CS388: Natural Language Processing

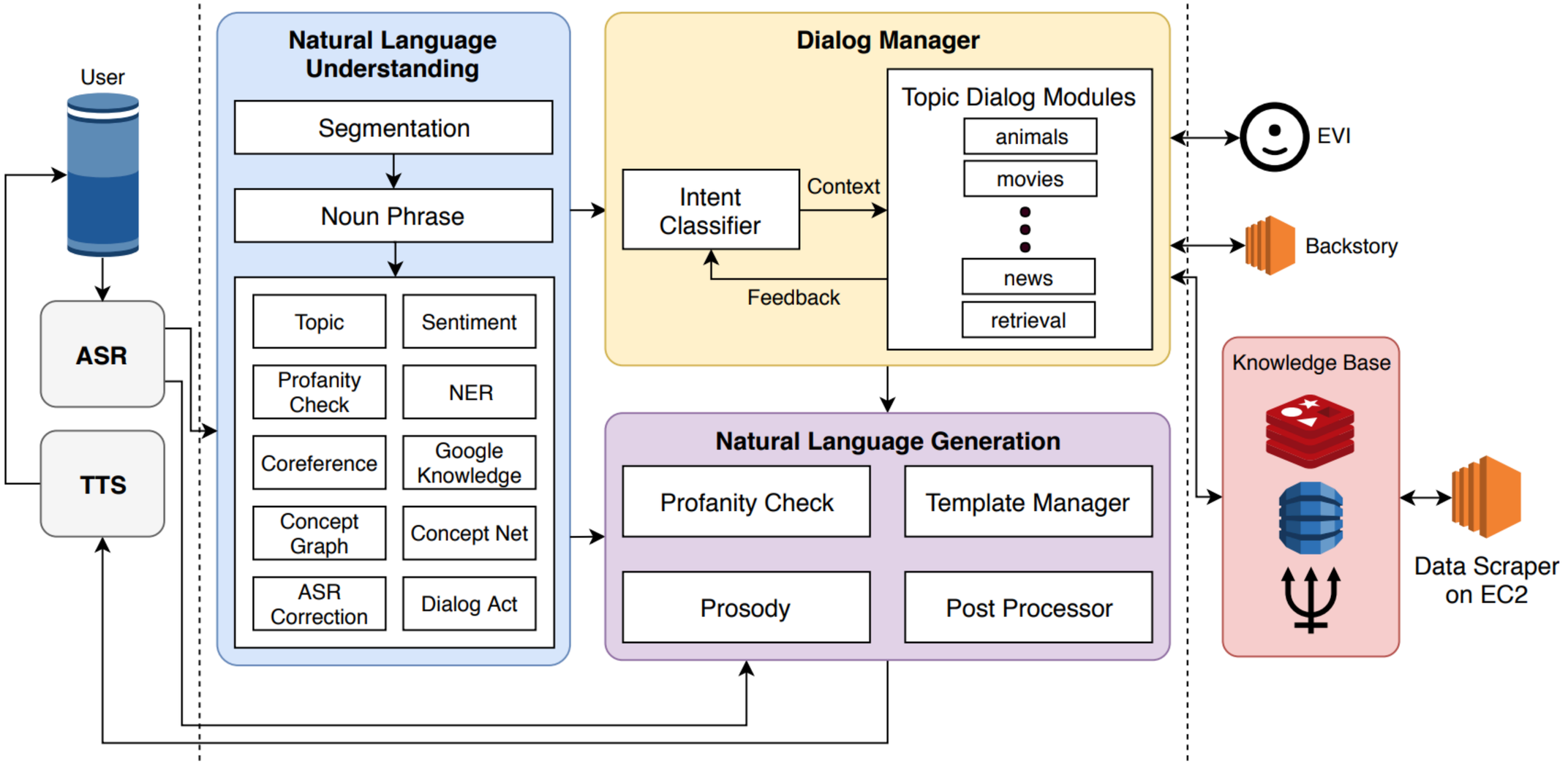
Lecture 21: Interpretation

Greg Durrett

Administrivia

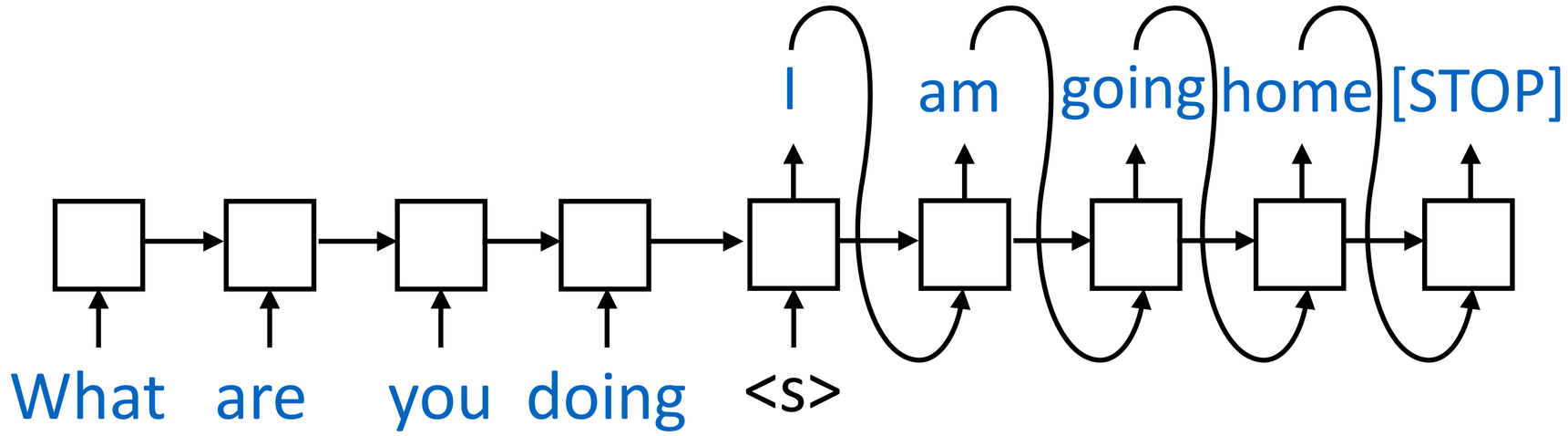
- Kenton Lee talk later today in Eunsol Choi's class
- P2 back soon

Recall: Dialogue Systems (Gunrock)



