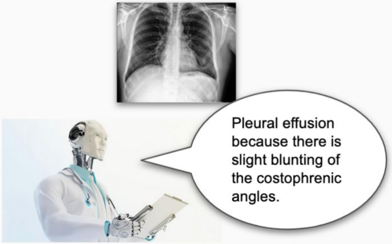
Can AI Models Generate Correct and Faithful Natural Language Explanations for Their Predictions?

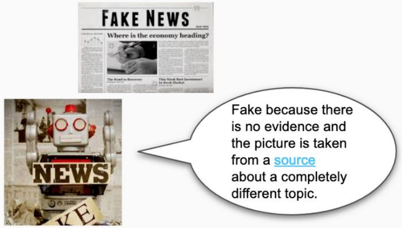
Oana-Maria Camburu

Senior Research Fellow

Leverhulme Early Career Fellowship

UCL NLP Group





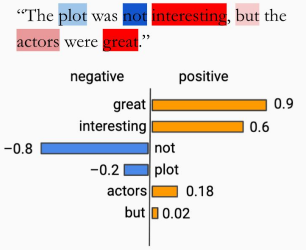
Outline

1. Introduction
2. AI with Natural Language Explanations, <u>bits and pieces</u> from:
    i. e-SNLI: Natural Language Inference with Natural Language Explanations (NeurIPS'18)
    ii. Faithfulness Tests for Natural Language Explanations (ACL'23)
    iii. Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (ACL'20)
    iv. KNOW How to Make Up Your Mind! Adversarially Detecting and Alleviating Inconsistencies in Natural Language Explanations (ACL'23)
    v. Explaining Chest X-ray Pathologies in Natural Language (MICCAI'22)
    vi. e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks (ICCV'21)
    vii. Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations (ICML'22)
3. Open Questions

Introduction

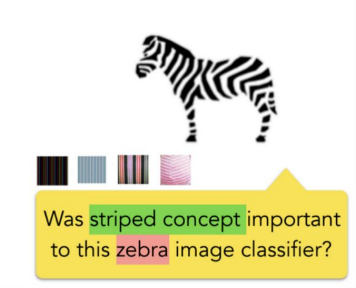
Introduction

Types of explanations



M. Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD, 2016.
S. Lundberg and S. Lee, A Unified Approach to Interpreting Model Predictions, NeurIPS, 2017.
M. Sundararajan, Axiomatic Attribution for Deep Networks, ICML, 2017.

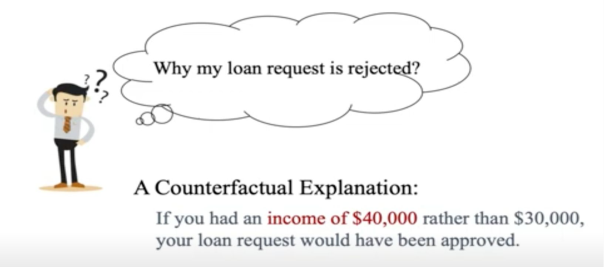
Feature importance



https://medium.com/intuit-engineering/navigating-the-sea-of-explainability-f6cc4631f473

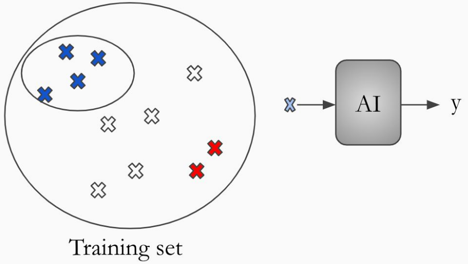
B. Kim et al., Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

Concept based



From https://www.youtube.com/watch?v=wVrJ5youWNU&ab_channel=IEEEVisualizationConference, March 2022

Counterfactuals



P. Koh and P. Liang, Understanding Black-box Predictions via Influence Functions, ICML, 2017.

Training examples

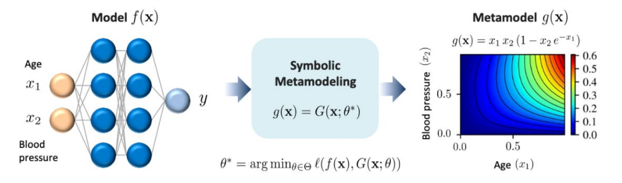


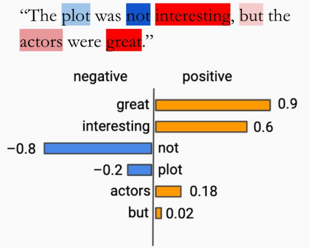
Figure 1: **Pictorial depiction of the symbolic metamodeling framework.** Here, the model $f(\mathbf{x})$ is a deep neural network (left), and the metamodel $g(\mathbf{x})$ is a closed-form expression $x_1\, x_2\, (1 - x_2\, \exp(-x_1))$ (right).

