Applied Natural Language Processing

Info 256

Lecture 17: Interpretability (Oct. 21, 2021)

David Bamman, UC Berkeley

Prediction

kaggle

Overview Data Code Discussion Leaderboard Rules **Join Competition** •••

In the money Gold Silver Bronze

#	△...	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	Toxic Crusaders			0.98856	171	4y
2	—	neongen & Computer sa...			0.98822	129	4y
3	▲ 3	Adversarial Autoenc...			0.98805	451	4y

Interpretability

- Lots of scenarios where you need to understand the decisions your model is making:
    - Is your classifier using the right information to make decisions? How robust and transferable is it to new data that does not look exactly like the training data?
    - Is your classifier using information not aligned with your ethical values?
    - You want to use your model to interrogate the differences between categories

Insight

What makes a haiku?

Whitecaps on the bay:
A broken signboard banging
In the April wind.

— Richard Wright

Long and So (2016), “Literary Pattern Recognition: Modernism between Close Reading and Machine Learning,” Critical Inquiry

Insight

What makes a haiku?

Three spirits came to me
And drew me apart
To where the olive boughs
Lay stripped upon the ground;
Pale carnage beneath bright mist.

— Ezra Pound

Long and So (2016), “Literary Pattern Recognition: Modernism between Close Reading and Machine Learning,” Critical Inquiry

Word	Label	Probability
sky = True	not-ha : haiku =	5.7 : 1.0
shall = True	not-ha : haiku =	5.0 : 1.0
sea = True	not-ha : haiku =	5.0 : 1.0
man = True	not-ha : haiku =	4.3 : 1.0
last = True	not-ha : haiku =	3.7 : 1.0
snow = True	haiku : not-ha =	3.7 : 1.0
earth = True	not-ha : haiku =	3.7 : 1.0
blue = True	not-ha : haiku =	3.7 : 1.0
pass = True	not-ha : haiku =	3.7 : 1.0
voice = True	haiku : not-ha =	3.7 : 1.0
white = True	not-ha : haiku =	3.0 : 1.0
house = True	haiku : not-ha =	3.0 : 1.0
child = True	not-ha : haiku =	3.0 : 1.0
give = True	not-ha : haiku =	3.0 : 1.0
lo = True	haiku : not-ha =	3.0 : 1.0
sun = True	not-ha : haiku =	3.0 : 1.0
life = True	not-ha : haiku =	2.3 : 1.0
full = True	haiku : not-ha =	2.3 : 1.0
things = True	haiku : not-ha =	2.3 : 1.0
morning = True	haiku : not-ha =	2.3 : 1.0

Long and So (2016), “Literary Pattern Recognition: Modernism between Close Reading and Machine Learning,” Critical Inquiry

Logistic regression

$$P(y = 1 \mid x, \beta) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{F} x_i \beta_i\right)}$$

-2.850	UNIGRAM_scientist
-2.798	UNIGRAM_earth
-2.127	UNIGRAM_alien
-1.919	UNIGRAM_mysterious
-1.897	UNIGRAM_dr.
-1.849	UNIGRAM_planet
-1.715	UNIGRAM_brain
-1.626	UNIGRAM_world
-1.570	UNIGRAM_robot
-1.565	UNIGRAM_space

2.808	UNIGRAM_love
1.826	UNIGRAM_wedding
1.783	UNIGRAM_relationship
1.620	UNIGRAM_her
1.589	UNIGRAM_money
1.486	UNIGRAM_she
1.457	UNIGRAM_men
1.437	UNIGRAM_marriage
1.437	UNIGRAM_college
1.416	UNIGRAM_marry

6.classification/Classification.ipynb

- Global explanations describe the behavior of an entire model.