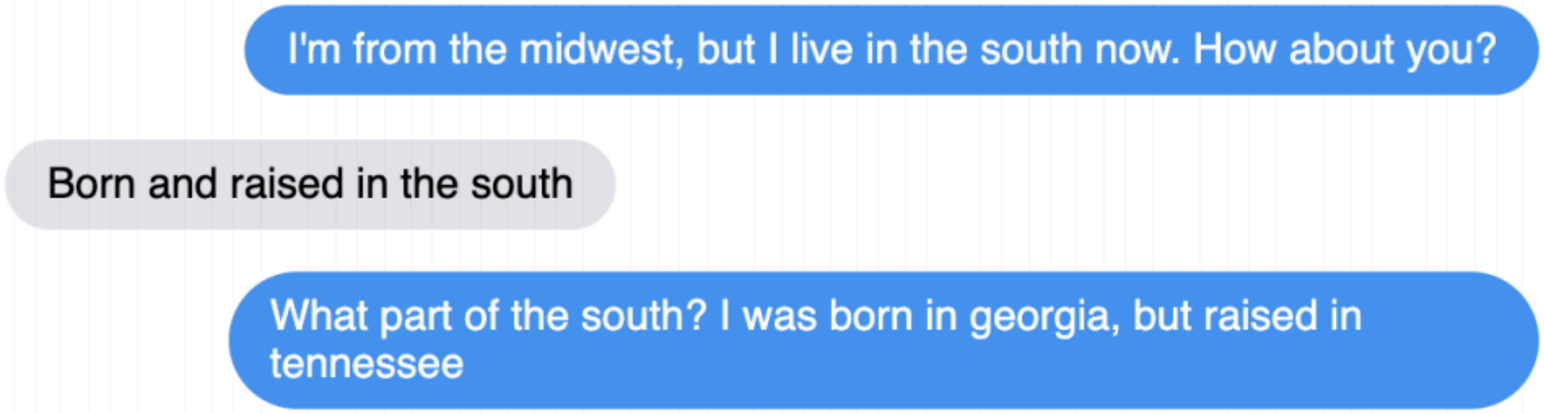
Chen … Zhou Yu (2018)

Recall: Dialogue Systems (Blender)



- Pre-train this model on >300GB of social media data (Meena/Blender)

- Fluent, on-topic generations, but system still gets into weird situations



Recall: Summarization Systems

Source Document (abbreviated)	BART Summary
The researchers examined three types of coral in reefs off the coast of Fiji ... The researchers found when fish were plentiful, they would eat algae and seaweed off the corals, which appeared to leave them more resistant to the bacterium Vibrio coralliilyticus, a bacterium associated with bleaching. The researchers suggested the algae, like warming temperatures, might render the corals' chemical defenses less effective, and the fish were protecting the coral by removing the algae.	Fisheries off the coast of Fiji are protecting coral reefs from the effects of global warming, according to a study in the journal Science.
Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."	Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.

- These look great! But they're not always factual

Lewis et al. (2019)

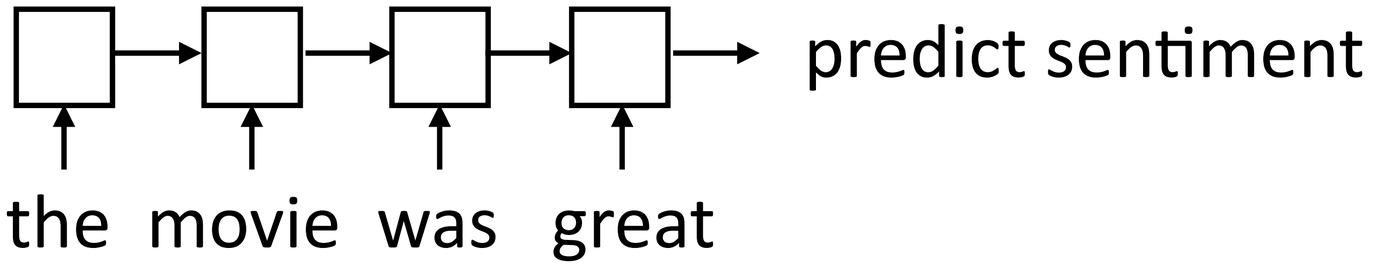
Today

- Interpreting neural networks: what does this mean and why should we care?
- Local explanations: erasure techniques
- Gradient-based methods
- Text-based explanations
- Evaluating explanations

Interpreting Neural Networks

Interpreting Neural Networks

- Neural models have complex behavior. How can we understand them?
- Sentiment w/LSTMs



- Looking at individual neurons usually doesn't tell us much
- Sentiment w/BERT: there are hundreds of attention computations… which ones actually mean something?

Interpreting Neural Networks

- Neural models have complex behavior. How can we understand them?
- Sentiment w/DANs:

	DAN	Ground Truth
this movie was not good	negative	negative
this movie was good	positive	positive
this movie was bad	negative	negative
the movie was not bad	negative	positive

  - Left side: predictions the model makes on individual words
  - Tells us how these words combine
- **How do we know why a neural network model made the prediction it made?**

Iyyer et al. (2015)

Why explanations?

- **Trust:** if we see that models are behaving in human-like ways and making human-like mistakes, we might be more likely to trust them and deploy them
- **Causality:** if our classifier predicts class *y* because of input feature *x*, does that tell us that *x* causes *y*? Not necessarily, but it might be helpful to know
- **Informativeness:** more information may be useful (e.g., predicting a disease diagnosis isn't that useful without knowing more about the patient's situation)
- **Fairness:** ensure that predictions are non-discriminatory

Lipton (2016)

Why explanations?

- Some models are naturally **transparent**: we can understand why they do what they do (e.g., a decision tree with <10 nodes)

- Explanations of more complex models

  - **Local explanations:** highlight what led to this classification decision. (Counterfactual: if these features were different, the model would've predicted a different class) — focus of this lecture

  - **Text explanations:** describe the model's behavior in language

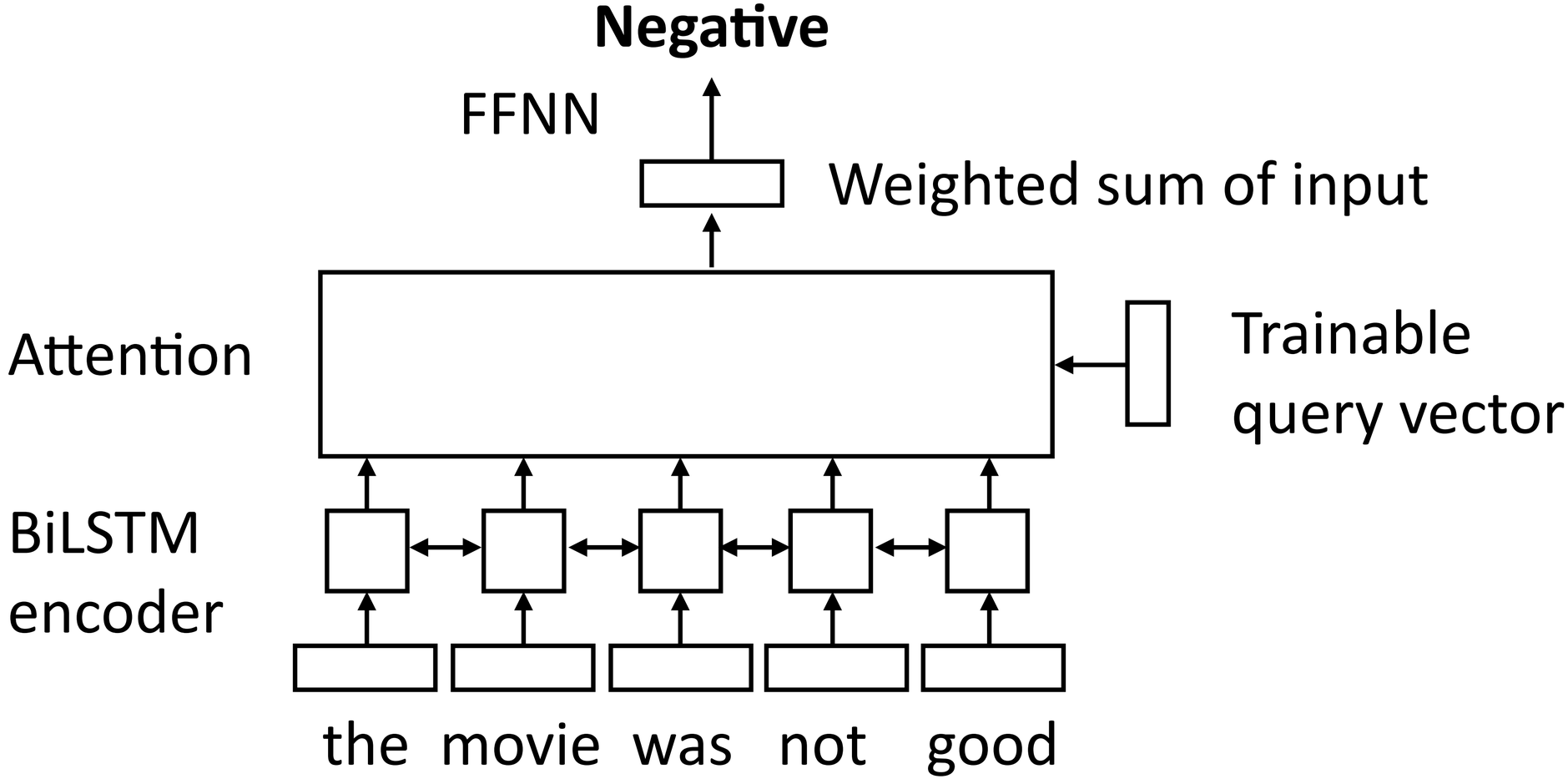
  - **Model probing:** auxiliary tasks, challenge sets, adversarial examples to understand more about how our model works

