Plan for today

- Part I: Natural Language Inference
  - Definition and background
  - Datasets
  - Models
  - Problems (leading to Part II)

- Part II: Interpretable NLP
  - Motivation
  - Major approaches
  - Detailed methods

Part I: Natural Language Inference

Xiaochuang Han

with content borrowed from Sam Bowman and Xiaodan Zhu

What is natural language inference?

Example

- Text (T): *The Mona Lisa, painted by Leonardo da Vinci from 1503-1506, hangs in Paris' Louvre Museum.*
- Hypothesis (H): *The Mona Lisa is in France.*

Can we draw an appropriate inference from T to H?

What is natural language inference?

“We say that T entails H if, typically, a human reading T would infer that H is most likely true.”

- Dagan et al., 2005

What is natural language inference?

Example

- Text (T): *The Mona Lisa, painted by Leonardo da Vinci from 1503-1506, hangs in Paris' Louvre Museum.*
- Hypothesis (H): *The Mona Lisa is in France.*

Requires compositional sentence understanding:

(1) The Mona Lisa (not Leonardo da Vinci) hangs in …
(2) Paris' Louvre Museum is in France.

Other names

Terminologies below mean the same:

- Natural language inference (NLI)
- Recognizing textual entailment (RTE)
- Local textual inference

Format

- A short passage, usually just one sentence, of text (T) / premise (P)
- A sentence of hypothesis (H)
- A label indicating whether we can draw appropriate inferences
  - 2-way: entailment | non-entailment
  - 3-way: entailment | neutral | contradiction

Data

Recognizing Textual Entailment (**RTE**) **1-7**

- Seven annual competitions (First PASCAL, then NIST)
- Some variation in format (2-way / 3-way), but about 5000 NLI-format examples total
- Premises (texts) drawn from naturally occurring text, often long or complex
- Expert-constructed hypotheses

P: *Cavern Club sessions paid the Beatles £15 evenings and £5 lunchtime.*
H: *The Beatles perform at Cavern Club at lunchtime.*
Label: entailment

Dagan et al., 2006 et seq.

Data

The Stanford NLI Corpus (**SNLI**)

- Premises derived from image captions (Flickr 30k), hypotheses created by crowdworkers
- About 550,000 examples; first NLI corpus to see encouraging results with neural networks

P: *A black race car starts up in front of a crowd of people.*
H: *A man is driving down a lonely road.*
Label: contradiction

Bowman et al., 2015

Data

Multi-genre NLI (**MNLI**)

- Multi-genre follow-up to SNLI: Premises come from ten different sources of written and spoken language, hypotheses written by crowdworkers
- About 400,000 examples

P: *yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual*
H: *August is a black out month for vacations in the company.*
Label: contradiction

Williams et al., 2018

Data

Crosslingual NLI (**XNLI**)

- A new development and test set for MNLI, translated into 15 languages
- About 7,500 examples per language
- Meant to evaluate cross-lingual transfer: Train on English MNLI, evaluate on another target languages

P: 让我告诉你，美国人最终如何看待你作为独立顾问的表现。
H: 美国人完全不知道您是独立律师。
Label: contradiction

Conneau et al., 2018

Data

SciTail

- Created by pairing statements from science tests with information from the web
- First NLI set built entirely on existing text
- About 27,000 pairs

P: *Cut plant stems and insert stem into tubing while stem is submerged in a pan of water.*
H: <u>Stems transport water to other parts of the plant through a system of tubes.</u>
Label: neutral

Khot et al., 2018

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.

entailment

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.

neutral

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

contradiction

Connections with other tasks

Question Answering: Given a question (premise), identify a text that entails an answer (hypothesis).

Information Retrieval: Given a query (hypothesis), identify texts that entail that query (premises).

Summarization: Given a text (premise) *T*, create or identify a text that *T* entails.

Summarization: Omit sentences that are entailed by others.

Machine translation: Mutual entailment between texts in different languages.

Bill MacCartney, Stanford CS224U slides

Some early methods

Some earlier NLI work involved learning with shallow features:

- Bag of words features on hypothesis
- Bag of word-pairs features to capture alignment
- Tree kernels
- Overlap measures like BLEU

These methods work surprisingly well, but not competitive on current benchmarks.

MacCartney, 2009; Stern and Dagan, 2012; Bowman et al. 2015

Some early methods

Much non-ML work on NLI involves **natural logic**:

- A formal logic for deriving entailments between sentences.
- Operates directly on parsed sentences (natural language), no explicit logical forms.
- Generally sound but far from complete — only supports inferences between sentences with clear structural parallels.
- Most NLI datasets aren't strict logical entailment, and require some unstated premises — this is hard.

Lakoff, 1970; Sánchez Valencia, 1991; MacCartney, 2009; Icard III and Moss, 2014; Hu et al., 2019

A bit more into natural logic

Monotonicity

- <u>Upward monotone</u>: preserve entailments from **subsets** to **supersets**.

- <u>Downward monotone</u>: preserve entailments from **supersets** to **subsets**.

- <u>Non-monotone</u>: do not preserve entailment in either direction.

Bill MacCartney, Stanford CS224U slides

A bit more into natural logic

Upward monotonicity in language

- Upward monotonicity is sort of the default for lexical items
- Most determiners (e.g., a, some, at least, more than)
- The **second** argument of every (e.g., every turtle **danced**)

Bill MacCartney, Stanford CS224U slides

A bit more into natural logic

Downward monotonicity in language

- Negations (e.g., not, n’t, never, no, nothing, neither)
- The **first** argument of every (e.g., every **turtle** danced)
- Conditional antecedents (if-clauses)

A bit more into natural logic

Edits that help preserve forward entailment:

- Deleting modifiers
- Changing specific terms to more general ones
- Dropping conjuncts, adding disjuncts

Edits that do not help preserve forward entailment:

- Adding modifiers
- Changing general terms to specific ones
- Adding conjuncts, dropping disjuncts

In downward monotone environments, the above are reversed.

Bill MacCartney, Stanford CS224U slides

A bit more into natural logic

Q: Which of the below contexts are **upward monotone**?

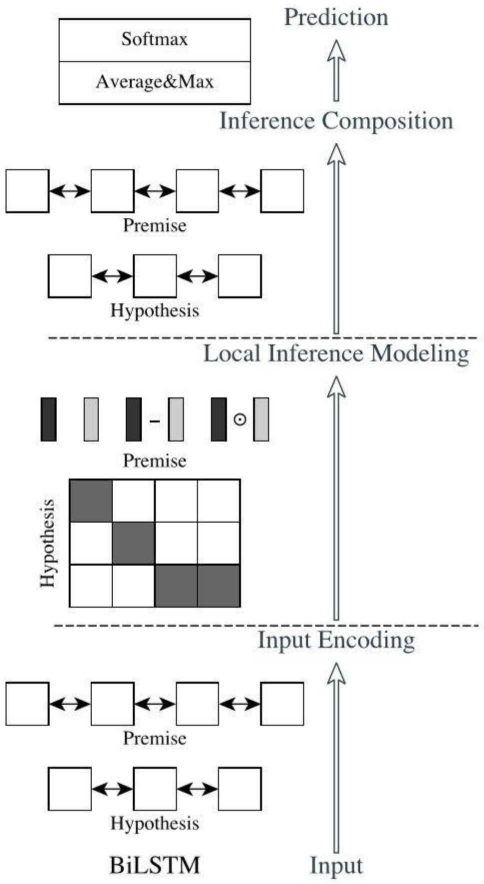
1. Some **<u>dogs</u>** are cute
2. Most **<u>cats</u>** meow
3. Some parrots **<u>talk</u>**

More recent methods

Deep learning models for NLI

- Baseline model with typical components
  - ESIM (Chen et al., 2017)
- Enhance with syntactic structures
  - HIM (Chen et al., 2017)
- Leverage unsupervised pretraining
  - BERT (Devlin et al., 2018)
- Enhance with semantic roles
  - SJRC (Zhang et al., 2019)

Enhanced Sequential Inference Models (ESIM)



Layer 3: Inference Composition/Aggregation

Perform composition/aggregation over local inference output to make the global judgement.

Layer 2: Local Inference Modeling

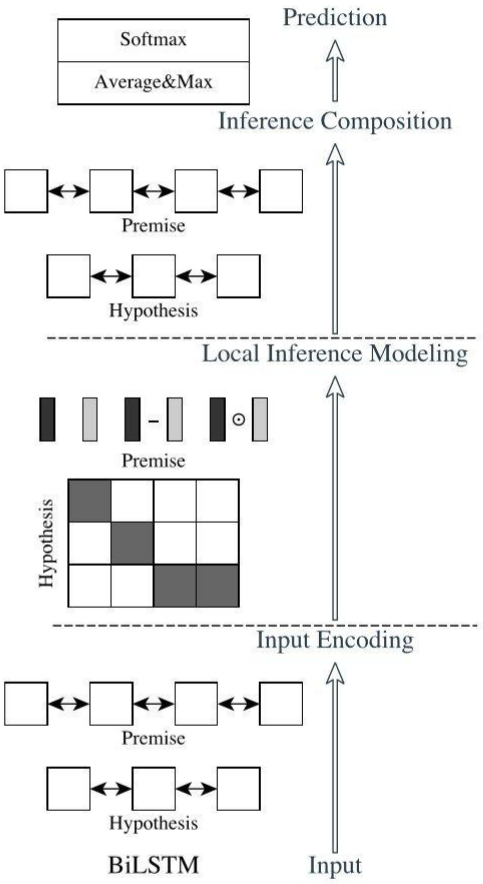
Collect information to perform “local” inference between words or phrases. (Some heuristics works well in this layer.)