A. Alaa and M. van der Shaar, Demystifying Black-box Models with Symbolic Metamodels, NeurIPS, 2019

Surrogate models

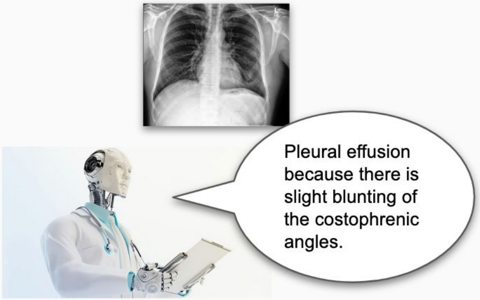
Introduction

Types of explanations

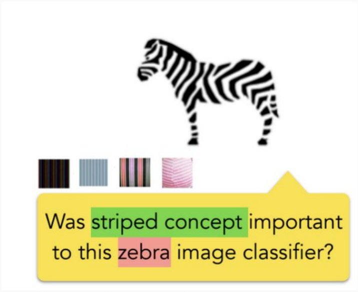


M. Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD, 2016.
S. Lundberg and S. Lee, A Unified Approach to Interpreting Model Predictions, NeurIPS, 2017.
M. Sundararajan, Axiomatic Attribution for Deep Networks, ICML, 2017.

Feature importance



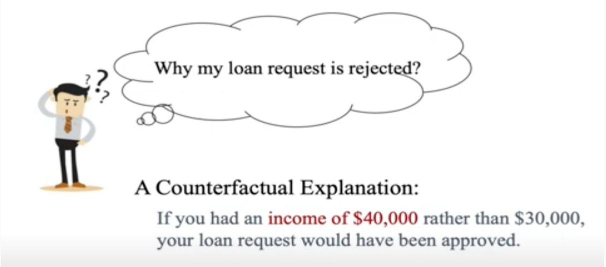
Natural Language Explanations (NLEs)



https://medium.com/intuit-engineering/navigating-the-sea-of-explainability-f6cc4631f473

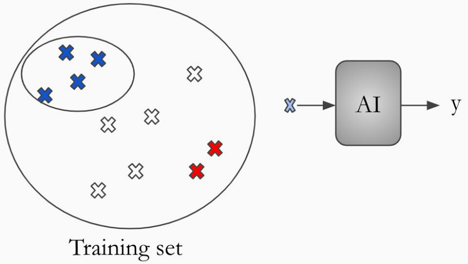
B. Kim et al., Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

Concept based



From https://www.youtube.com/watch?v=wVrJ5youWNU&ab_channel=IEEEVisualizationConference, March 2022

Counterfactuals



P. Koh and P. Liang, Understanding Black-box Predictions via Influence Functions, ICML, 2017.

Training examples

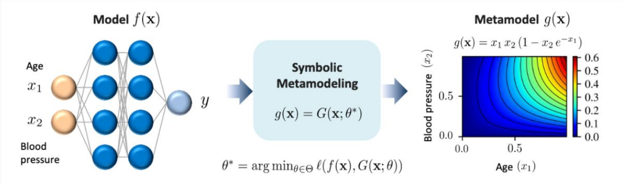


Figure 1: **Pictorial depiction of the symbolic metamodeling framework.** Here, the model $f(\mathbf{x})$ is a deep neural network (left), and the metamodel $g(\mathbf{x})$ is a closed-form expression $x_1\, x_2\, (1 - x_2 \exp(-x_1))$ (right).

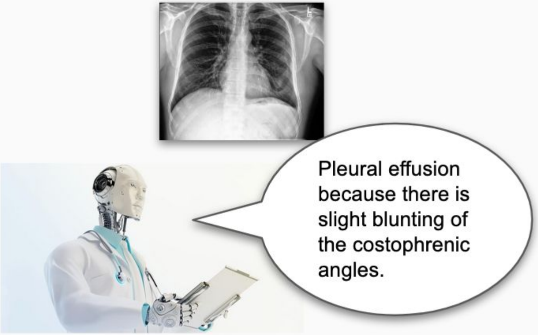
A. Alaa and M. van der Shaar, Demystifying Black-box Models with Symbolic Metamodels, NeurIPS, 2019

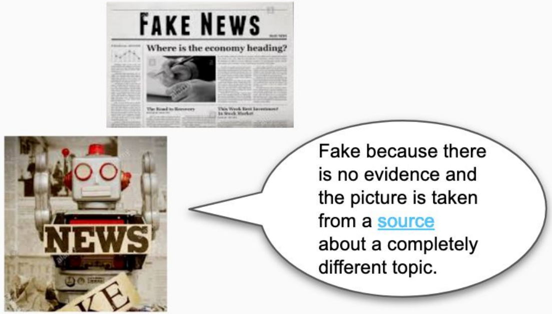
Surrogate models

Natural Language Explanations (NLEs)

Models

- **generate** NLEs for their predictions at deployment time
- (**learn** from NLEs for the ground-truth answers at training/prompting time)





Natural Language Explanations (NLEs)

Motivation

- **Human-intelligible explanations**. Kaur et al. (2020): “**few** of our participants [**197 data scientists**] were able to **accurately describe** the visualizations output by these tools [feature importance]” and “**data scientists over-trust** and **misuse** interpretability tools”.

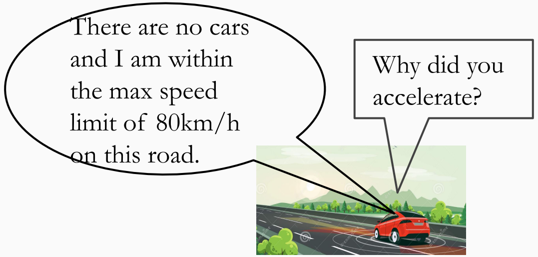
H. Kaur et al., Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning, CHI 2020.