Lipton (2016); Belinkov and Glass (2018)

Local Explanations

(which parts of the input were responsible for the model's prediction on this particular data point?)

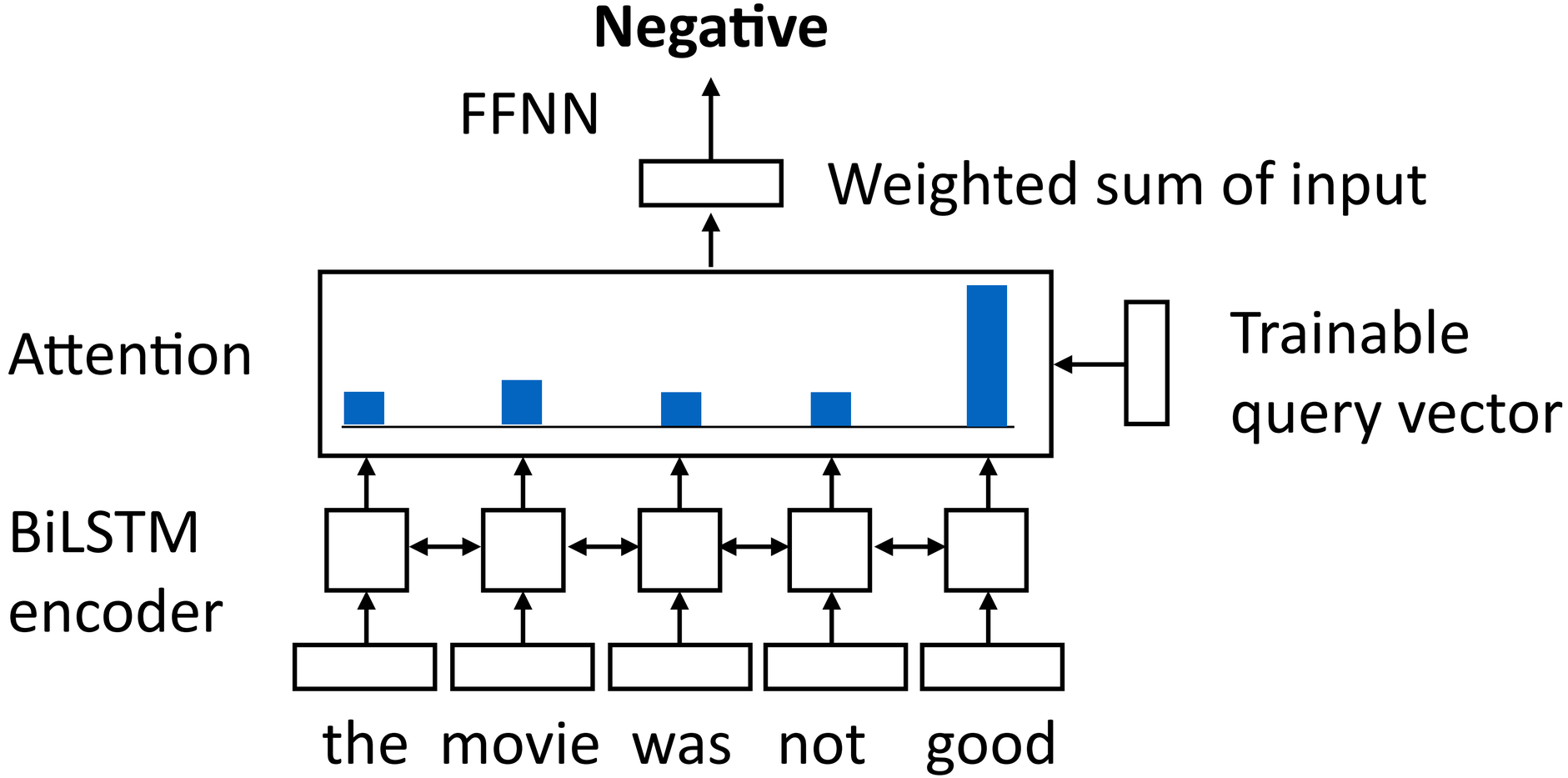
Sentiment Analysis with Attention



- Similar to a DAN model, but (1) extra BiLSTM layer; (2) attention layer instead of just a sum

Jain and Wallace (2019)

Attention Analysis



- Attention places most mass on *good* — did the model ignore *not*?
- What if we removed *not* from the input?

Jain and Wallace (2019)

Local Explanations

- An explanation could help us answer counterfactual questions: if the input were ***x'*** instead of ***x***, what would the output be?

	Model
that movie was not great , in fact it was terrible !	—
that movie was not ____ , in fact it was terrible !	—
that movie was ____ great , in fact it was ____ !	+

- Attention can't necessarily help us answer this!

Erasure Method

- Delete each word one by and one and see how prediction prob changes

that movie was not great , in fact it was terrible !	— prob = 0.97
___ movie was not great , in fact it was terrible !	— prob = 0.97
that ____ was not great , in fact it was terrible !	— prob = 0.98
that movie ____not great, in fact it was terrible !	— prob = 0.97
that movie was ___ great, in fact it was terrible !	— prob = 0.8
that movie was not ____, in fact it was terrible !	— prob = 0.99

Erasure Method

- Output: highlights of the input based on how strongly each word affects the output

  that movie was not great , in fact it was terrible !