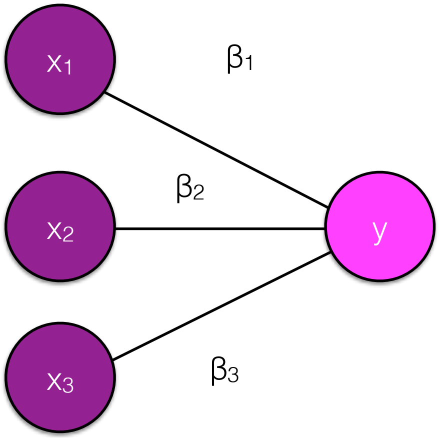
Local explanation

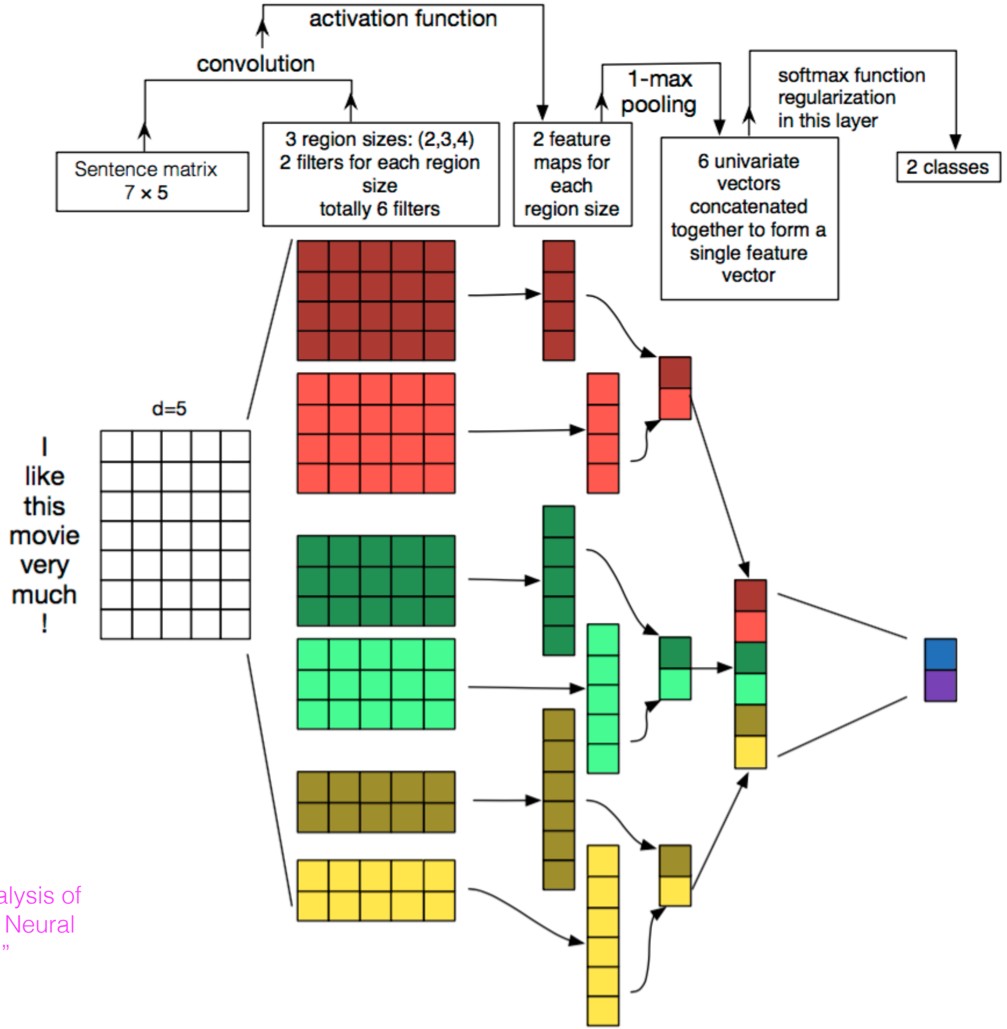
- Local explanations explain the classification decision for a single data point.
- What's the minimal set of features for a given data point that, if removed, would lead us to predict the opposite class? [Martens and Provost 2014]

"Dr. Strangelove, is a 1964 black comedy film that satirizes the Cold War fears of a nuclear conflict between the Soviet Union and the United States." → SCIENCE FICTION