Natural Language Explanations (NLEs)

Motivation

- **Human-intelligible explanations**. Kaur et al. (2020): “**few** of our participants [**197 data scientists**] were able to **accurately describe** the visualizations output by these tools [feature importance]” and “**data scientists over-trust** and **misuse** interpretability tools”.

- Allow for **comprehensive** justifications, **filling in reasoning and background knowledge** that is not present in the input.

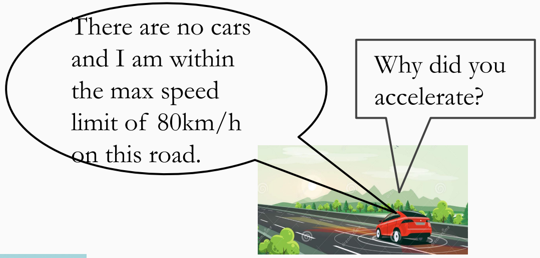


Natural Language Explanations (NLEs)

Motivation

- **Human-intelligible explanations**. Kaur et al. (2020): "**few** of our participants [**197 data scientists**] were able to **accurately describe** the visualizations output by these tools [feature importance]" and "**data scientists over-trust** and **misuse** interpretability tools".

- Allow for **comprehensive** justifications, **filling in reasoning and background knowledge** that is not present in the input.



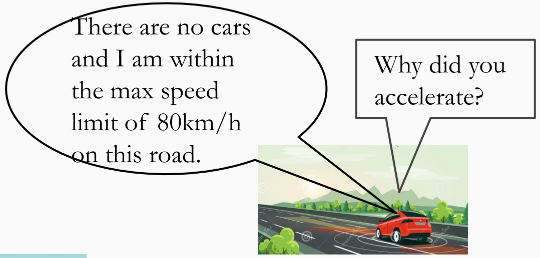
- **Adapt to the audience.**

Natural Language Explanations (NLEs)

Motivation

- **Human-intelligible explanations**. Kaur et al. (2020): "**few** of our participants [**197 data scientists**] were able to **accurately describe** the visualizations output by these tools [feature importance]" and "**data scientists over-trust** and **misuse** interpretability tools".

- Allow for **comprehensive** justifications, **filling in reasoning and background knowledge** that is not present in the input.



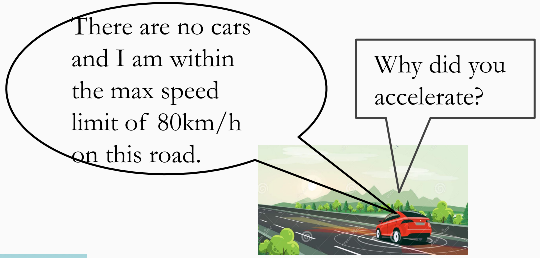
- **Adapt to the audience.**

- **Additional rich signal** at training/prompting time may lead to better performance and robustness. Humans don't learn just from labelled examples.

Natural Language Explanations (NLEs)

Motivation

- **Human-intelligible explanations**. Kaur et al. (2020): "**few** of our participants [**197 data scientists**] were able to **accurately describe** the visualizations output by these tools [feature importance]" and "**data scientists over-trust** and **misuse** interpretability tools".
- Allow for **comprehensive** justifications, **filling in reasoning and background knowledge** that is not present in the input.



- **Adapt to the audience.**

- **Additional rich signal** at training/prompting time may lead to better performance and robustness. Humans don't learn just from labelled examples.

Natural Language Explanations (NLEs)

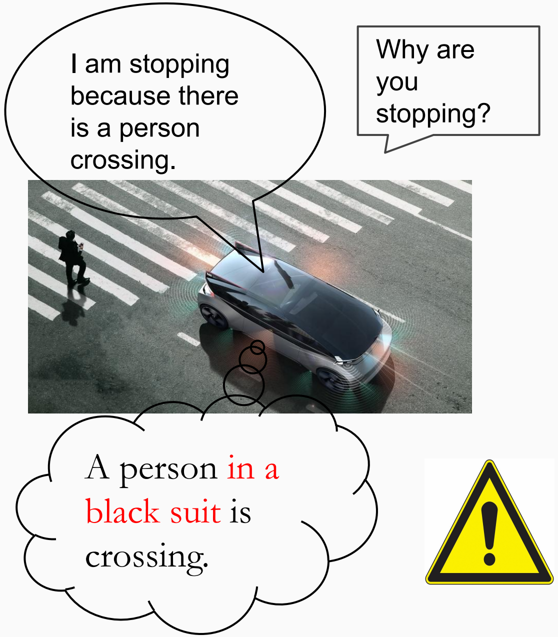
Criteria

- **Correctness**: Does the explanation give the correct reasons for the correct prediction?

Natural Language Explanations (NLEs)

Criteria

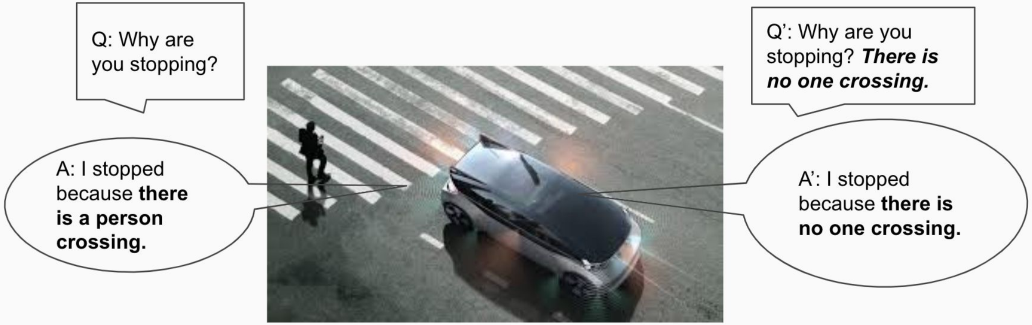
- **Correctness**: Does the explanation give the correct reasons for the correct prediction?
- **Faithfulness:** Is the explanation faithful to the decision-making process of the model?