  - *not* contributed to predicting the negative class (removing it made it less negative), great contributed to predicting the positive class (removing it made it more negative)

- Will this work well?
  - Inputs are now unnatural, model may behave in “weird” ways
  - Saturation: if there are two features that each contribute to negative predictions, removing each one individually may not do much

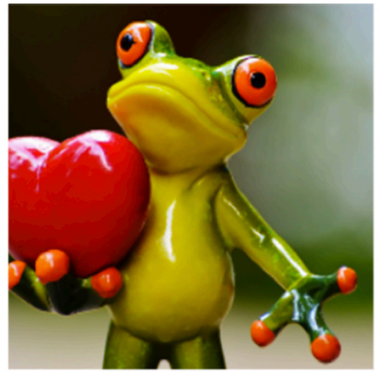
LIME

- Locally-interpretable, model-agnostic explanations (LIME)

- Similar to erasure method, but we're going to delete collections of things at once
  - Can lead to more realistic input (although people often just delete words with it)
  - More scalable to complex settings

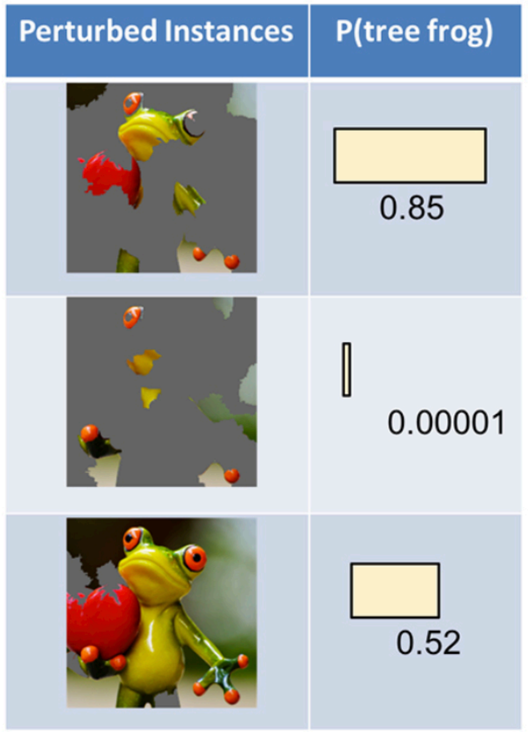
Ribeiro et al. (2016)

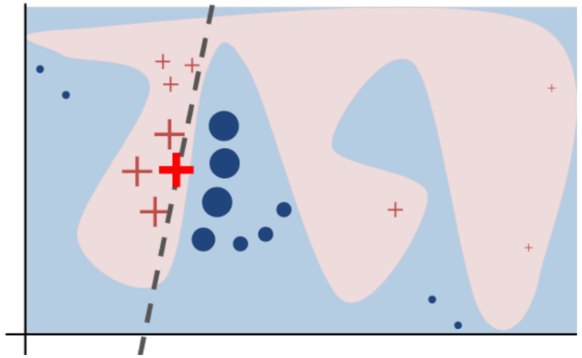
LIME



Original Image

Interpretable Components

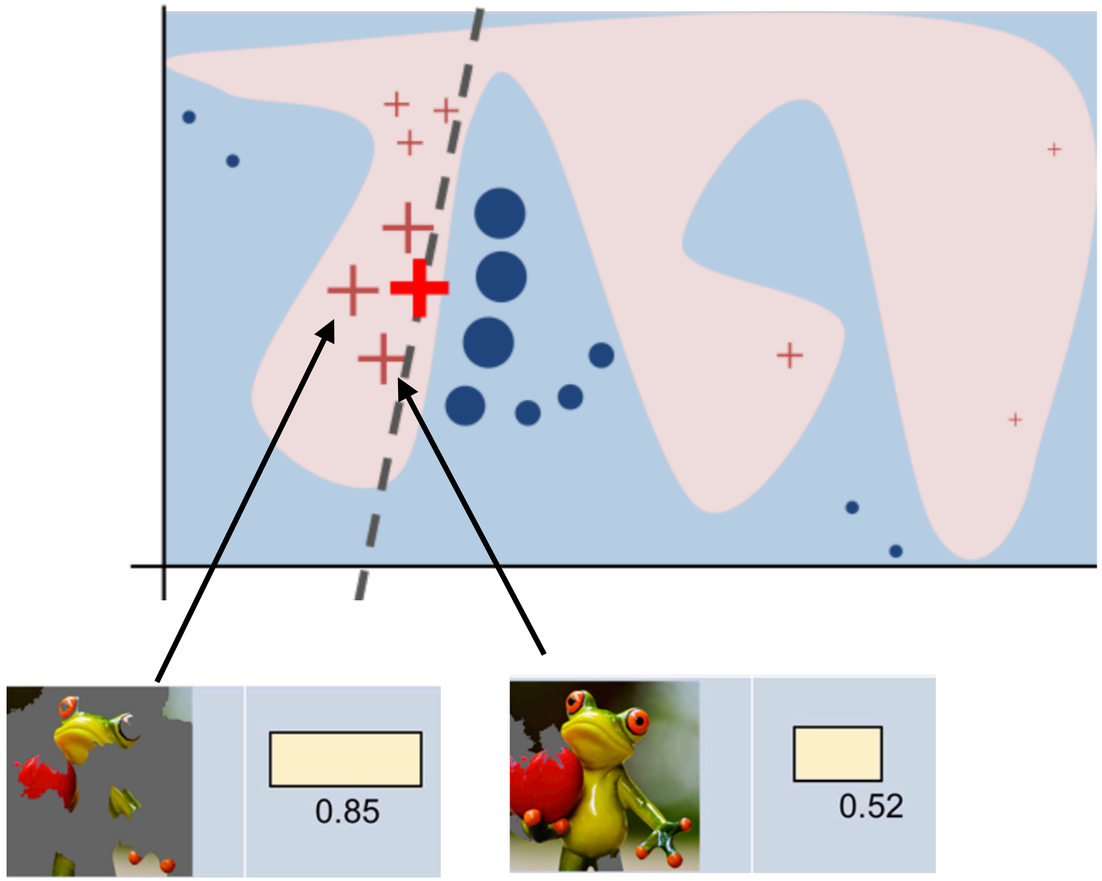




- Break input into components (for text: could use words, phrases, sentences, …)
- Check predictions on subsets of those
- Now we have model predictions on perturbed examples

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

LIME (cont'd)



- This is what the model is doing on perturbed examples of the input
- Now we train a classifier to predict **the model's behavior** based on **what subset of the input it sees**
- The weights of that classifier tell us which parts of the input are important

LIME (cont'd)

- This secondary classifier's **weights** now give us highlights on the input

The movie is mediocre, maybe even bad.	**Negative** 99.8%
The movie is mediocre, maybe even ~~bad~~.	**Negative** 98.0%
The movie is ~~mediocre~~, maybe even bad.	**Negative** 98.7%
The movie is ~~mediocre~~, maybe even ~~bad~~.	**Positive** 63.4%
The movie is ~~mediocre~~, ~~maybe~~ even ~~bad~~.	**Positive** 74.5%
The ~~movie~~ is mediocre, maybe even ~~bad~~.	**Negative** 97.9%

The movie is mediocre, maybe even bad.