Layer 1: Input Encoding

ESIM uses BiLSTM, but different architectures can be used here, e.g., transformer-based, ELMo, densely connected CNN, tree-based models, etc.

Chen et al., 2017

Enhanced Sequential Inference Models (ESIM)



Layer 3: Inference Composition/Aggregation

Perform composition/aggregation over local inference output to make the global judgement.

Layer 2: Local Inference Modeling

Collect information to perform "local" inference between words or phrases. (Some heuristics works well in this layer.)

Layer 1: Input Encoding

ESIM uses BiLSTM, but different architectures can be used here, e.g., transformer-based, ELMo, densely connected CNN, tree-based models, etc.

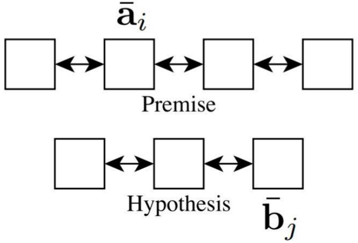
[Chen et al., 2017]

Encoding premise and hypothesis

- For a premise sentence **a** and a hypothesis sentence **b**:

$$\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_{\ell_a})$$
$$\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_{\ell_b})$$

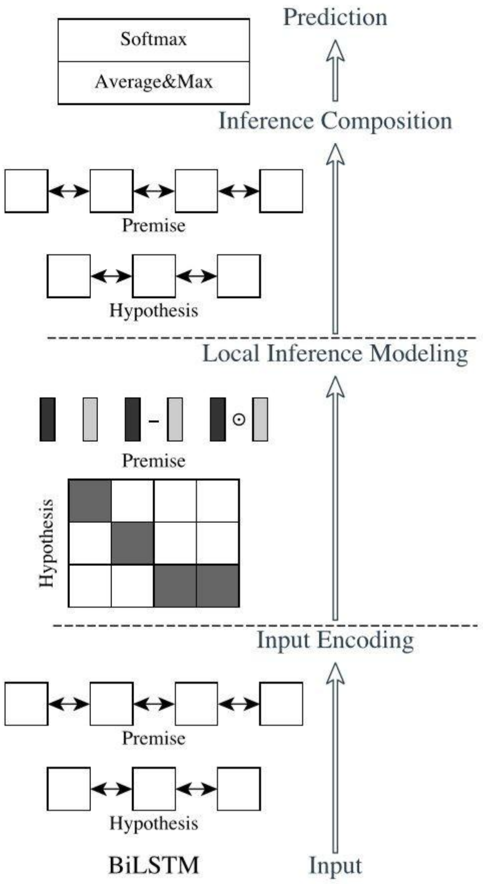
we can apply different encoders (e.g., here BiLSTM)**:**



$$\bar{\mathbf{a}}_i = \{\mathrm{BiLSTM}(\mathbf{a})\}_i, i \in (1, \ldots, \ell_a)$$
$$\bar{\mathbf{b}}_j = \{\mathrm{BiLSTM}(\mathbf{b})\}_j, j \in (1, \ldots, \ell_b)$$

where ā_i denotes the output vector of BiLSTM at the position i of premise, which encodes word a_i and its context.

Enhanced Sequential Inference Models (ESIM)



Layer 3: Inference Composition/Aggregation

Perform composition/aggregation over local inference output to make the global judgement.

Layer 2: Local Inference Modeling

Collect information to perform "local" inference between words or phrases. (Some heuristics works well in this layer.)

Layer 1: Input Encoding

ESIM uses BiLSTM, but different architectures can be used here, e.g., transformer-based, ELMo, densely connected CNN, tree-based models, etc.

[Chen et al., 2017](#)

Local inference modeling

Premise *Two **dogs** are running through a field*

Hypothesis *There are **animals** outdoors*

Attention content

$$\tilde{\mathbf{a}}(\text{“dogs”})$$
$$=0.05 \times \text{“There”} + 0.05 \times \text{“are”}$$
$$+\,0.8 \times \text{“animals”} + 0.1 \times \text{“outdoors”}$$

Attention Weights

Local inference modeling

- The (cross-sentence) attention content is computed along both the premise-to-hypothesis and hypothesis-to-premise direction.

$$\tilde{\mathbf{a}}_i = \sum_{j=1}^{\ell_b} \frac{\exp(e_{ij})}{\sum_{k=1}^{\ell_b} \exp(e_{ik})} \bar{\mathbf{b}}_j$$

$$\tilde{\mathbf{b}}_j = \sum_{i=1}^{\ell_a} \frac{\exp(e_{ij})}{\sum_{k=1}^{\ell_a} \exp(e_{kj})} \bar{\mathbf{a}}_i$$

where,

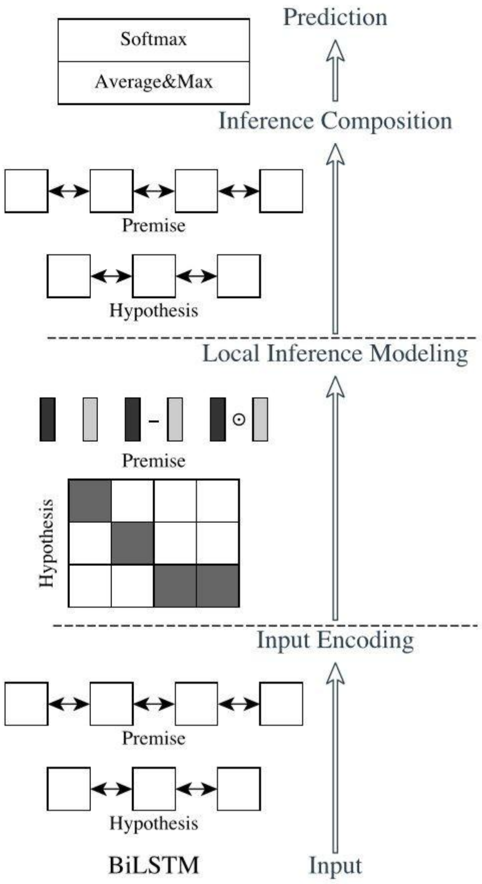
$$e_{ij} = \bar{\mathbf{a}}_i^T \bar{\mathbf{b}}_j$$

Local inference modeling

- With soft alignment ready, we can collect local inference information.
- Note that in various NLI models, the following heuristics have shown to work very well:

$$\mathbf{m_a} = [\bar{\mathbf{a}}; \tilde{\mathbf{a}}; \bar{\mathbf{a}} - \tilde{\mathbf{a}}; \bar{\mathbf{a}} \odot \tilde{\mathbf{a}}]$$
$$\mathbf{m}_b = [\bar{\mathbf{b}}; \tilde{\mathbf{b}}; \bar{\mathbf{b}} - \tilde{\mathbf{b}}; \bar{\mathbf{b}} \odot \tilde{\mathbf{b}}]$$

Enhanced Sequential Inference Models (ESIM)



Layer 3: Inference Composition/Aggregation

Perform composition/aggregation over local inference output to make the global judgement.

Layer 2: Local Inference Modeling

Collect information to perform "local" inference between words or phrases. (Some heuristics works well in this layer.)

Layer 1: Input Encoding

ESIM uses BiLSTM, but different architectures can be used here, e.g., transformer-based, ELMo, densely connected CNN, tree-based models, etc.

[Chen et al., 2017]

Inference composition / aggregation

- The next component is to perform composition/aggregation over local inference knowledge collected above.
- BiLSTM can be used here to perform “composition” over local inference:

$$\mathbf{v_a} = \mathrm{BiLSTM}(\mathbf{m_a})$$

$$\mathbf{v_b} = \mathrm{BiLSTM}(\mathbf{m_b})$$

where

$$\mathbf{m_a} = [\bar{\mathbf{a}}; \tilde{\mathbf{a}}; \bar{\mathbf{a}} - \tilde{\mathbf{a}}; \bar{\mathbf{a}} \odot \tilde{\mathbf{a}}]$$

$$\mathbf{m}_b = [\bar{\mathbf{b}}; \tilde{\mathbf{b}}; \bar{\mathbf{b}} - \tilde{\mathbf{b}}; \bar{\mathbf{b}} \odot \tilde{\mathbf{b}}]$$