Natural Language Explanations (NLEs)

Criteria

- **Consistency:** Is the model consistent in its explanations across instances?



Inconsistency → Incorrectness OR Unfaithfulness

"Trust is built with consistency"
Lincoln Chafee.

Natural Language Explanations (NLEs)

Criteria

- **Correctness**: Does the explanation give the correct reasons for the correct prediction?
- **Faithfulness:** Is the explanation faithful to the decision-making process of the model?
- **Consistency:** Is the model consistent in its explanations across instances?

…

Correctness: Does the explanation give the correct reasons for the correct prediction?

e-SNLI

- Train (~550k): 1 NLE / instance; Dev and Test (~10k): 3 NLEs / instance + annotation of salient tokens

Premise:
Two women are embracing while holding to go packages.

Hypothesis:
Two women are holding food in their hands.

Label:
Entailment

Explanation: Holding to go packages implies that there is food in it.

Premise:
A black race car starts up in front of a crowd of people.

Hypothesis:
A man is driving down a lonely road.

Label:
Contradiction

Explanation: A road can't be lonely if there is a crowd of people.

Premise:
A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.

Hypothesis:
A man is repainting a garage

Label:
Neutral

Explanation: It is not clear whether the man is repainting the garage or not.

Is e-SNLI too easy for current LLMs?

Likely yes, but out of 8 random examples from e-SNLI, ChatGPT got 2 wrong.

e-SNLI: Contradiction, “If a man is wearing a backwards baseball cap, then his bald head will not be exposed and burned by the sun.”

You have to solve the task of natural language inference and explain your answer. Given a premise and a hypothesis, if the premise implies the hypothesis, the answer is "entailment", if the hypothesis contradicts the premise the answer is “contradiction”, does not logically imply the hypothesis, the answer is "neutral". Write a brief reason to explain the answer. Here is an example.

Premise: A woman is in the park.
Hypothesis: A person is outdoors.
Answer: entailment
Explanation: A woman is a person and the park implies the outdoors.

Premise: A woman is in the park.
Hypothesis: The woman is indoors.
Answer: contradiction
Explanation: The woman is in the park so not indoors

Premise: A woman is in the park.
Hypothesis: The blond woman is outdoors.
Answer: neutral
Explanation: The premise doesn’t mention anything about the hair color of the woman.

Premise: Man in boat wearing athletic shorts, a backwards baseball cap and sunglasses unhooks a fish he has caught while fishing.
Hypothesis: His bald head is being burned by the sun.
Answer:

Answer: entailment

Explanation: The premise describes a man in a boat who is wearing athletic shorts, a backwards baseball cap, and sunglasses. The action of the man unhooks a fish he has caught while fishing. Based on this information, it can be inferred that the man is exposed to the sun. The hypothesis states that his bald head is being burned by the sun, which aligns with the implication from the premise. Therefore, the premise implies the hypothesis, leading to an answer of "entailment."

Is e-SNLI too easy for current LLMs?

Likely yes, but out of 8 random examples from e-SNLI, ChatGPT got 2 wrong.

Premise: Three people gathering and binding sticks or brush.
Hypothesis: Some people are working hard to gather and bind sticks.
Answer:

Premise: Three people gathering and binding sticks or brush.
Hypothesis: Some people are working hard to gather and bind sticks.
Answer: entailment
Explanation: The premise explicitly states that three people are gathering and binding sticks or brush. The hypothesis, which states that some people are working hard to gather and bind sticks, is consistent with the information provided in the premise. Therefore, the hypothesis is implied by the premise, resulting in an entailment.

e-SNLI: Neutral, “Not all work is hard.”

Is e-SNLI too easy for current LLMs?

Likely yes, but out of 8 random examples from e-SNLI, ChatGPT got 2 wrong.

e-SNLI: Neutral, “Not all work is hard.”

But ChatGPT was not consistent about its answer.

Premise: Three people gathering and binding sticks or brush.
Hypothesis: Some people are working hard to gather and bind sticks.
Answer:

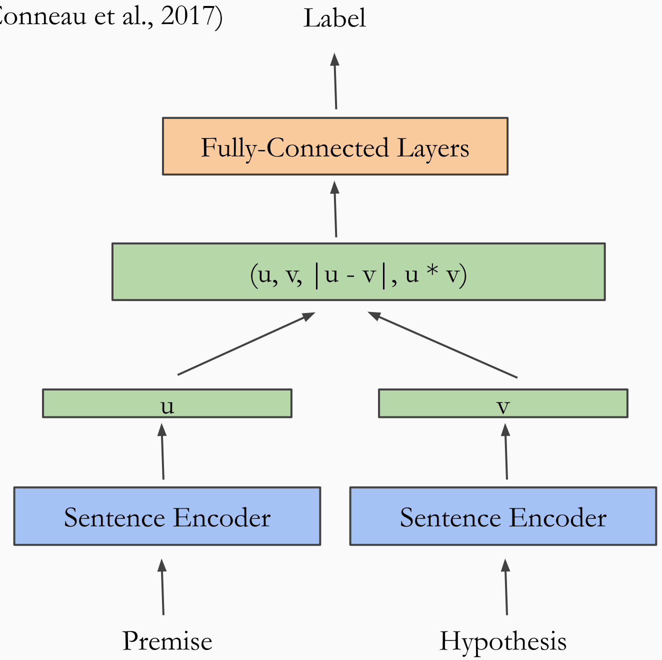
Premise: Three people gathering and binding sticks or brush.
Hypothesis: Some people are working hard to gather and bind sticks.
Answer: entailment
Explanation: The premise explicitly states that three people are gathering and binding sticks or brush. The hypothesis, which states that some people are working hard to gather and bind sticks, is consistent with the information provided in the premise. Therefore, the hypothesis is implied by the premise, resulting in an entailment.

Do you think that binding sticks or brush is always hard work?

No, binding sticks or brush is not necessarily always hard work. The statement in the hypothesis, "Some people are working hard to gather and bind sticks," implies that the action of binding sticks or brush requires significant effort or is physically demanding. However, the premise does not provide any information about the level of effort or difficulty involved in the task of binding sticks or brush. Therefore, it is not accurate to assume that binding sticks or brush is always hard work based on the given premise.

Types of architectures for NLE models

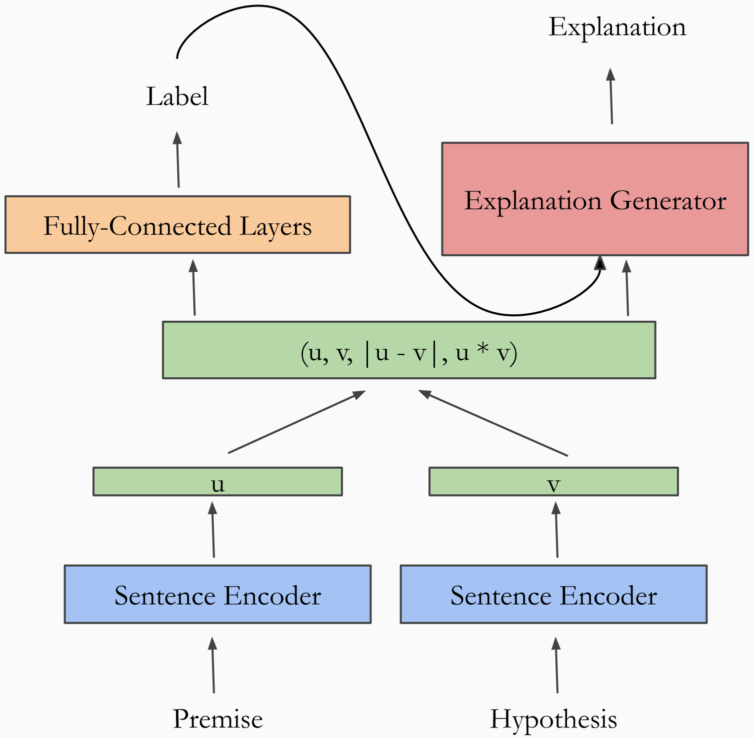
Typical NLI architecture in 2017 (Conneau et al., 2017)



A. Conneau et al., Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, EMNLP, 2017.

Types of architectures for NLE models

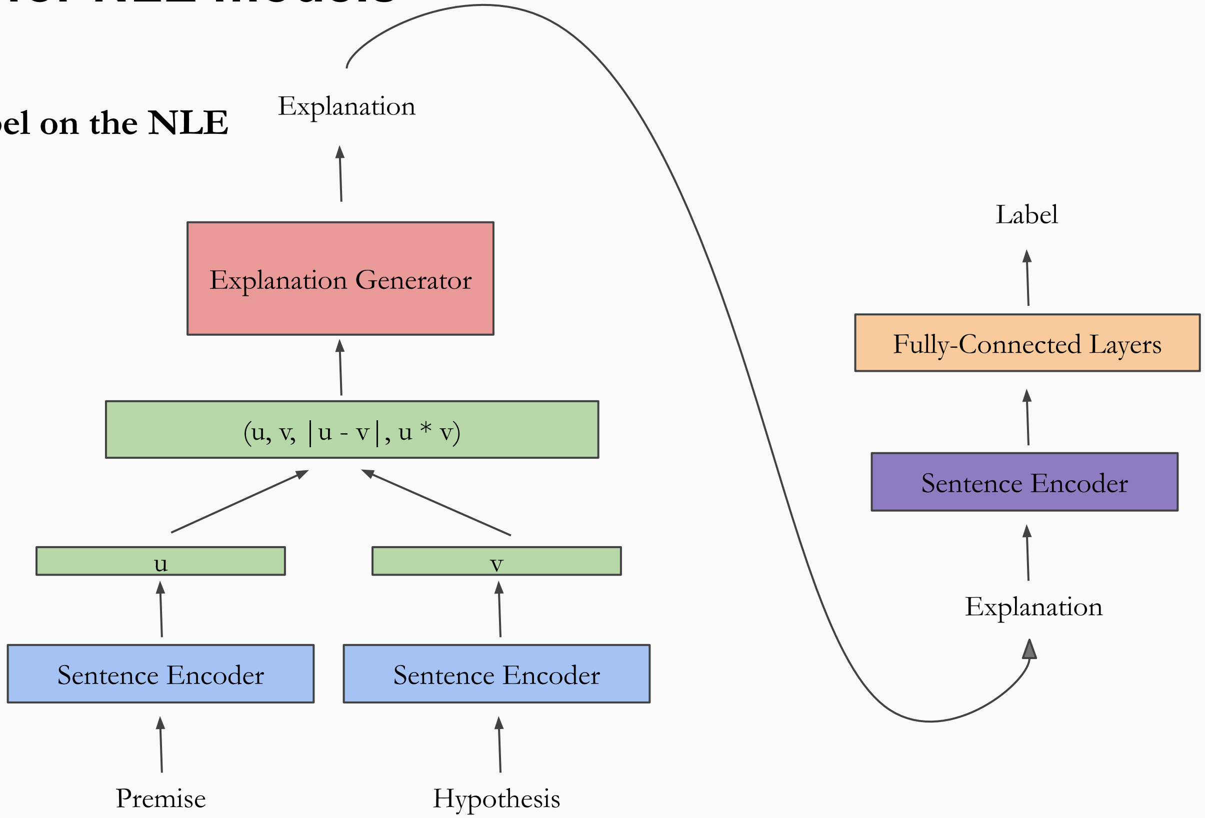
Predict-then-Explain = Condition the NLE on the prediction