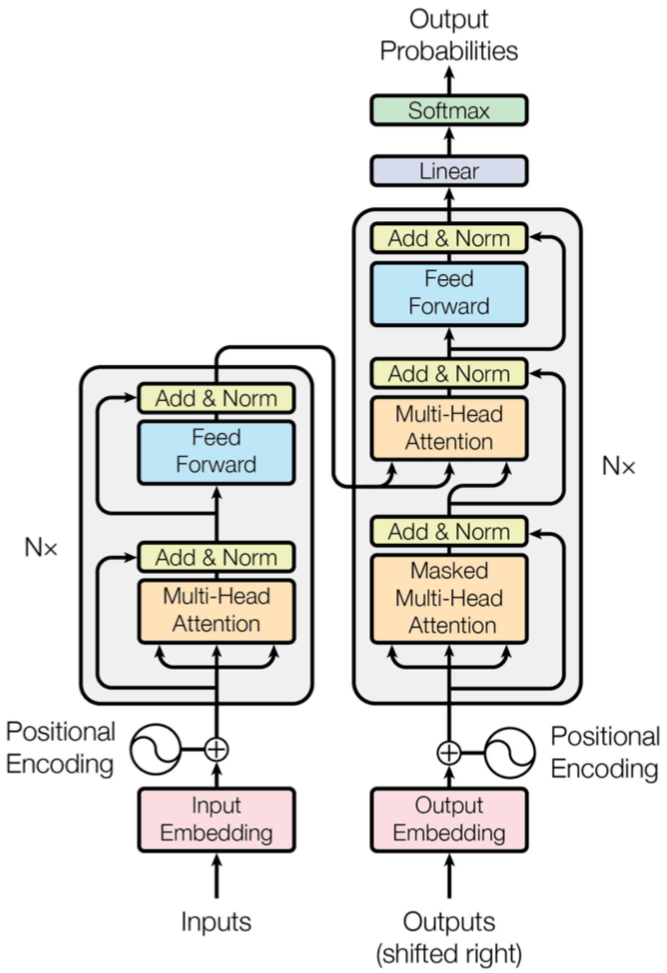
-2.850	UNIGRAM_scientist
-2.798	UNIGRAM_earth
-2.127	UNIGRAM_alien
-1.919	UNIGRAM_mysterious
-1.897	UNIGRAM_dr.
-1.849	UNIGRAM_planet
-1.715	UNIGRAM_brain
-1.626	UNIGRAM_world
-1.570	UNIGRAM_robot
-1.565	UNIGRAM_space
2.808	UNIGRAM_love
1.826	UNIGRAM_wedding
1.783	UNIGRAM_relationship
1.620	UNIGRAM_her
1.589	UNIGRAM_money
1.486	UNIGRAM_she
1.457	UNIGRAM_men
1.437	UNIGRAM_marriage
1.437	UNIGRAM_college
1.416	UNIGRAM_marry

Logistic regression





Zhang and Wallace 2016, “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification”



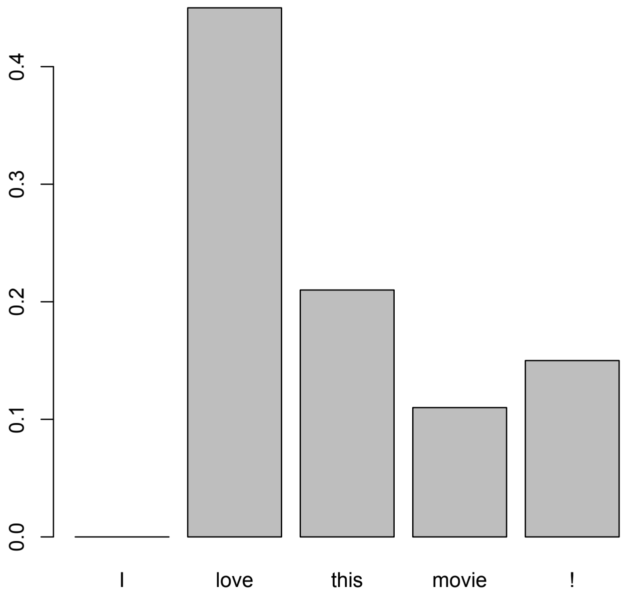
Vaswani et al. (2017), “Attention in All You Need”

Interpretability

- Intrinsic methods use information about the model to provide an interpretation (e.g., attention weights); post-hoc methods tend to be model-agnostic.

Intrinsic methods

- When used for explanation, attention is an intrinsic method — a *component* of the model itself is used to provide the explanation (here, the distribution of attention weights over the input).



Attention

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore	after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore
original α	adversarial $\tilde{\alpha}$
$f(x\|\alpha, \theta) = 0.01$	$f(x\|\tilde{\alpha}, \theta) = 0.01$

[Submitted on 26 Feb 2019 (v1), last revised 8 May 2019 (this version, v3)]

Attention is not Explanation

Sarthak Jain, Byron C. Wallace

Base model	brilliant	and	moving	performances	by	tom	and	peter	finch
Jain and Wallace (2019)	brilliant	and	moving	performances	by	tom	and	peter	finch
Our adversary	brilliant	and	moving	performances	by	tom	and	peter	finch

Figure 2: Attention maps for an IMDb instance (all predicted as positive with score $>$ 0.998), showing that in practice it is difficult to learn a distant adversary which is consistent on all instances in the training set.