Wallace, Gardner, Singh
Interpretability Tutorial at EMNLP 2020

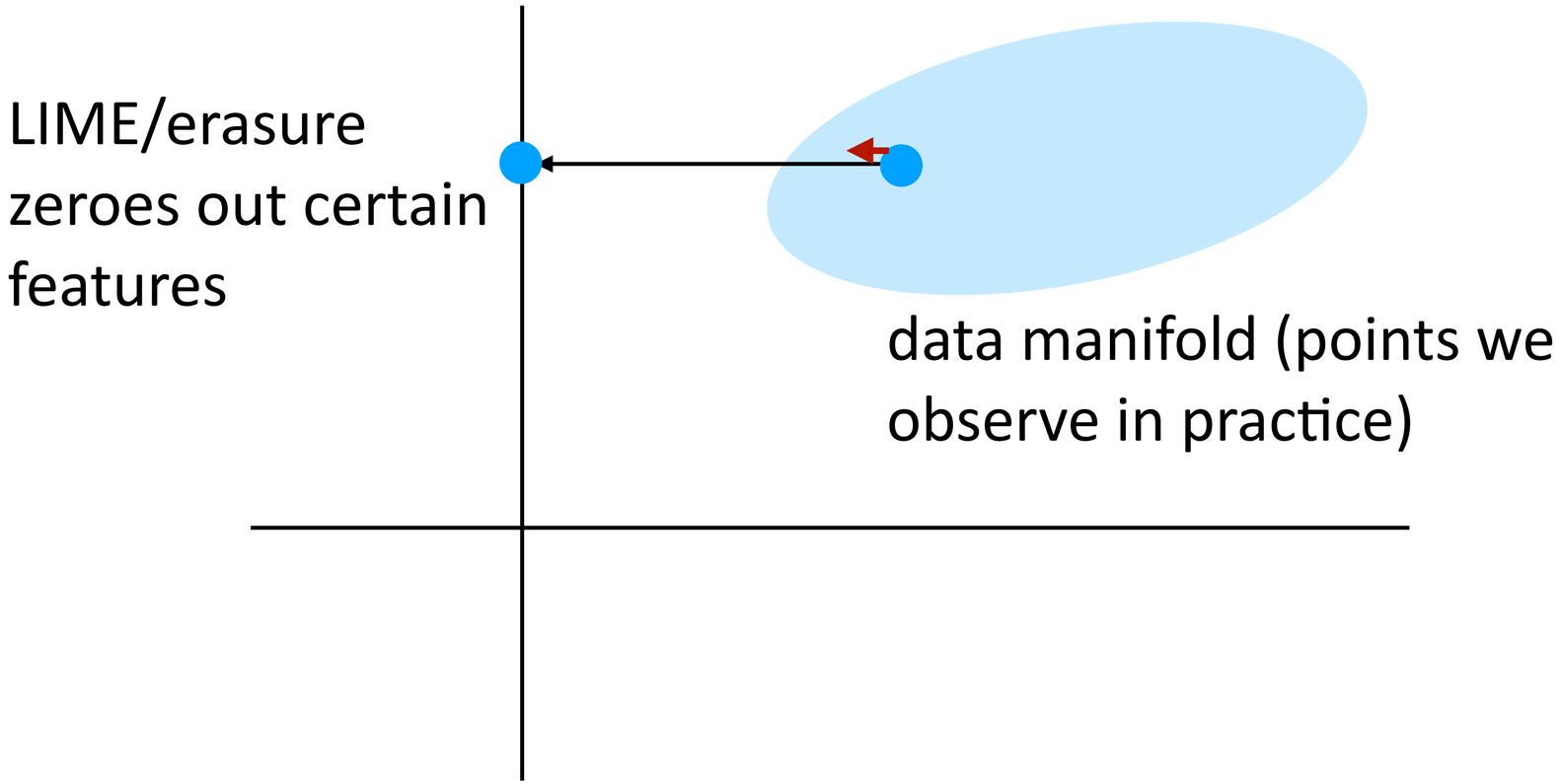
Problems with LIME

- Lots of moving parts here: what perturbations to use? what model to train? etc.
- Expensive to call the model all these times
- Linear assumption about interactions may not be reliable

Gradient-based Methods

Problems with LIME

- Problem: fully removing pieces of the input may cause it to be very unnatural



- Alternative approach: look at what this perturbation does locally right around the data point using gradients

Gradient-based Methods

score = weights * features
(or an NN)

Learning a model

Compute derivative of score with respect to weights: how can changing weights improve score of correct class?

Gradient-based Explanations

Compute derivative of score with respect to ***features***: how can changing ***features*** improve score of correct class?