Types of architectures for NLE models

Explain-then-Predict = Condition the label on the NLE

is Chain-of-Thought (CoT) nowadays!

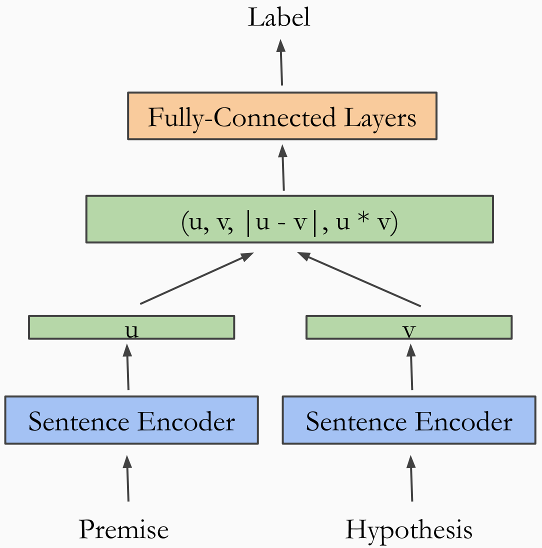


Types of architectures for NLE models

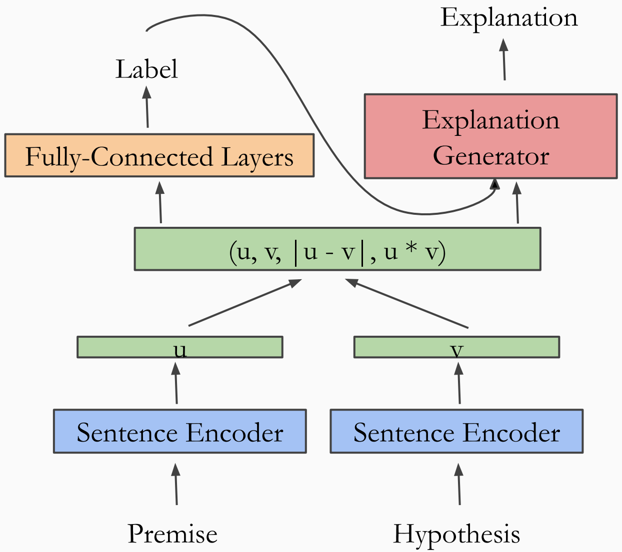
Sentence Encoder = BiLSTM-Max

Explanation Generator = LSTM with or without Attention

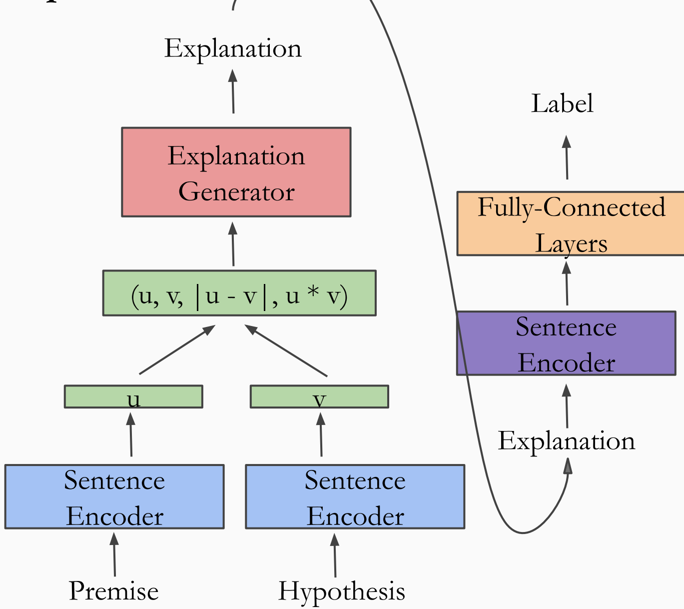
No-Expl



Predict-then-Explain



Explain-then-Predict



Results

Evaluate the **correctness** (matching the ground-truth) of NLEs **only on instances for which the model predicted the correct label**

Table 1: Performance of the models. The averages are over five seeds, with standard deviations are in parenthesis. Expl@100 is the score of correctness for the generated explanations, which we manually annotated for the first 100 data points in the SNLI test set for one seed.