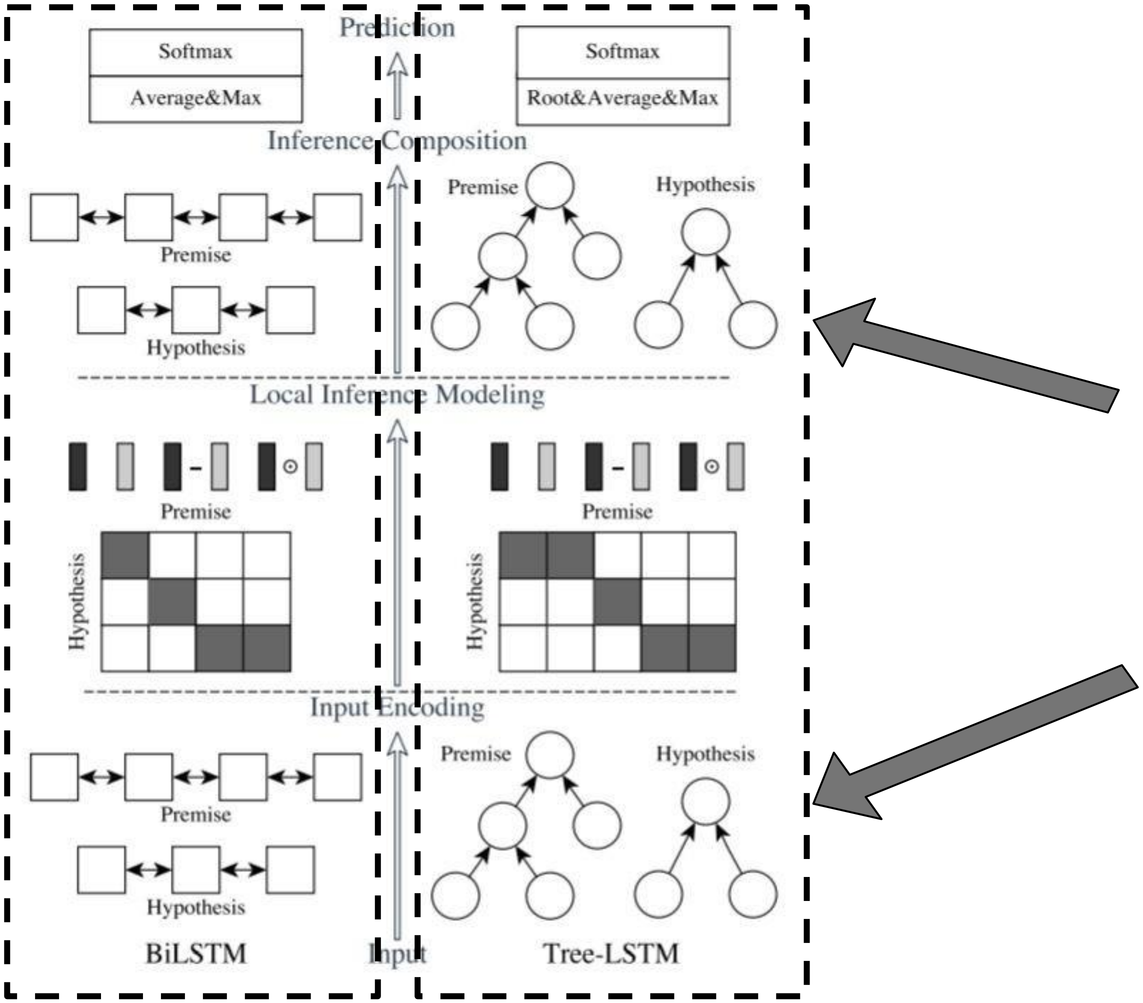
- Then by concatenating the average and max-pooling of m_a and m_b, we obtain a vector v which is fed to a classifier.

Performance of ESIM on SNLI

Model	#Para.	Train	Test
(1) Handcrafted features (Bowman et al., 2015)	-	99.7	78.2
(2) 300D LSTM encoders (Bowman et al., 2016)	3.0M	83.9	80.6
(3) 1024D pretrained GRU encoders (Vendrov et al., 2015)	15M	98.8	81.4
(4) 300D tree-based CNN encoders (Mou et al., 2016)	3.5M	83.3	82.1
(5) 300D SPINN-PI encoders (Bowman et al., 2016)	3.7M	89.2	83.2
(6) 600D BiLSTM intra-attention encoders (Liu et al., 2016)	2.8M	84.5	84.2
(7) 300D NSE encoders (Munkhdalai and Yu, 2016a)	3.0M	86.2	84.6
(8) 100D LSTM with attention (Rocktäschel et al., 2015)	250K	85.3	83.5
(9) 300D mLSTM (Wang and Jiang, 2016)	1.9M	92.0	86.1
(10) 450D LSTMN with deep attention fusion (Cheng et al., 2016)	3.4M	88.5	86.3
(11) 200D decomposable attention model (Parikh et al., 2016)	380K	89.5	86.3
(12) Intra-sentence attention + (11) (Parikh et al., 2016)	580K	90.5	86.8
(13) 300D NTI-SLSTM-LSTM (Munkhdalai and Yu, 2016b)	3.2M	88.5	87.3
(14) 300D re-read LSTM (Sha et al., 2016)	2.0M	90.7	87.5
(15) 300D btree-LSTM encoders (Paria et al., 2016)	2.0M	88.6	87.6
(16) 600D ESIM	4.3M	92.6	88.0

Models enhanced with syntactic structures

- Syntax has been used in many non-neural NLI/RTE systems (MacCartney, 2009; Dagan et al. 2013).
- How to explore syntactic structures in NN-based NLI systems? Several typical models:
    - Hierarchical Inference Models (**HIM**) (Chen et al., 2017)
    - Stack-augmented Parser-Interpreter Neural Network (**SPINN**) (Bowman et al., 2016)
    - Tree-Based CNN (**TBCNN**) (Mou et al., 2016)



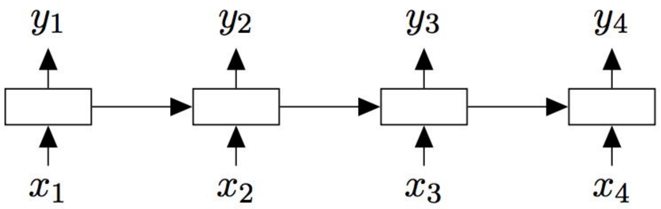
ESIM **HIM**

Parse information can be considered in different phases of NLI.

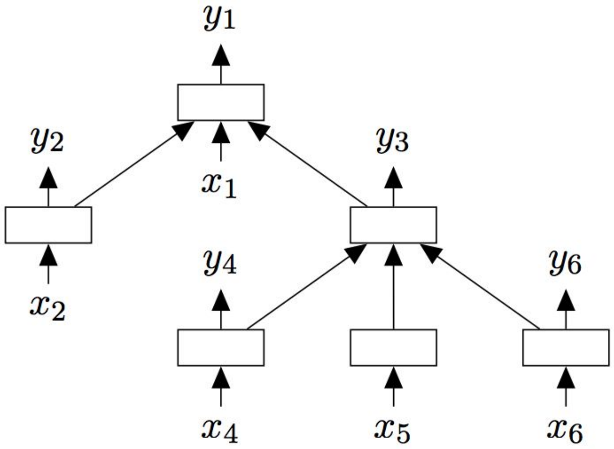
[Chen et al. '17]

Tree LSTM

Chain LSTM

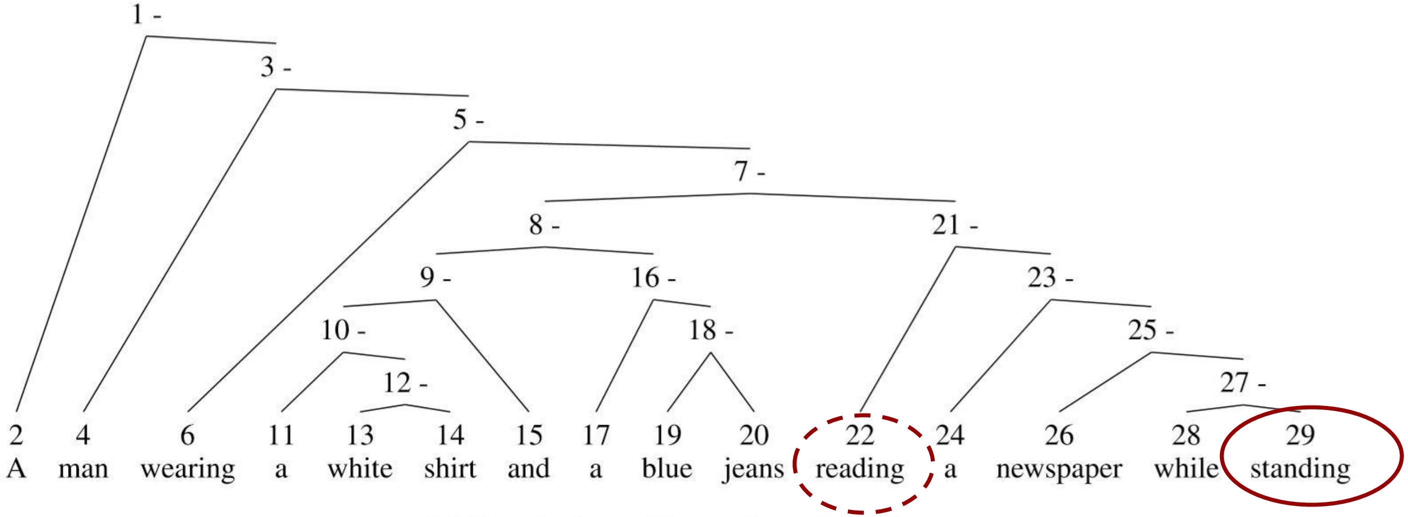


Tree LSTM

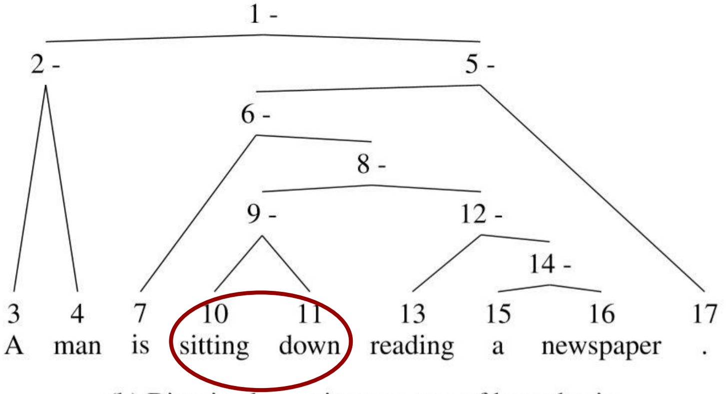


E.g., max branching N=3

Tai et al., 2015



(a) Binarized constituency tree of premise



(b) Binarized constituency tree of hypothesis

- Attention weights showed that the tree models aligned “sitting down” with “standing” and the classifier relied on that to make the correct judgement.
- The sequential model, however, soft-aligned “sitting” with both “reading” and “standing” and confused the classifier.