[Submitted on 13 Aug 2019 (v1), last revised 5 Sep 2019 (this version, v2)]

Attention is not not Explanation

Sarah Wiegreffe, Yuval Pinter

Interpretability

- Plausibility: an explanation should be understandable by people and convincing to them.

- Fidelity (faithfulness): an explanation should reflect the underlying decision process a model made in making its prediction.

Jacovi and Goldberg (2020)

Post-hoc Interpretability

- Input features
- Adversarial examples
- Natural language explanations

Madsen et al. 2021

Input Features

- How important is a given token in the input for the prediction that's made?

Madsen et al. 2021

Gradient

x = inputs

$$\frac{d}{dx} f(x)_c$$

the output of the full model for class c

- The gradient in general measures how much the output of a function changes with respect to a change in the input → how important that input is for the final decision for a particular class.

Madsen et al. 2021

Gradient

Logistic regression

$$P(Y = y \mid X = x; \beta) = \frac{\exp\left(x^\top \beta_y\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(x^\top \beta_{y'}\right)}$$

Linear regression

$$y = x^\top \beta$$

$$\frac{\partial}{\partial x_i} x^\top \beta = \beta_i$$

This is the method of interpretability we've been using all along for linear models

Logistic regression

$$P(y = 1 \mid x, \beta) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{F} x_i \beta_i\right)}$$

-2.850	UNIGRAM_scientist
-2.798	UNIGRAM_earth
-2.127	UNIGRAM_alien
-1.919	UNIGRAM_mysterious
-1.897	UNIGRAM_dr.
-1.849	UNIGRAM_planet
-1.715	UNIGRAM_brain
-1.626	UNIGRAM_world
-1.570	UNIGRAM_robot
-1.565	UNIGRAM_space

2.808	UNIGRAM_love
1.826	UNIGRAM_wedding
1.783	UNIGRAM_relationship
1.620	UNIGRAM_her
1.589	UNIGRAM_money
1.486	UNIGRAM_she
1.457	UNIGRAM_men
1.437	UNIGRAM_marriage
1.437	UNIGRAM_college
1.416	UNIGRAM_marry

6.classification/Classification.ipynb

Integrated gradient

- The gradient method can violate “sensitivity” — that if x and x’ have different predictions and differ only in feature f, then f should be given high attribution — for neural components where gradients are flat (e.g., ReLU).

- The method of integrated gradients addresses this by additional introduction a baseline — another data point b that the feature importance of x is calculated with respect to — and integrating the gradients for all points along the path between x and b.

- For NLP, the baseline can just be a neutral data point — e.g., all [PAD] tokens.

Madsen et al. 2021

Adversarial examples

- Adversarial examples are data points that a classifier predicts incorrectly *and* that appear to be similar to data points a classifier predicts correctly.

A dark dystopian noir and Brad Pitt was terrific	→	positive
A dark dystopian noir and Brad Pritt was terrific	→	negative

- These examples help provide interpretability by surfacing the aspects of an input that would cause a prediction to be different if they were changed.

HotFlip

- One way of finding such adversarial examples is to find the inputs that would lead to the greatest change in the resulting loss — e.g., for a training data point <x, y=1>, a model that may predict 0.99 for original input x (so small loss); what token t can be we change from v to ~v in x to make it predict 0 (and so have high loss)?

$$\mathcal{L}\left(y, \tilde{x}_{t:v\rightarrow\tilde{v}}\right) - \mathcal{L}(y, x) \approx \frac{\partial \mathcal{L}(y, x)}{\partial x_{t,\tilde{v}}} - \frac{\partial \mathcal{L}(y, x)}{\partial x_{t,v}}$$

The difference in losses between the original input x and an altered one ~x

Is about equal to the difference in loss gradients with respect to each of those different inputs

Madsen et al. 2021; Ebrahimi et al. 2018

HotFlip

$$\text{HotFlip}(x) = \arg\max_{\tilde{x}_t: v \to \tilde{v}} \frac{\partial \mathcal{L}(y, x)}{\partial x_{t,\tilde{v}}} - \frac{\partial \mathcal{L}(y, x)}{\partial x_{t,v}}$$

- We can compute these gradients for every token in the input and select the ones that lead to the greatest change.