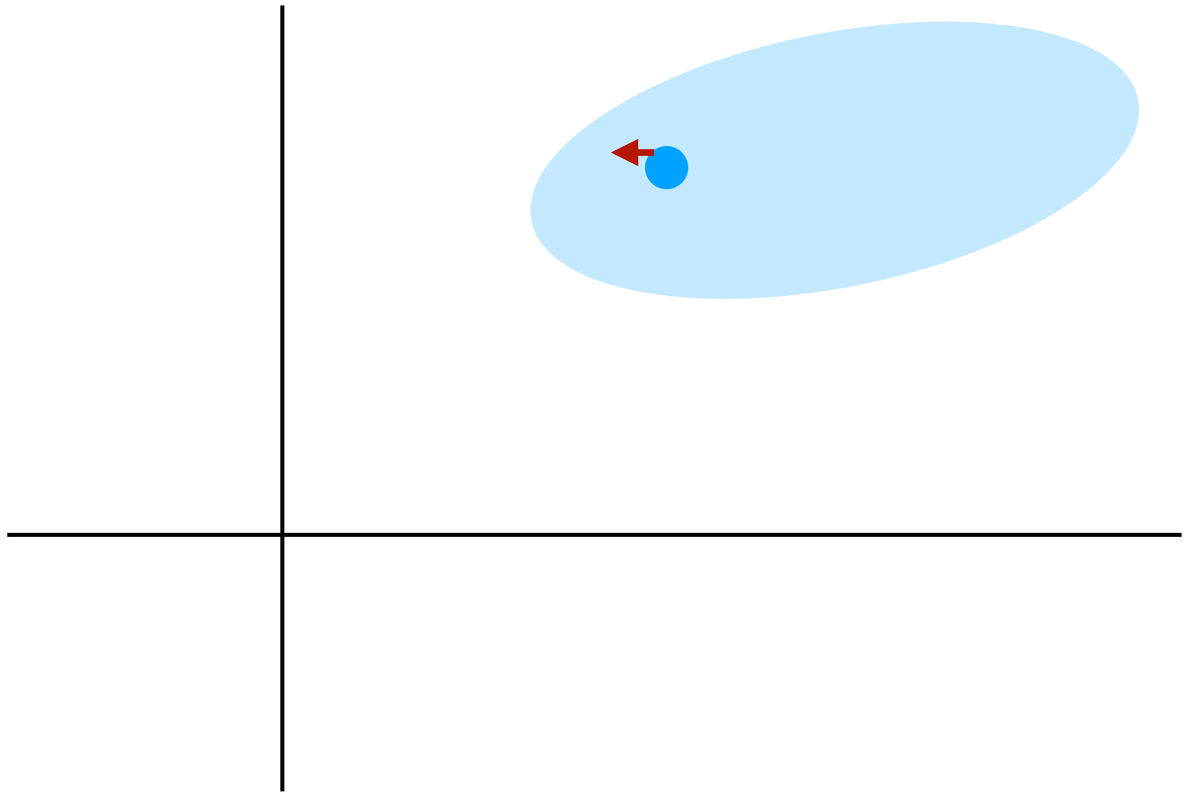
Gradient-based Methods

- Originally used for images

S_c = score of class c

I_0 = current image

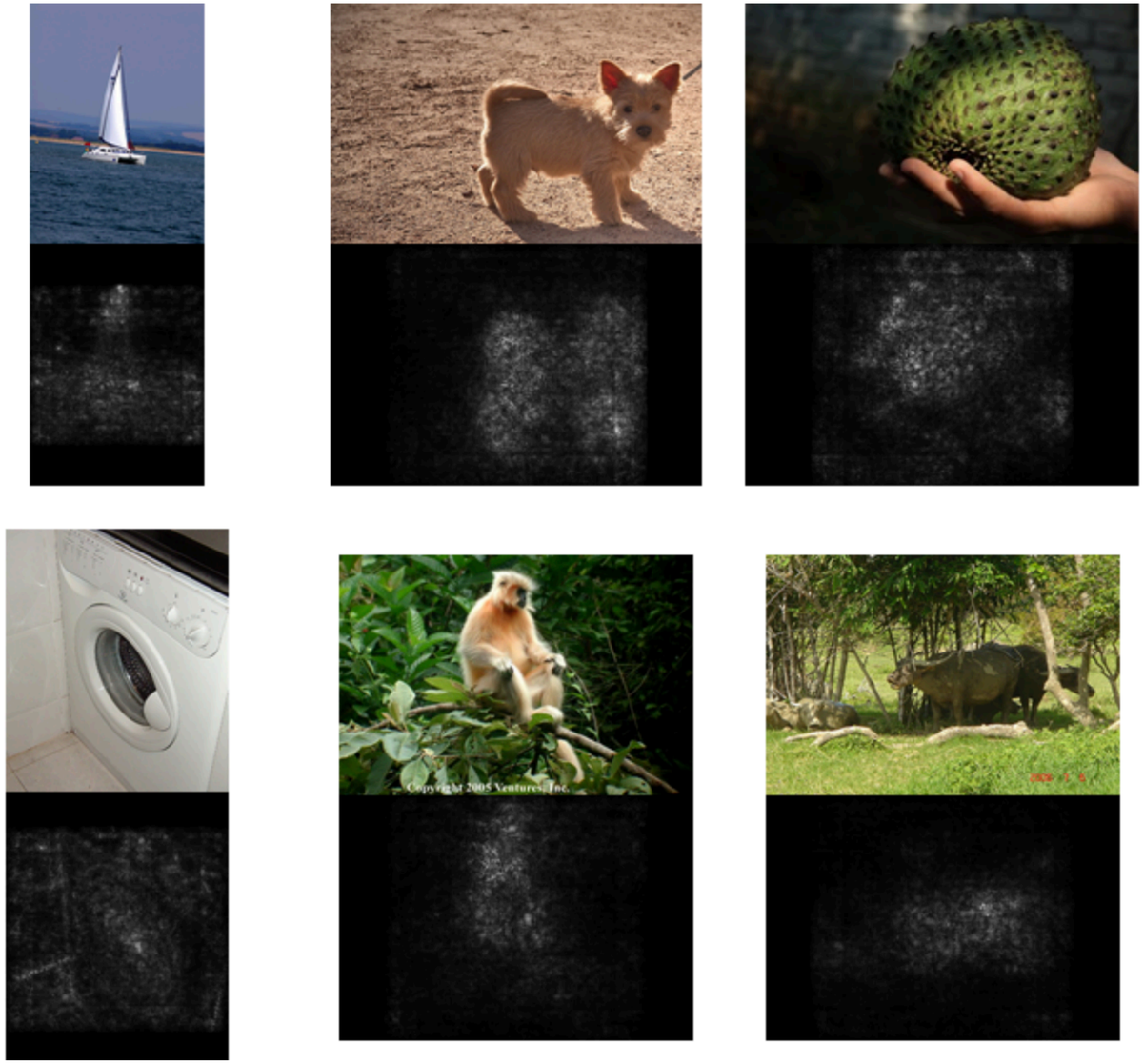
$$w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0}$$



- Higher gradient magnitude = small change in pixels leads to large change in prediction
- For words: “pixels” are coordinates of each word’s vector, sum these up to get the importance of that word

Simonyan et al. (2013)

Gradient-based Methods



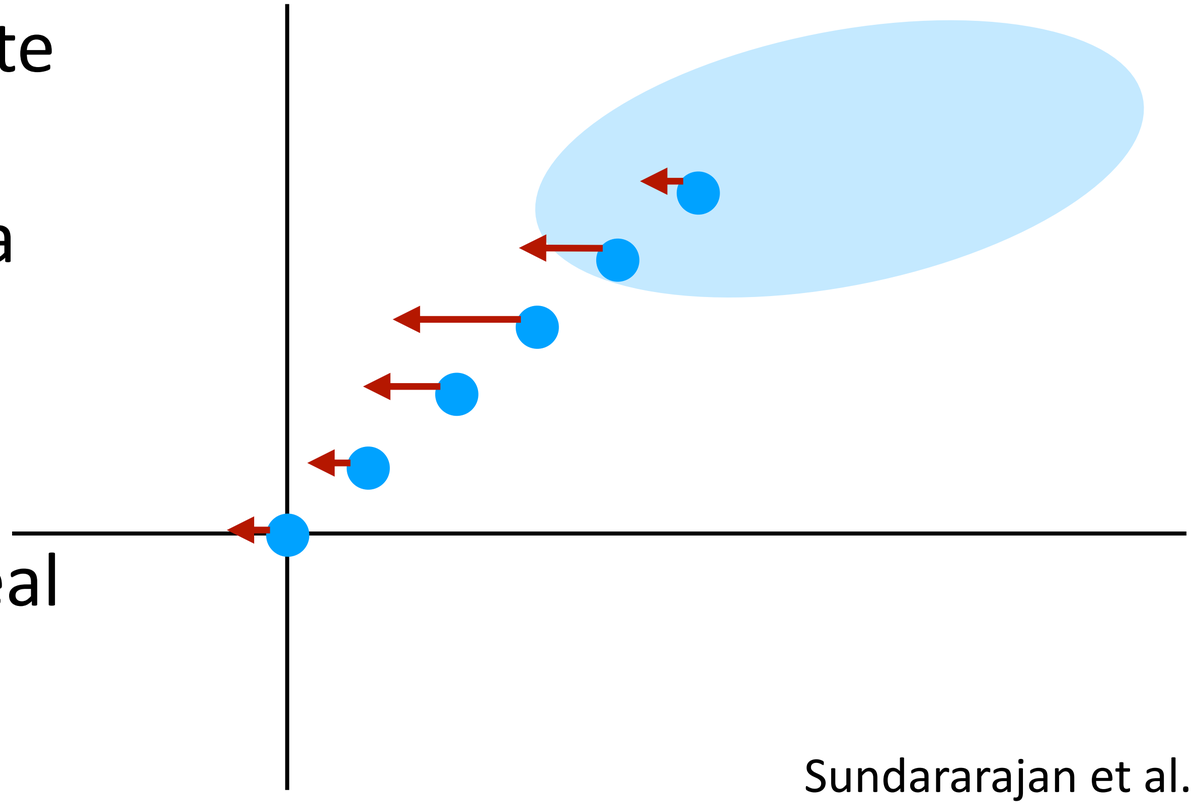
Simonyan et al. (2013)