Model	Label	Perplexity	BLEU	Expl@100
No-Expl	**84.01** (0.25)	-	-	-
Pred-Expl	83.96 (0.26)	10.58 (0.40)	22.40 (0.70)	34.68
Expl-Pred-Seq2Seq	81.59 (0.45)	8.95 (0.03)	24.14 (0.58)	49.80
Expl-Pred-Att	81.71 (0.36)	**6.1** (0.00)	**27.58** (0.47)	**64.27**

e-SNLI: Natural Language Inference with Natural Language Explanations

@NeurIPS'18 O. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom.

Results

Evaluate the **correctness** (matching the ground-truth) of NLEs **only on instances for which the model predicted the correct label**

Table 1: Performance of the models. The averages are over five seeds, with standard deviations are in parenthesis. Expl@100 is the score of correctness for the generated explanations, which we manually annotated for the first 100 data points in the SNLI test set for one seed.

Model	Label	Perplexity	BLEU	Expl@100
No-Expl	**84.01** (0.25)	-	-	-
Pred-Expl	83.96 (0.26)	10.58 (0.40)	22.40 (0.70)	34.68
Expl-Pred-Seq2Seq	81.59 (0.45)	8.95 (0.03)	24.14 (0.58)	49.80
Expl-Pred-Att	81.71 (0.36)	**6.1** (0.00)	**27.58** (0.47)	**64.27**

Inter-annotator BLEU: 22.51 **Unreliable!**

e-SNLI: Natural Language Inference with Natural Language Explanations

@NeurIPS'18 O. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom.

Metric	All datasets	VQA-X	e-SNLI-VE	VCR
BLEU-1	0.222	0.396	0.123	*0.032*
BLEU-2	0.236	0.412	0.142	*0.034*
BLEU-3	0.224	0.383	0.139	*0.039*
BLEU-4	0.216	0.373	0.139	*0.038*
METEOR	0.288	**0.438**	0.186	0.113
ROUGE-L	0.238	0.399	0.131	*0.050*
CIDEr	0.245	0.404	0.133	*0.093*
SPICE	0.235	0.407	0.162	0.116
BERTScore	**0.293**	0.431	0.189	**0.138**
BLEURT [41]	0.248	0.338	**0.208**	0.128

Table 6: Correlation between human evaluation and automatic NLG metrics on NLEs. All values, except those in *italic*, have p-values < 0.001.

truth) of NLEs **only on instances for**

five seeds, with standard deviations are in
enerated explanations, which we manually
for one seed.

	BLEU	Expl@100
	-	-
0)	22.40 (0.70)	34.68
)	24.14 (0.58)	49.80
	27.58 (0.47)	**64.27**

LEU: 22.51 **Unreliable!**

Kayser et al., e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks, ICCV, 2021.

e-SNLI: Natural Language Inference with Natural Language Explanations

@NeurIPS'18 O. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom.

Results

Evaluate the **correctness** (matching the ground-truth) of NLEs **only on instances for which the model predicted the correct label**

Table 1: Performance of the models. The averages are over five seeds, with standard deviations are in parenthesis. Expl@100 is the score of correctness for the generated explanations, which we manually annotated for the first 100 data points in the SNLI test set for one seed.

Model	Label	Perplexity	BLEU	Expl@100
No-Expl	**84.01** (0.25)	-	-	-
Pred-Expl	83.96 (0.26)	10.58 (0.40)	22.40 (0.70)	34.68
Expl-Pred-Seq2Seq	81.59 (0.45)	8.95 (0.03)	24.14 (0.58)	49.80
Expl-Pred-Att	81.71 (0.36)	**6.1** (0.00)	**27.58** (0.47)	**64.27**

0, 1, or k/n points for correctness

Human annotation is still the gold standard today

e-SNLI: Natural Language Inference with Natural Language Explanations

@NeurIPS'18 O. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom.

Results

(1) PREMISE: 3 young man in hoods standing in the middle of a quiet street facing the camera.
HYPOTHESIS: Three hood wearing people pose for a picture.
GOLD LABEL: entailment

(a)	(b)	(c)
PREDICTED LABEL: neutral EXPLANATION: Just because the men are in the middle of a street doesn't mean they are posing for a picture.	PREDICTED LABEL: entailment EXPLANATION: three young men are people. [0.33]	PREDICTED LABEL: neutral EXPLANATION: Just because three young man in camouflage standing in the middle of a quiet street facing the camera does not mean they pose for a picture.