Performance of HIM on SNLI

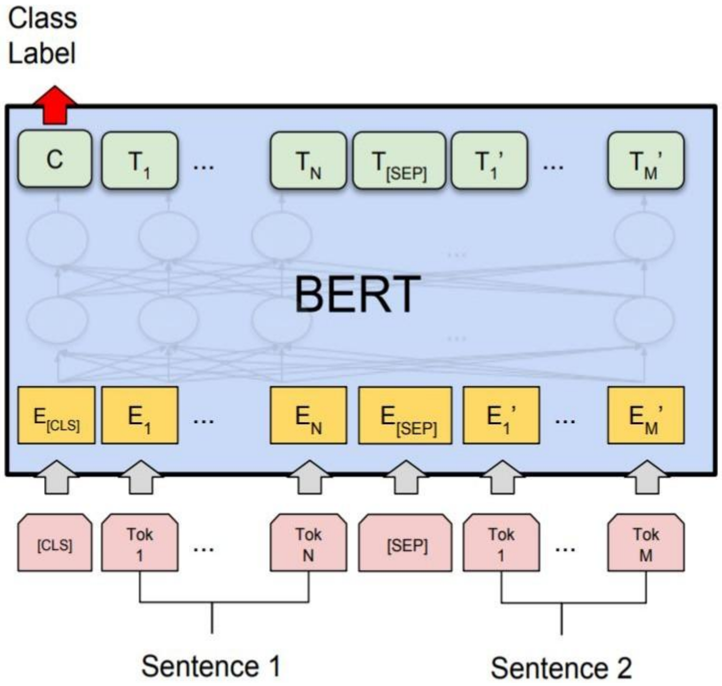
Model	#Para.	Train	Test
(1) Handcrafted features (Bowman et al., 2015)	-	99.7	78.2
(2) 300D LSTM encoders (Bowman et al., 2016)	3.0M	83.9	80.6
(3) 1024D pretrained GRU encoders (Vendrov et al., 2015)	15M	98.8	81.4
(4) 300D tree-based CNN encoders (Mou et al., 2016)	3.5M	83.3	82.1
(5) 300D SPINN-PI encoders (Bowman et al., 2016)	3.7M	89.2	83.2
(6) 600D BiLSTM intra-attention encoders (Liu et al., 2016)	2.8M	84.5	84.2
(7) 300D NSE encoders (Munkhdalai and Yu, 2016a)	3.0M	86.2	84.6
(8) 100D LSTM with attention (Rocktäschel et al., 2015)	250K	85.3	83.5
(9) 300D mLSTM (Wang and Jiang, 2016)	1.9M	92.0	86.1
(10) 450D LSTMN with deep attention fusion (Cheng et al., 2016)	3.4M	88.5	86.3
(11) 200D decomposable attention model (Parikh et al., 2016)	380K	89.5	86.3
(12) Intra-sentence attention + (11) (Parikh et al., 2016)	580K	90.5	86.8
(13) 300D NTI-SLSTM-LSTM (Munkhdalai and Yu, 2016b)	3.2M	88.5	87.3
(14) 300D re-read LSTM (Sha et al., 2016)	2.0M	90.7	87.5
(15) 300D btree-LSTM encoders (Paria et al., 2016)	2.0M	88.6	87.6
(16) 600D ESIM	4.3M	92.6	88.0
(17) HIM (600D ESIM + 300D Syntactic tree-LSTM)	7.7M	93.5	88.6

More recent methods

Deep learning models for NLI

- Baseline model with typical components
  - ESIM (Chen et al., 2017)
- Enhance with syntactic structures
  - HIM (Chen et al., 2017)
- Leverage unsupervised pretraining
  - BERT (Devlin et al., 2018)
- Enhance with semantic roles
  - SJRC (Zhang et al., 2019)

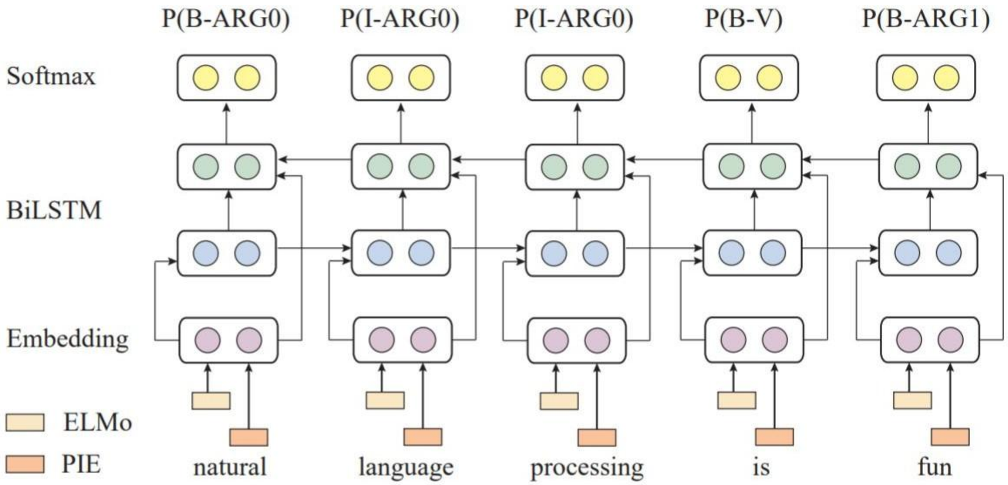
Models leveraging unsupervised pretraining



- Pretrained models can leverage large unannotated datasets, which have brought forward the state of the art of NLI and many other tasks.
  - See Peters et al., 2017, Radford et al., 2018, Devlin et al., 2018 for more details.
- E.g., BERT achieves a 90.4% accuracy on SNLI.

[Devlin et al. ‘18]

Models enhanced with semantic roles



Semantic role labeler.

- Recent research (Zhang et al., 2019) incorporated Semantic Role Labeling (SRL) into NLI and found it improved the performance.
- The proposed model simply concatenated SRL embedding into word embedding.

[Zhang et al. '19](#)

Models enhanced with semantic roles

Model	Accuracy (%)
DIIN	88.0
DR-BiLSTM	88.5
CAFE	88.5
MAN	88.3
KIM	88.6
DMAN	88.8
ESIM + TreeLSTM	88.6
ESIM + ELMo	88.7
DCRCN	88.9
LM-Transformer	89.9
MT-DNN†	91.1
Baseline (ELMo)	88.4
+ SRL	89.1
Baseline ($BERT_{BASE}$)	89.2
+ SRL	89.6
Baseline ($BERT_{LARGE}$)	90.4
+ SRL	**91.3**

Accuracy on SNLI

Zhang et al. '19

Artifacts in NLI

Example 1

- P:
- H: Someone is not crossing the road.
- Entailment? Neutral? Contradiction?

Example 2

- P:
- H: Someone is outside.
- Entailment? Neutral? Contradiction?

Artifacts in NLI

Example 1

- P:
- H: Someone is not crossing the road.
- Entailment? Neutral? **Contradiction**?

Example 2

- P:
- H: Someone is outside.
- **Entailment**? Neutral? Contradiction?

Artifacts in NLI

Entailment indicators

- Generic words (*animal, instrument, outdoors*)

Neutral indicators

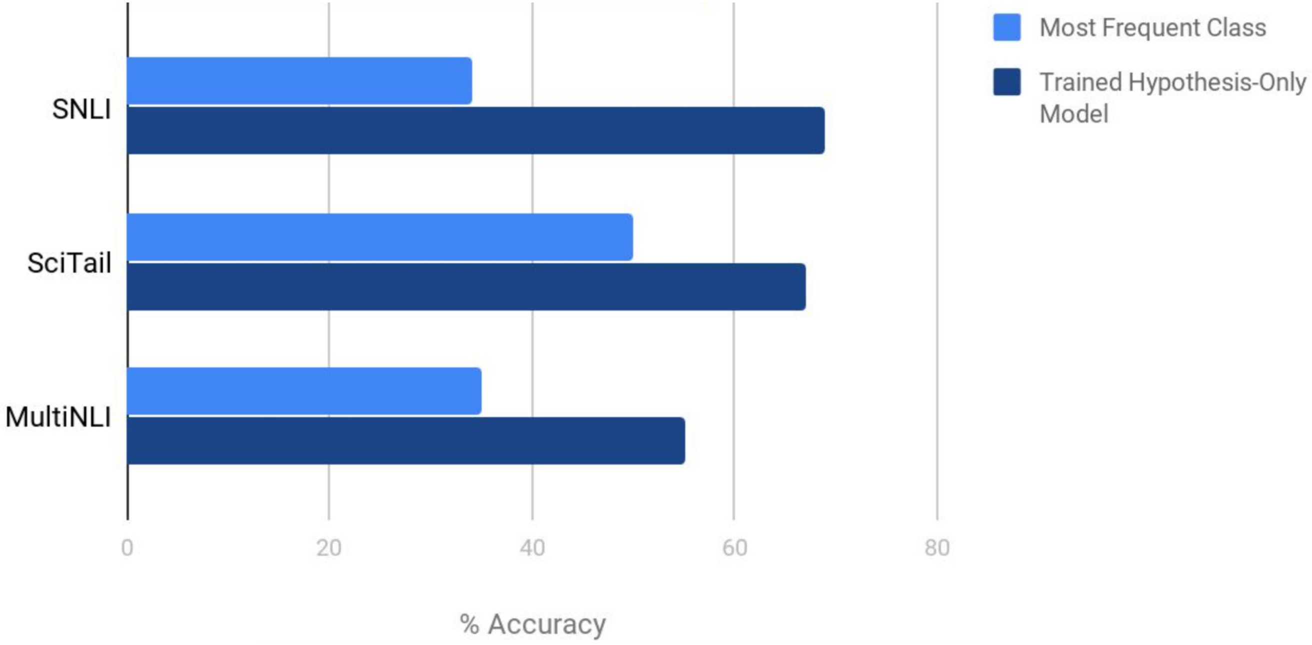
- Modifiers (*tall, sad, popular*) and superlatives (*first, favorite, most*)

Contradiction indicators

- Negation words (*nobody, no, never, nothing*)

Gururangan et al., 2018

Artifacts in NLI



- Models can do moderately well on NLI datasets without looking at the premise.