Integrated Gradients

- Suppose you have prediction = A OR B for features A and B. Changing either feature doesn't change the prediction, but changing both would. Gradient-based method says neither is important

- Integrated gradients: compute gradients along a path from the origin to the current data point, aggregate these to learn feature importance

- Intermediate points can reveal new info about features



Sundararajan et al. (2017)

Integrated Gradients

$$\text{IntegratedGrads}_i^{approx}(x) ::= (x_i - x_i') \times \Sigma_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

Scale by total distance

Compute gradient at the *k*th point along the way w.r.t. the ith feature

Average over the m steps

x'_i = "baseline" — all PAD or MASK tokens (MASK usually works better

- Can be expensive: requires calling forward() and backward() at *m* steps along the way

Sundararajan et al. (2017)

Integrated Gradients

- Question type classification task:

how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Sundararajan et al. (2017)

Comparison

(Answer = Stanford University)

Question: Where did the Broncos practice for the Super Bowl ?
Passage: The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott . The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott .

(d) Erasure exact search optima.

Question: Where did the Broncos practice for the Super Bowl ?
Passage: The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott . The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott .

(a) Integrated Gradient (Sundararajan et al., 2017).

- Are these good explanations?

De Cao et al. (2020)

Text Explanations

Explanations of Bird Classification

Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

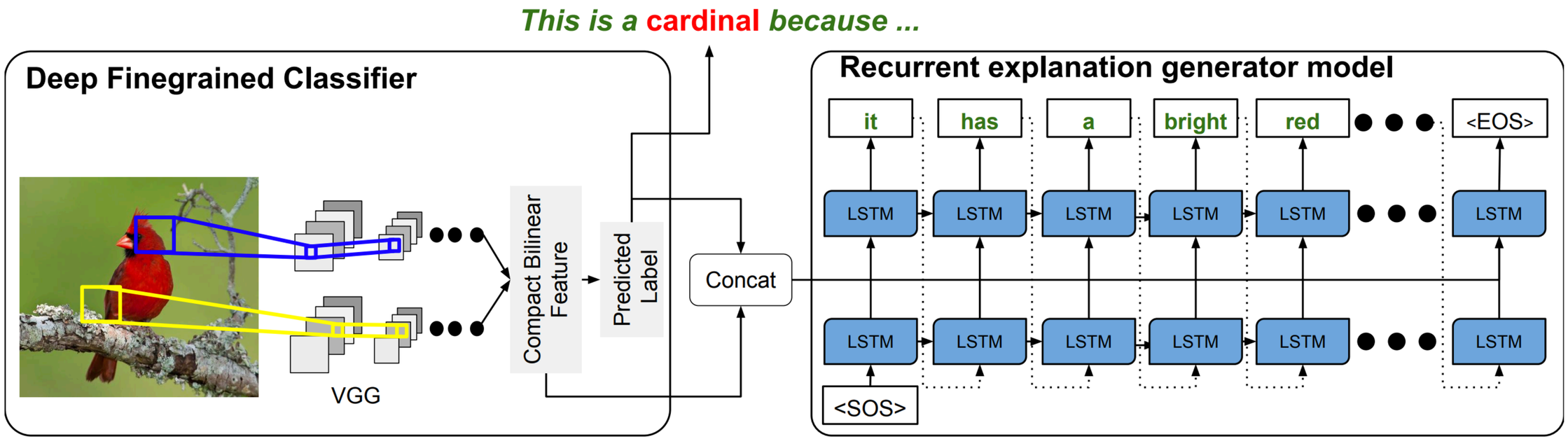
Laysan Albatross

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

- An explanation should be relevant to both the class and the image
- Are these features *really* what the model used?

Hendricks et al. (2016)

Explanations of Bird Classification



- Are these features *really* what the model used? The decoder looks at the image, but what it reports may not truly reflect the model's decision-making
- More likely to produce plausible (look good to humans) but unfaithful explanations!

Hendricks et al. (2016)

