text{gen}()$ and $\text{enc}()$
Evaluation: Parsing Pathology Report

**Category:**

- **Accession Number:** <unk>
- **Report Status:** Final
- **Type:** Surgical Pathology
- **Pathology Report:**
  LEFT BREAST ULTRASOUND GUIDED CORE NEEDLE BIOPSY SE ... INVASIVE DUCTAL CARCINOMA poorly differentiated modified Bloom Richardson grade III III measuring at least 0 7cm in this limited specimen Central hyalinization is present within the tumor mass but no necrosis is noted No lymphovascular invasion is identified No in situ carcinoma is present Special studies were performed at an outside institution with the following results not reviewed ESTROGEN RECEPTOR NEGATIVE PROGESTERONE RECEPTOR NEGATIVE ...

- **IDC:**
  - **F-score:** 98%

- **LCIS:**
  - **F-score:** 97%

- **LVI:**
  - **F-score:** 84%
Evaluation: Question Retrieval

Dataset: question posts from *AskUbuntu* forum (dos Santos et al., 2015; Lei et al., 2016) question pairs annotated as similar by users

Task: optimize neural representations such that distance between similar questions is small

Rationales:

what is the easiest way to install all the media codec available for ubuntu ? i am having issues with multiple applications prompting me to install codecs before they can play my files . how do i install media codecs ?

please any one give the solution for this whenever i try to convert the rpm file to deb file i always get this problem error : <unk> : not an rpm package ( or package manifest ) error executing `` lang=c rpm -qp -- queryformat % { name } <unk> ' " : at <unk> line 489 thanks . converting rpm file to debian file
Conclusion

• We present a prototype framework for rationalizing model predictions, and evaluate it quantitatively and qualitatively on various applications

• A lot of interesting future work

**aggregation**

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...

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

poured into a snifter, produces a small coffee head that reduces quickly, black as night, pretty typical imp, roasted malts hit on the nose, a little sweet chocolate follows, big toasty character on the taste...

**vision**

• good looking
• heavy palate
• chocolate smell

**improve training**

(variance reduction)

... ...
Conclusion

• We present a prototype framework for rationalizing model predictions, and evaluate it quantitatively and qualitatively on various applications

• A lot of interesting future work

Thank you!

Code & data available at

https://github.com/taolei87/rcnn/tree/master/code/rationale
Past Approaches

• How model interpretability is achieved?
  - learning sparse, interpretable word vectors
    (Faruqui et al., 2015; Herbert and Vecchi, 2015)
  - learning interpretable approximations of neural networks
    (e.g. “if-then” rules, linear classifiers and decision trees)
    (Thrun, 1995; Craven and Shavlik, 1996; Ribeiro et al., 2016)
Past Approaches

• How rationales (evidence) are defined / obtained?

  ▸ improve text classifications using rationale annotations (e.g. when training data is limited)

  *(Zaidan et al., 2007; Marshall et al., 2015; Zhang et al., 2016)*
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...
Rationales for Product Review

“T’ve got my old 37” or so tube television on it and it’s holding up really well.”
Ali_NY | 1 reviewer made a similar statement

“And it looks darn good!!”
Andrea C. Pappalardo

“It’s the perfect height and length for my daughter’s room.”
Momster 1975 | 2 reviewers made a similar statement

“Rationales improve user experience”

Rationales improve user experience