[Poliak et al., 2018](#)

Artifacts in NLI

Heuristic Analysis for NLI Systems (**HANS**) dataset

- Three syntactic heuristics that can be falsely manipulated by NLI models: **lexical overlap**, **subsequence**, and **constituent**.

Heuristic	Premise	Hypothesis
Lexical overlap heuristic	The banker near the judge saw the actor.	The banker saw the actor.
	The lawyer was advised by the actor.	The actor advised the lawyer.
	The doctors visited the lawyer.	The lawyer visited the doctors.
	The judge by the actor stopped the banker.	The banker stopped the actor.

McCoy et al., 2019

Artifacts in NLI

Heuristic Analysis for NLI Systems (**HANS**) dataset

- Three syntactic heuristics that can be falsely manipulated by NLI models: **lexical overlap**, **subsequence**, and **constituent**.

entailment

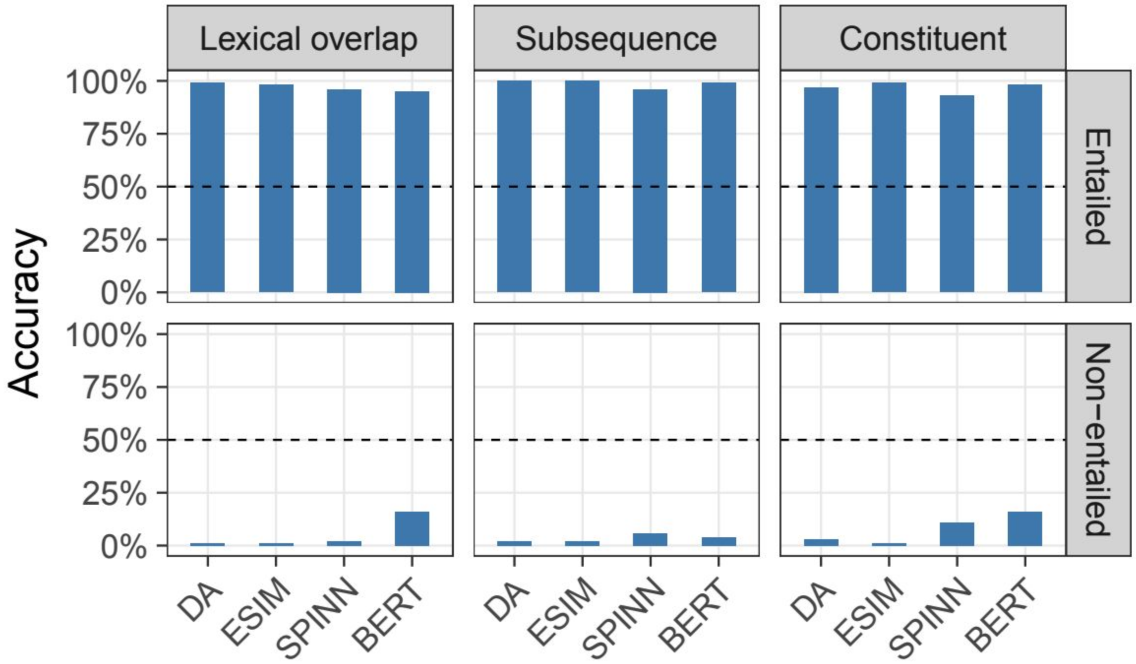
Heuristic	Premise	Hypothesis
Lexical overlap heuristic	The banker near the judge saw the actor.	The banker saw the actor.
	The lawyer was advised by the actor.	The actor advised the lawyer.
	The doctors visited the lawyer.	The lawyer visited the doctors.
	The judge by the actor stopped the banker.	The banker stopped the actor.

non-entailment

McCoy et al., 2019

Artifacts in NLI

Heuristic Analysis for NLI Systems (**HANS**) dataset



McCoy et al., 2019

Artifacts in NLI

Knowing that NLI models are vulnerable to data artifacts, a natural next question could be:

- Why does an NLI model make *each* entailment / non-entailment prediction?
  - Not all examples have indicative words like “animals” or “outdoors”, or satisfy the heuristics.

- Why does an NLP model make each of its decision?

Questions?

Part II: Interpretable NLP

Xiaochuang Han

with content borrowed from Byron Wallace and Sarthak Jain

Why is interpretability important?

Geoffrey Hinton
@geoffreyhinton

Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

3:37 PM · Feb 20, 2020 · Twitter Web App

1.1K Retweets **614** Quote Tweets **5.2K** Likes

Defining interpretability?

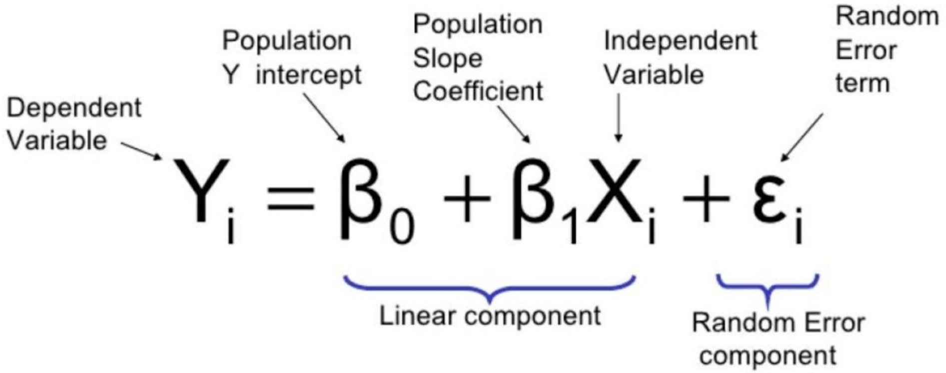
- There is no standard definition :)

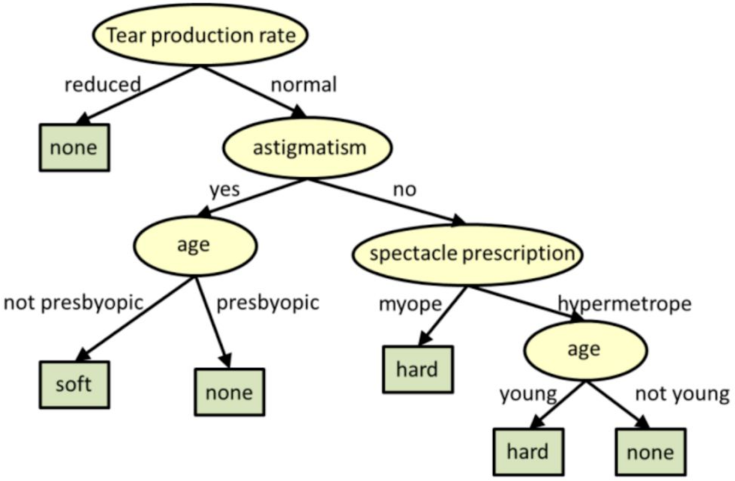
- Ability to explain or to present a model in understandable terms to humans (Doshi-Velez and Kim, 2017).

- It depends on the target audience.

What does interpretation look like?

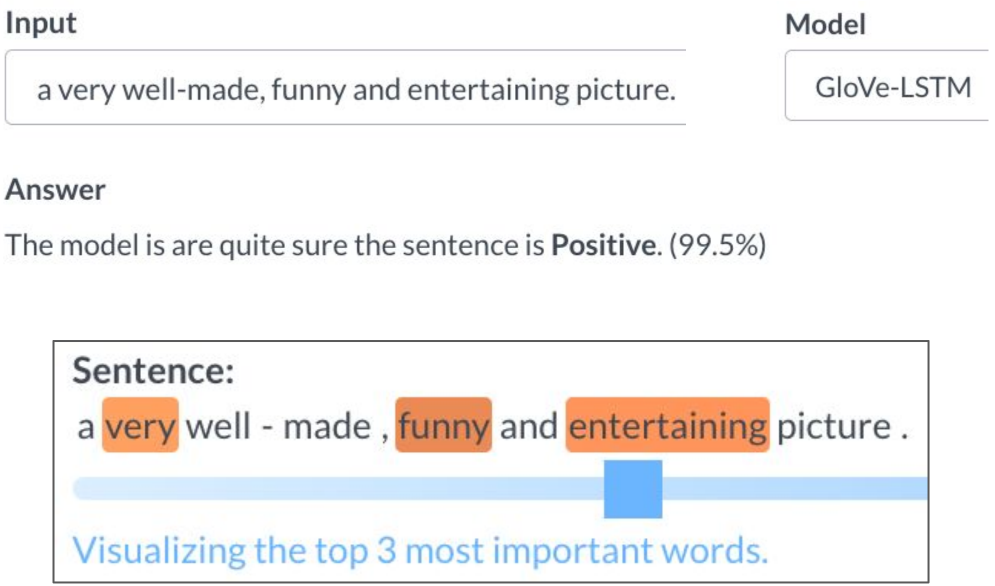
- In pre-deep learning models, some models are considered “interpretable”.