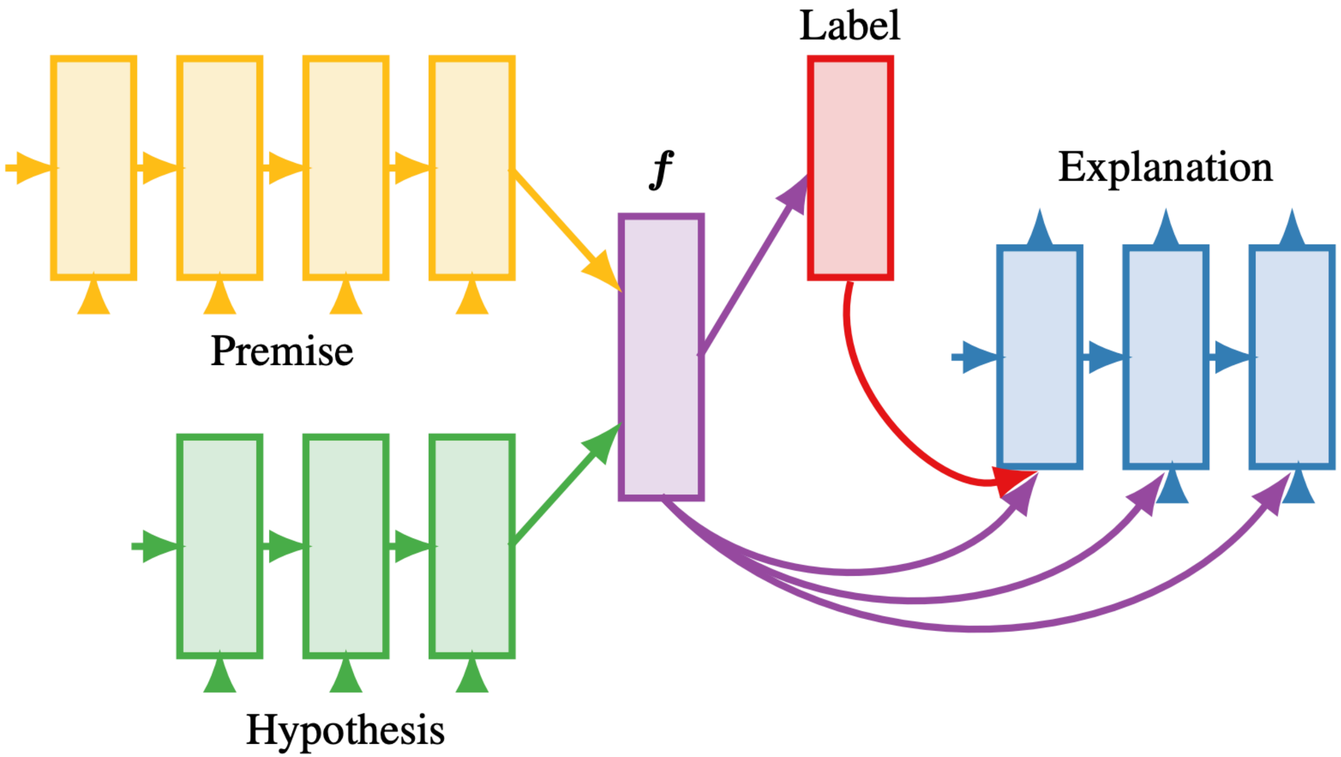
Explanations of NLI

Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

- How do we use this information? If we produce a network to predict it, does that make it an actual explanation of what's happening?

Camburu et al. (2019)

Explanations of NLI



- Information from *f* is fed into the explanation LSTM, but **no constraint that this must be used**. Different coordinates from *f* could predict label and explanations

Evaluating Explanations

Faithfulness vs. Plausibility

- Suppose our model is a bag-of-words model with the following:

  the = -1, movie = -1, good = +3, bad =0

  the movie was good prediction score=+1

  the movie was bad prediction score=-2

- Suppose explanation returned by LIME is:

  the movie was good

  the movie was bad

- Is this a “correct" explanation?

Faithfulness vs. Plausibility

- *Plausible* explanation: matches what a human would do

  the movie was good the movie was bad

  - Maybe useful to explain a task to a human, but it's not what the model is really doing!

- *Faithful* explanation: actually reflects the behavior of the model

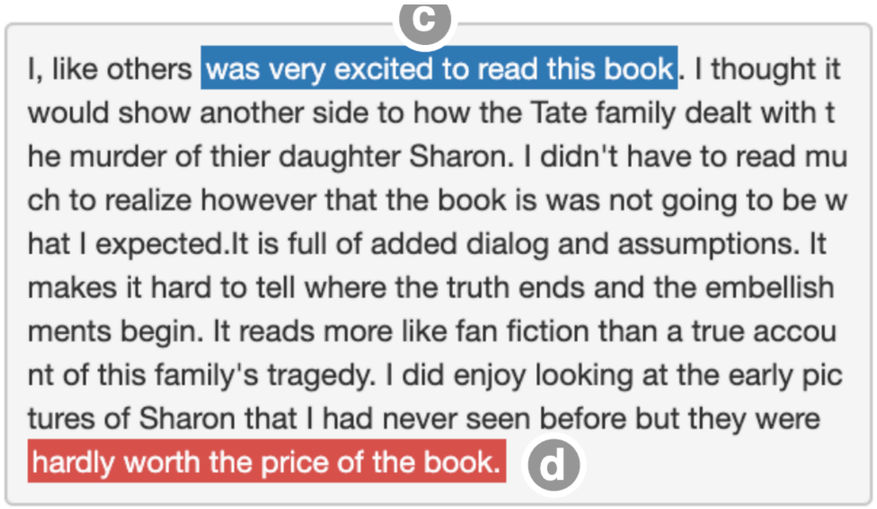
  the movie was good the movie was bad

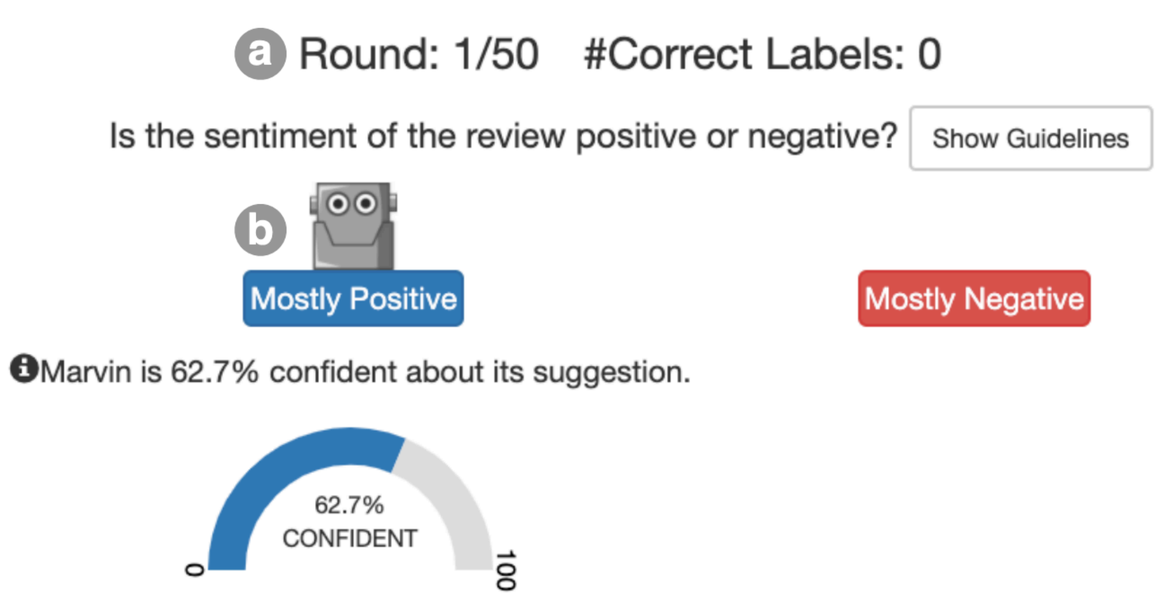
  - We usually prefer faithful explanations; non-faithful explanations are actually deceiving us about what our models are doing!
  - Rudin: *Stop Explaining Black Box Models for High-Stakes Decisions and Use Interpretable Models Instead*

Evaluating Explanations

- Nguyen (2018): delete words from the input and see how quickly the model flips its prediction?
  - Downside: not a “real” use case
- Hase and Bansal (2020): counterfactual simulatability: user should be able to predict what the model would do in another situation
  - Hard to evaluate

Evaluating Explanations





- Human is trying to label the sentiment. The AI provides its prediction to try to help. Does the human-AI team beat human/AI on their own?
- AI provides both an explanation for its prediction (blue) and also a possible counterargument (red)
- Do these explanations help the human? Slightly, but **AI is still better**
- No positive results on "human-AI teaming" with explanations Bansal et al. (2020)

Packages

- AllenNLP Interpret: https://allennlp.org/interpret
- Captum (Facebook): https://captum.ai/
- LIT (Google): https://ai.googleblog.com/2020/11/the-language-interpretability-tool-lit.html

Takeaways

- Many other ways to do explanation:
  - Probing tasks: we looked at these for ELMo, do vectors capture information about part-of-speech tags?
  - Diagnostic test sets (“unit tests” for models)
  - Building models that are explicitly interpretable (decision trees)
- Input attribution methods can be useful for visualization (consider using these for your final project!)