Madsen et al. 2021; Ebrahimi et al. 2018

HotFlip

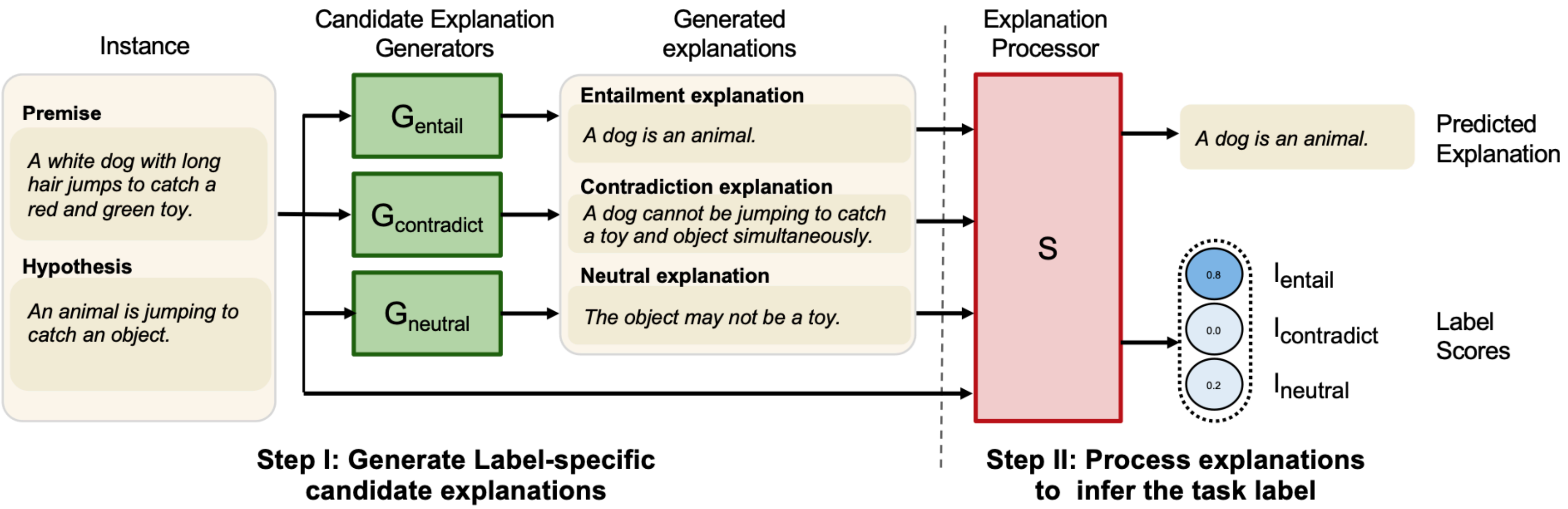
- Remember adversarial examples still need to be semantically similar to an original example.
- HotFlip contains the token swaps to be only among pairs of words that have a cosine similarity > 0.80.
- Semantically equivalent adversaries (Ribeiro et al. 2018) incorporate a paraphrase model to further satisfy this constraint.

Natural language explanations

A dark dystopian noir and the acting was terrific	→	positive

“People like good acting”

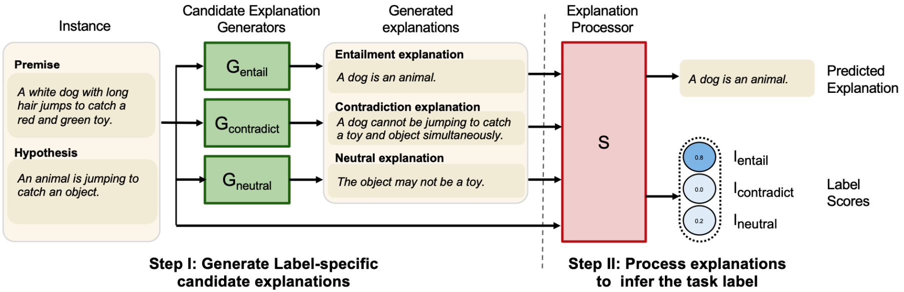
Intrinsic NL explanations



Kumar and Talukdar 2020

Intrinsic NL explanations

- Train a model to generate explanations for each possible class (in NLI: entail, contradict, neutral) on human-created explanations.



- Classifier inputs the text (hypothesis + premise) and the explanation in order to make a prediction about the class.

Kumar and Talukdar 2020

CAGE

- Solicit human-created explanations for answers in the Commonsense question answering dataset (CQA).
- Fine-tune GPT-2 on the question, answer, and explanation.
- Explanations do not necessarily need to be faithful to the model decision-making process.

Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

Rajani et al. 2019

Activity

`9.neural/Interpretability`

- Explore using integrated gradients to uncover what tokens in the input are most important for contributing to the model prediction.