What does interpretation look like?

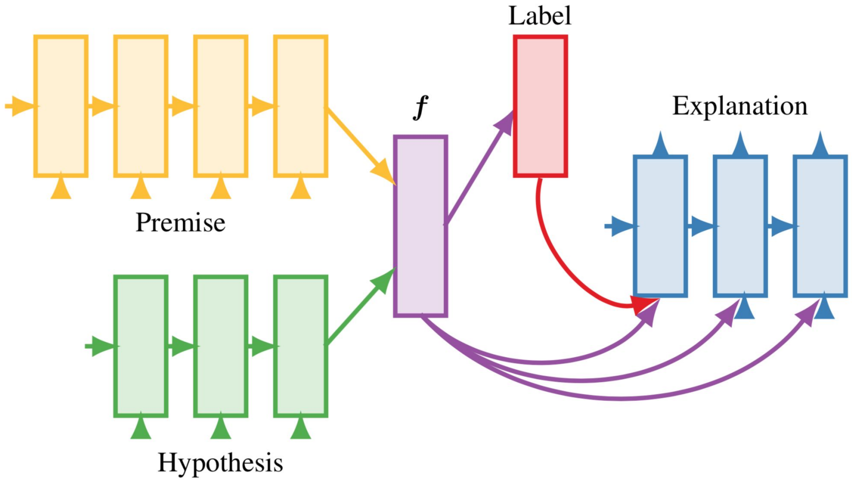
- Heatmap visualization over input
  - AllenNLP Interpret demo (Wallace et al., 2019)



What does interpretation look like?

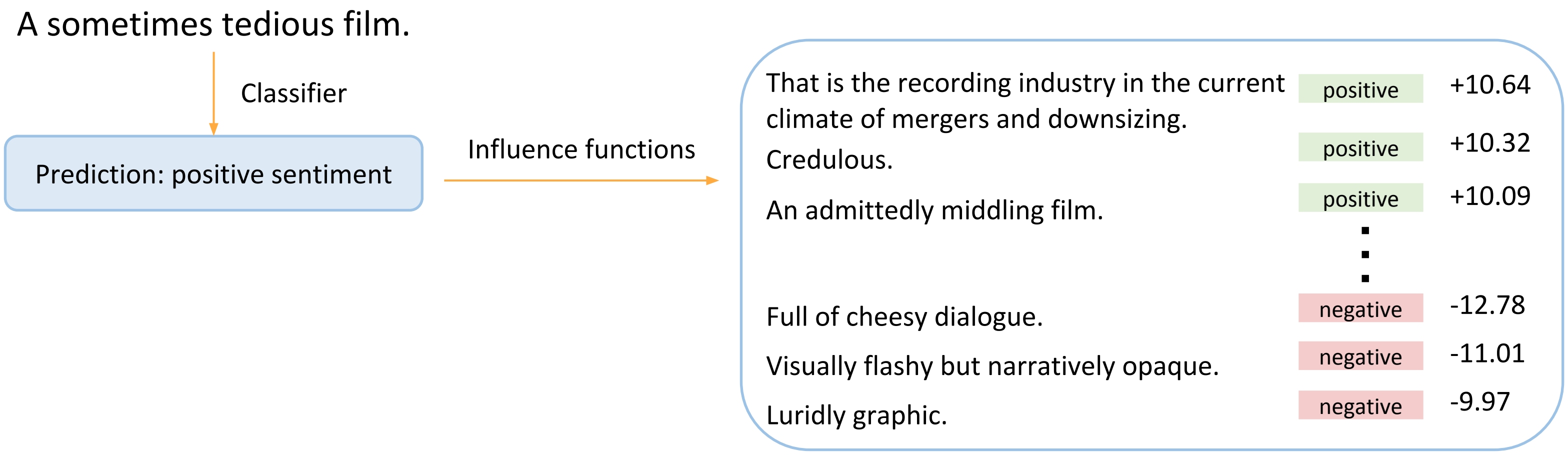
- Generate rationales as text
  - e-SNLI (Camburu et al., 2018)

Premise: A man in an orange vest leans over a pickup truck.
Hypothesis: A man is touching a truck.
Label: entailment
Explanation: Man leans over a pickup truck implies that he is touching it.



What does interpretation look like?

- Explain with influential training examples
  - Influence functions (Koh and Liang, 2017; Han et al., 2020)



Influential examples in the training corpus

Some properties of interpretations

- Faithfulness
  - How to provide explanations that accurately represent the true reasoning behind the model's final decision.
- Plausibility
  - Is the explanation correct or something we can believe is true, given our current knowledge of the problem?
- Understandable
  - Can I put it in terms that end user without in-depth knowledge of the system can understand?
- Stability
  - Does similar instances have similar interpretations?

Some categories of interpretations

Local vs. Global

- Do we explain individual prediction?
- Do we explain entire model?

Inherent vs. Post-hoc

- Is the explainability built into the model?
- Is the model black-box and we use external method to try to understand it?

Some categories of interpretations

Local vs. Global

- Do we explain individual prediction?
    - Heatmaps, rationales, influential training examples, …
- Do we explain entire model?
    -

Inherent vs. Post-hoc

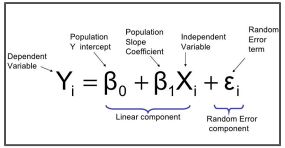
- Is the explainability built into the model?
    -
- Is the model black-box and we use external method to try to understand it?
    -

Some categories of interpretations

Local vs. Global

- Do we explain individual prediction?
  - Heatmaps, rationales, influential training examples, …

- Do we explain entire model?
  - Linear models, …



Inherent vs. Post-hoc

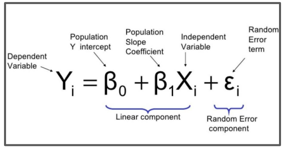
- Is the explainability built into the model?
- Is the model black-box and we use external method to try to understand it?

Some categories of interpretations

Local vs. Global

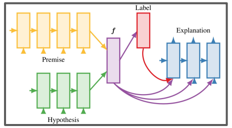
- Do we explain individual prediction?
  - Heatmaps, rationales, influential training examples, …

- Do we explain entire model?
  - Linear models, …



Inherent vs. Post-hoc

- Is the explainability built into the model?
  - Linear models, rationales, …



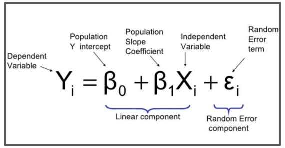
- Is the model black-box and we use external method to try to understand it?
  -

Some categories of interpretations

Local vs. Global

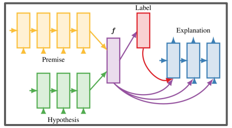
- Do we explain individual prediction?
  - Heatmaps, rationales, influential training examples, …

- Do we explain entire model?
  - Linear models, …

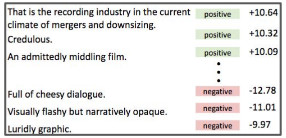


Inherent vs. Post-hoc

- Is the explainability built into the model?
  - Linear models, rationales, …



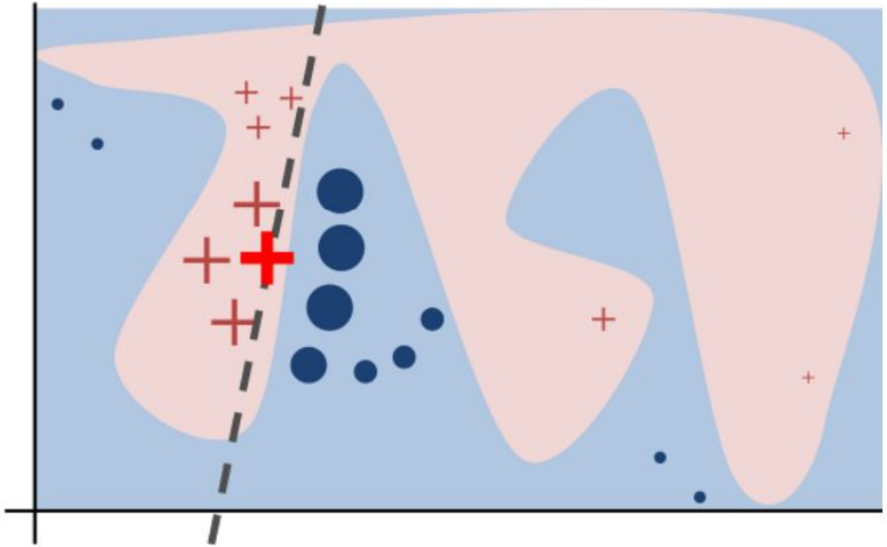
- Is the model black-box and we use external method to try to understand it?
  - **Heatmaps**, **influential training examples**, …



That is the recording industry in the current climate of mergers and downsizing.	positive	+10.64
Credulous.	positive	+10.32
An admittedly middling film.	positive	+10.09
⋮	⋮	
Full of cheesy dialogue.	negative	-12.78
Visually flashy but narratively opaque.	negative	-11.01
Luridly graphic.	negative	-9.97

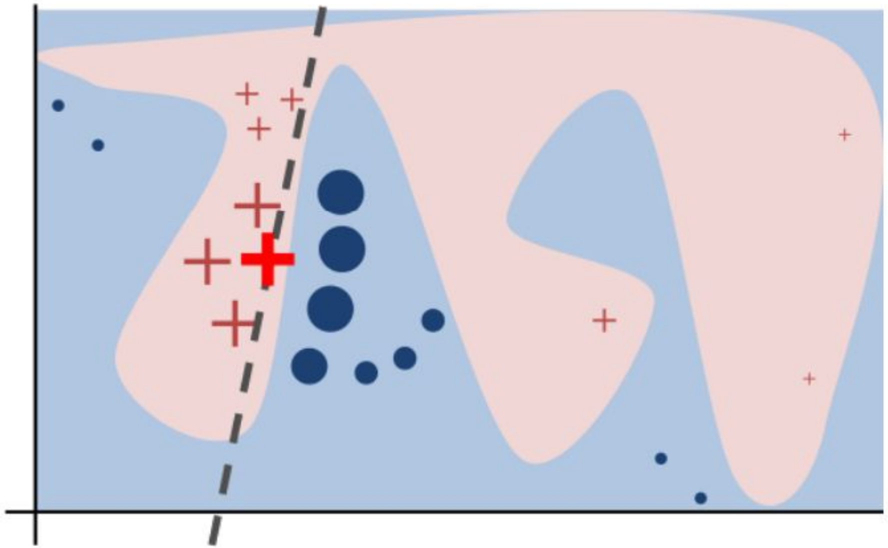
Local Interpretable Model-agnostic Explanations (**LIME**)

- Approximate a black-box model using linear models
- Cannot do this globally, but what about locally?
    - Ribeiro et al., 2016



Local Interpretable Model-agnostic Explanations (**LIME**)

- Approximate a black-box model using linear models
- Cannot do this globally, but what about locally?
    - Ribeiro et al., 2016

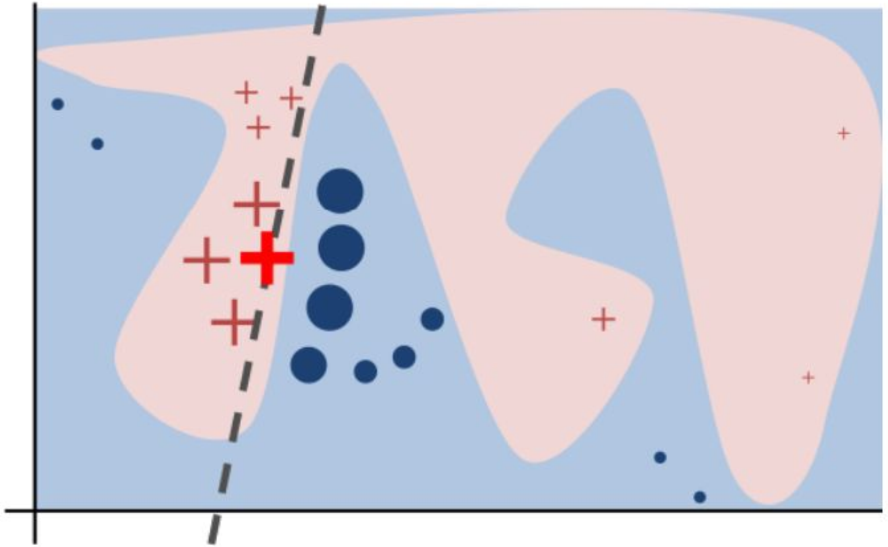


$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Local Interpretable Model-agnostic Explanations (**LIME**)

- Approximate a black-box model using linear models
- Cannot do this globally, but what about locally?
  - Ribeiro et al., 2016

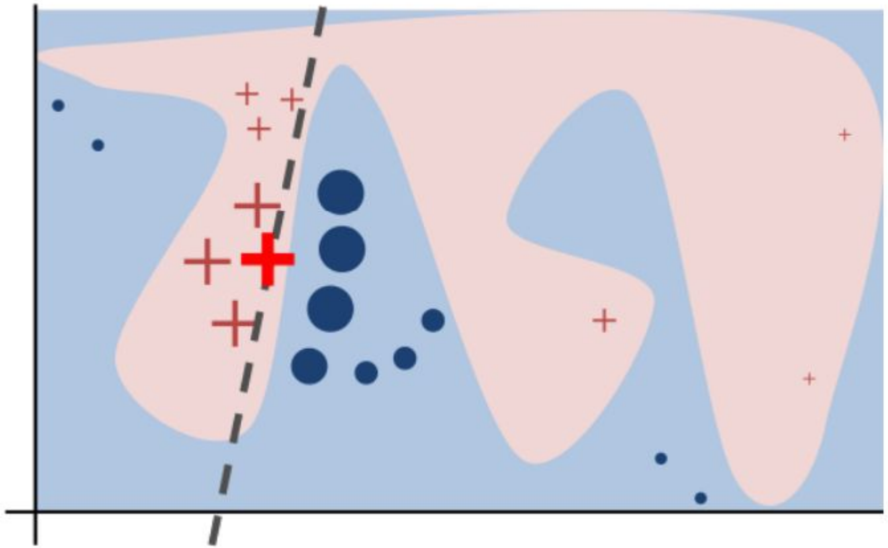


$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \overbrace{\pi_x(z)}^{\text{similarity kernel}} \Big(\overbrace{f(z)}^{\text{black-box classifier}} - \overbrace{g(z')}^{\text{linear model}} \Big)^2$$

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Local Interpretable Model-agnostic Explanations (**LIME**)

- Approximate a black-box model using linear models
- Cannot do this globally, but what about locally?
    - Ribeiro et al., 2016



similarity kernel

black-box classifier

linear model

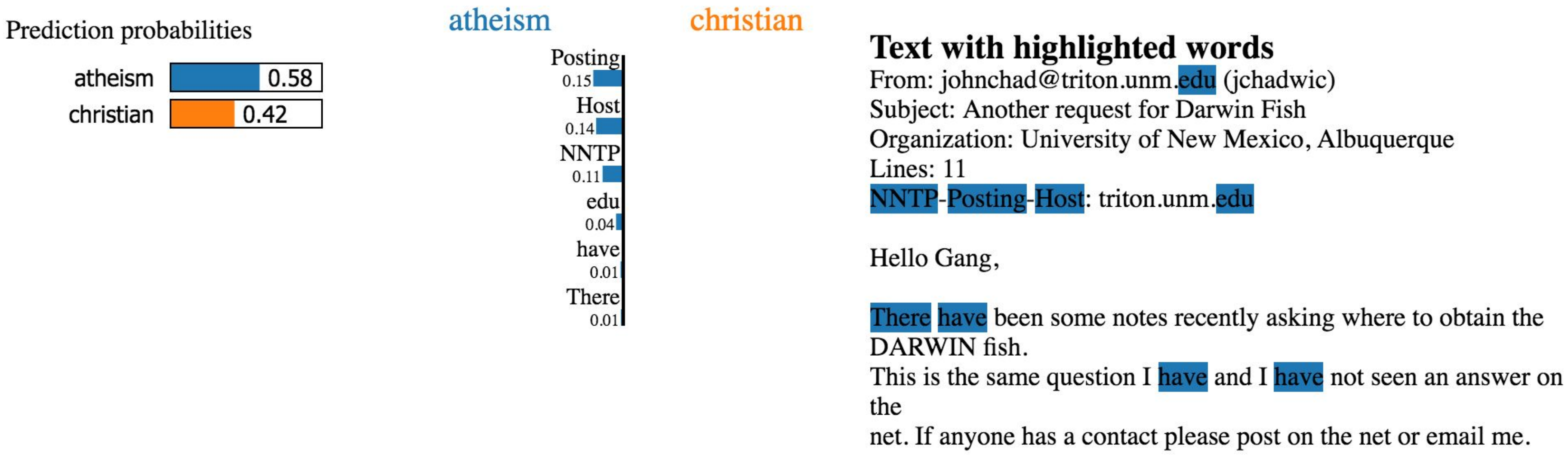
$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z') \right)^2$$

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Match interpretable model to black box

Control complexity of the interpretable model

Local Interpretable Model-agnostic Explanations (**LIME**)



An example LIME interpretation for a test input

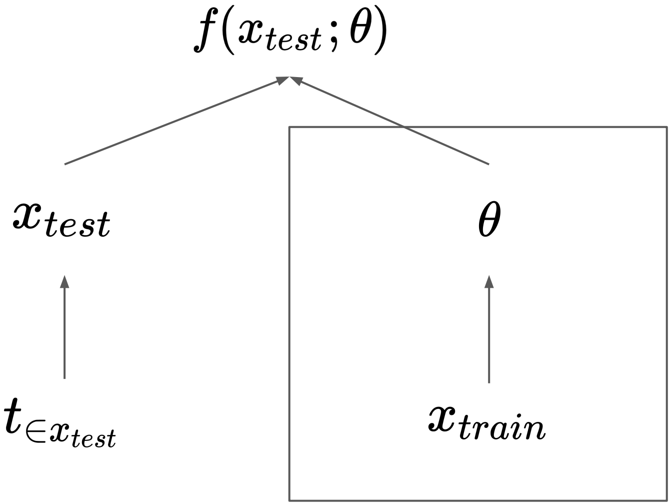
More heatmap methods

- Gradient-based saliency maps
  - Simonyan et al., 2014; Shrikumar et al., 2017; Sundararajan et al., 2017; Smilkov et al., 2017
- SHAP
  - Lundberg and Lee, 2017
- Attention scores?
  - Jain and Wallace, 2019; Wiegreffe and Pinter, 2019

Another perspective

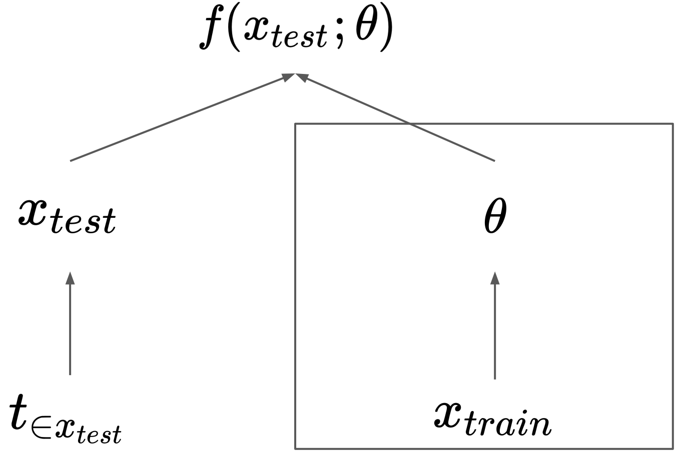
$$
\begin{tikzcd}
 & f(x_{test};\theta) \\
x_{test} \arrow[ur] & \\
t_{\in x_{test}} \arrow[u] &
\end{tikzcd}
$$

Another perspective



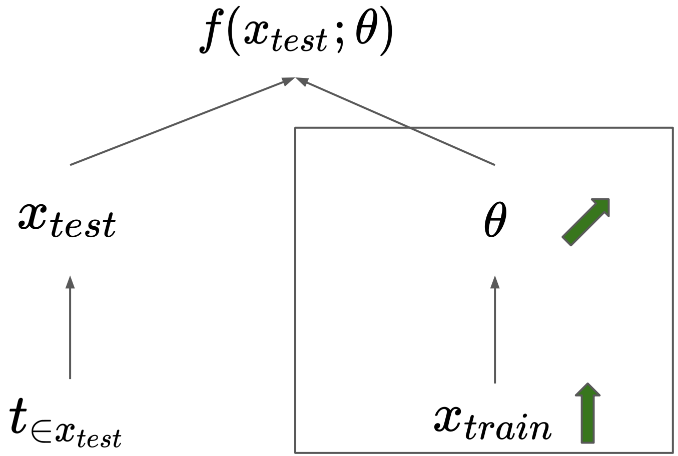
Influence functions

- The black-box model learns a set of parameters that minimize the training loss, which comes from all the training examples equally (i.i.d.).



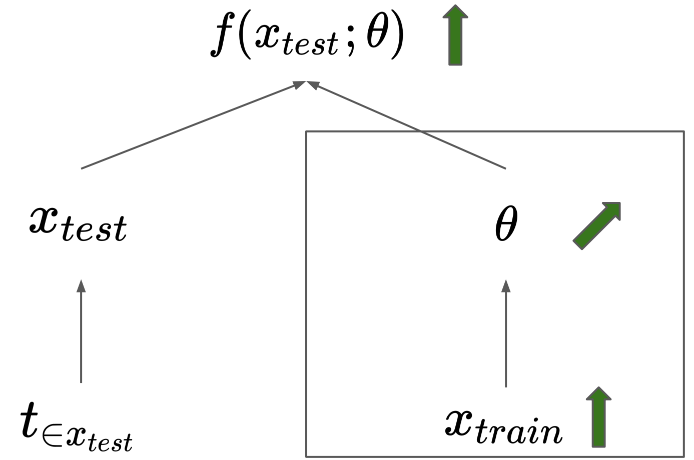
Influence functions

- The black-box model learns a set of parameters that minimize the training loss, which comes from all the training examples equally (i.i.d.).
- If we upweight a single training example, the potential model parameters would change.



Influence functions

- The black-box model learns a set of parameters that minimize the training loss, which comes from all the training examples equally (i.i.d.).
- If we upweight a single training example, the potential model parameters would change.
- The decision (probability) on the test input would also change, which can be attributed back to that training example.



Influence functions

1. How would an upweight to a training example (x_i, y_i) change the learned model parameters?
    - i.e., taking a single Newton step from the originally learned θ

1

$$\frac{d\hat{\theta}}{d\epsilon_i} = -\left(\frac{1}{n}\sum_{j=1}^{n} \nabla_\theta^2 \mathcal{L}(x_j, y_j, \hat{\theta})\right)^{-1} \nabla_\theta \mathcal{L}(x_i, y_i, \hat{\theta})$$

Influence functions

1. How would an upweight to a training example (x_i, y_i) change the learned model parameters?
   - i.e., taking a single Newton step from the originally learned θ
2. How would this change in the model parameters change the model decision?

1 $$\frac{d\hat{\theta}}{d\epsilon_i} = -\left(\frac{1}{n}\sum_{j=1}^{n} \nabla_\theta^2 \mathcal{L}(x_j, y_j, \hat{\theta})\right)^{-1} \nabla_\theta \mathcal{L}(x_i, y_i, \hat{\theta})$$

2 $$\frac{d\mathcal{L}_{\hat{y}}}{d\epsilon_i} = \nabla_\theta \mathcal{L}_{\hat{y}} \cdot \frac{d\hat{\theta}}{d\epsilon_i}$$

Influence functions

1. How would an upweight to a training example (x_i, y_i) change the learned model parameters?
    - i.e., taking a single Newton step from the originally learned θ
2. How would this change in the model parameters change the model decision?
3. A training example that leads to a more confident test decision / lower test loss is more (positively) influential.

$$\boxed{1} \quad \frac{d\hat{\theta}}{d\epsilon_i} = -\left(\tfrac{1}{n}\textstyle\sum_{j=1}^{n} \nabla_\theta^2 \mathcal{L}(x_j, y_j, \hat{\theta})\right)^{-1} \nabla_\theta \mathcal{L}(x_i, y_i, \hat{\theta})$$

$$\boxed{2} \quad \frac{d\mathcal{L}_{\hat{y}}}{d\epsilon_i} = \nabla_\theta \mathcal{L}_{\hat{y}} \cdot \frac{d\hat{\theta}}{d\epsilon_i}$$

$$\boxed{3} \quad s((x_i, y_i)) = -\frac{d\mathcal{L}_{\hat{y}}}{d\epsilon_i}$$

Influence functions example (back to NLI)

P: The manager was encouraged by the secretary.
H: The secretary encouraged the manager.
[entailment]

Test input, from HANS

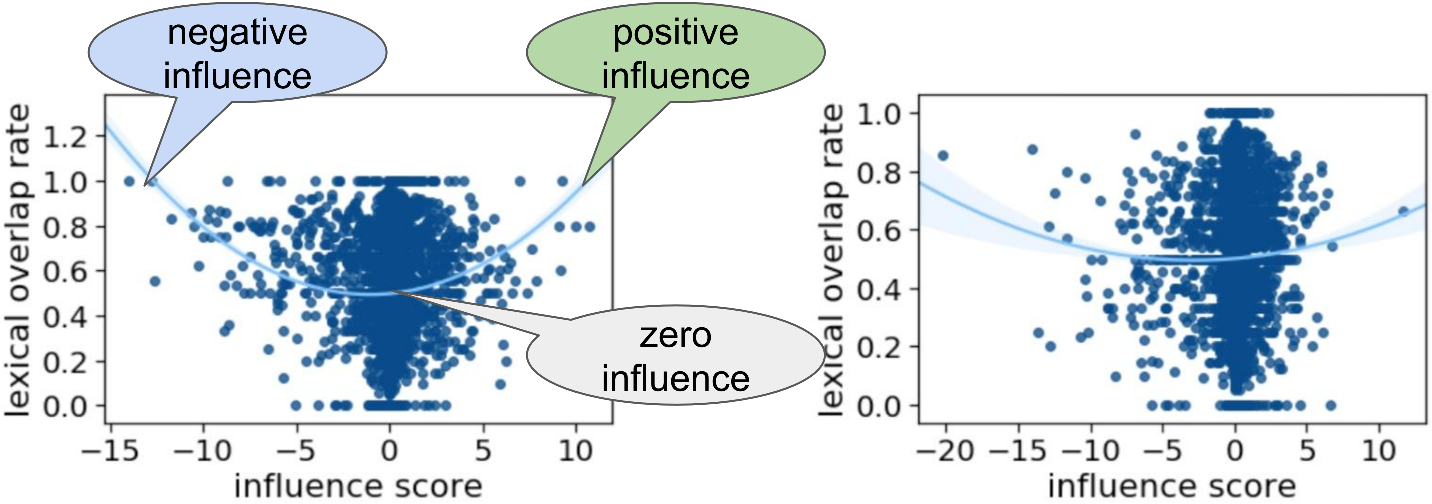
P: Because you’re having fun.
H: Because you’re having fun.
[entailment]

P: Do it now, think ’bout it later.
H: Don’t think about it now, just do it.
[entailment]

Most influential training examples, from MNLI

“Why does our model makes an entailment decision?”

Influence functions example (back to NLI)



Avg coef for HANS: $+3.28 \times 10^{-3}$

Avg coef for MNLI: $+0.65 \times 10^{-3}$

See more details in [Han et al., 2020]

Still a very open question

- What types of interpretations should we adopt for different models, tasks, and groups of users?

- Recent trend in continuous stress tests (non-i.i.d.) for NLP models indicates that the models might not be as robust as they first seem. Does good interpretability translate to more robust models?

Plan for today

- Part I: Natural Language Inference
  - Definition and background
  - Datasets (**RTE, SNLI, MNLI, XNLI, SciTail**)
  - Models (**Natural logic, ESIM, ESIM+Tree LSTM, BERT, BERT+SRL**)
  - Problems (**Data artifacts, challenge set HANS**)
- Part II: Interpretable NLP
  - Motivation
  - Major approaches (**Heatmaps, rationale generation, explain with training examples**)
  - Detailed methods (**LIME, influence functions**)
- Questions?