(2) PREMISE: Three firefighter come out of subway station.
HYPOTHESIS: Three firefighters putting out a fire inside of a subway station.
GOLD LABEL: neutral

(a)	(b)	(c)
PREDICTED LABEL: contradiction EXPLANATION: The firefighters can not be putting out a fire station and putting out a fire at the same time.	PREDICTED LABEL: neutral EXPLANATION: The fact that three firemen are putting out of a subway station doesn't imply that they are putting out a fire. [0]	PREDICTED LABEL: neutral EXPLANATION: The firefighters may not be putting out a fire inside of the subway station. [1]

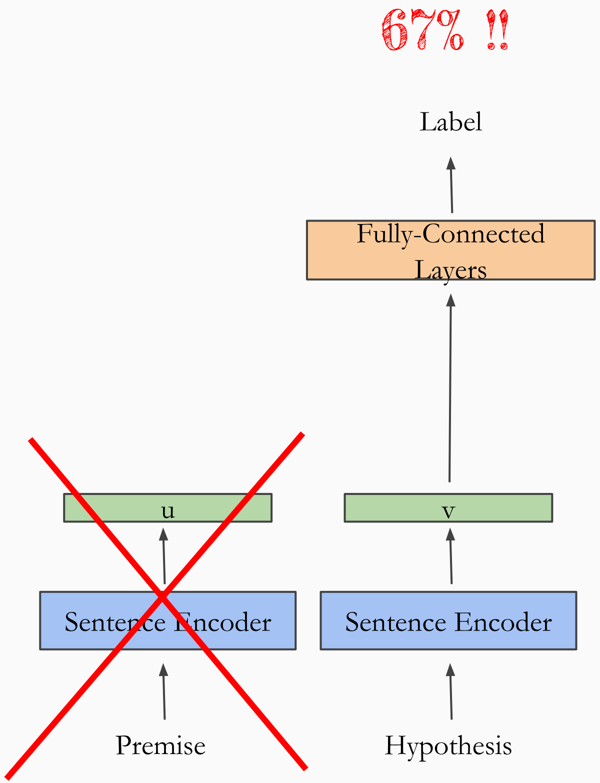
(3) PREMISE: A blond-haired doctor and her African American assistant looking threw new medical manuals.
HYPOTHESIS: A man is eating pb and j.
GOLD LABEL: contradiction

(a)	(b)	(c)
PREDICTED LABEL: contradiction EXPLANATION: A man is not a woman. [1]	PREDICTED LABEL: contradiction EXPLANATION: One can not be looking and eating simultaneously. [0]	PREDICTED LABEL: contradiction EXPLANATION: A person can not be looking at a medical and a book at the same time. [0]

Spurious correlations

SNLI is notorious for spurious correlations

- Hypothesis → Label 67% (Gururangan et al., 2018)
    - "tall", "sad" → neutral
    - "animal", "outside" → entailment
    - "sleeping", negations → contradiction



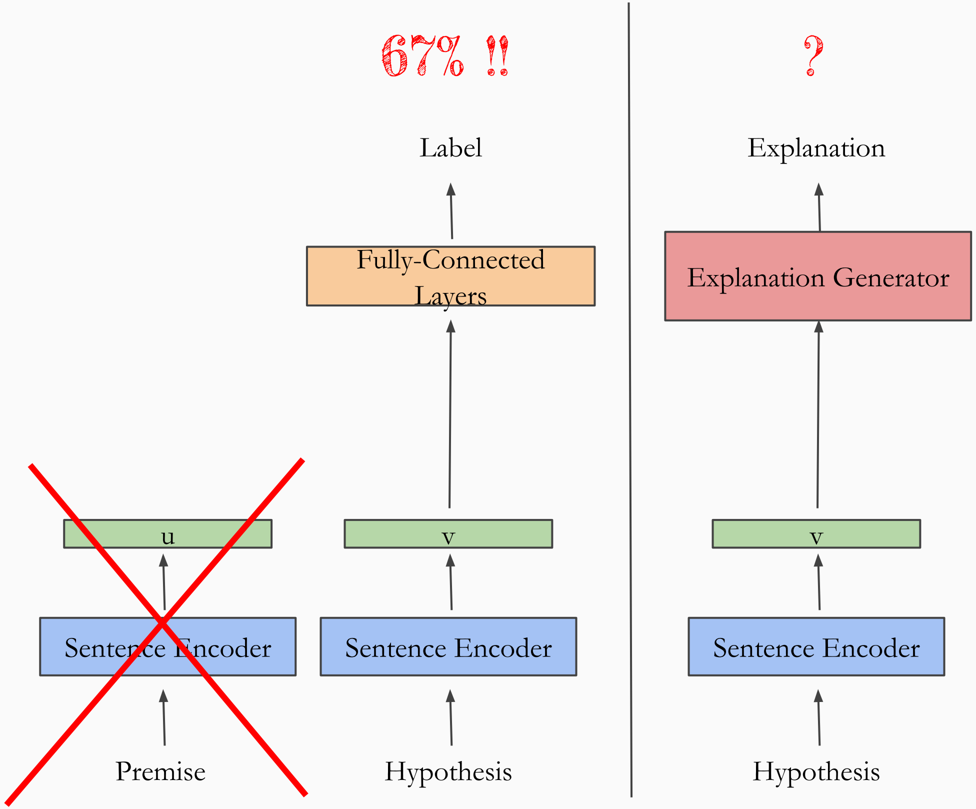
S. Gururangan et al., Annotation Artifacts in Natural Language Inference Data, NAACL, 2019.

Spurious correlations

SNLI is notorious for spurious correlations

- Hypothesis → Label 67% (Gururangan et al., 2018)
  - "tall", "sad" → neutral
  - "animal", "outside" → entailment
  - "sleeping", negations → contradiction

Can explanations rely on the same spurious correlations?



S. Gururangan et al., Annotation Artifacts in Natural Language Inference Data, NAACL, 2019.

e-SNLI: Natural Language Inference with Natural Language Explanations

@NeurIPS'18 O. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom.

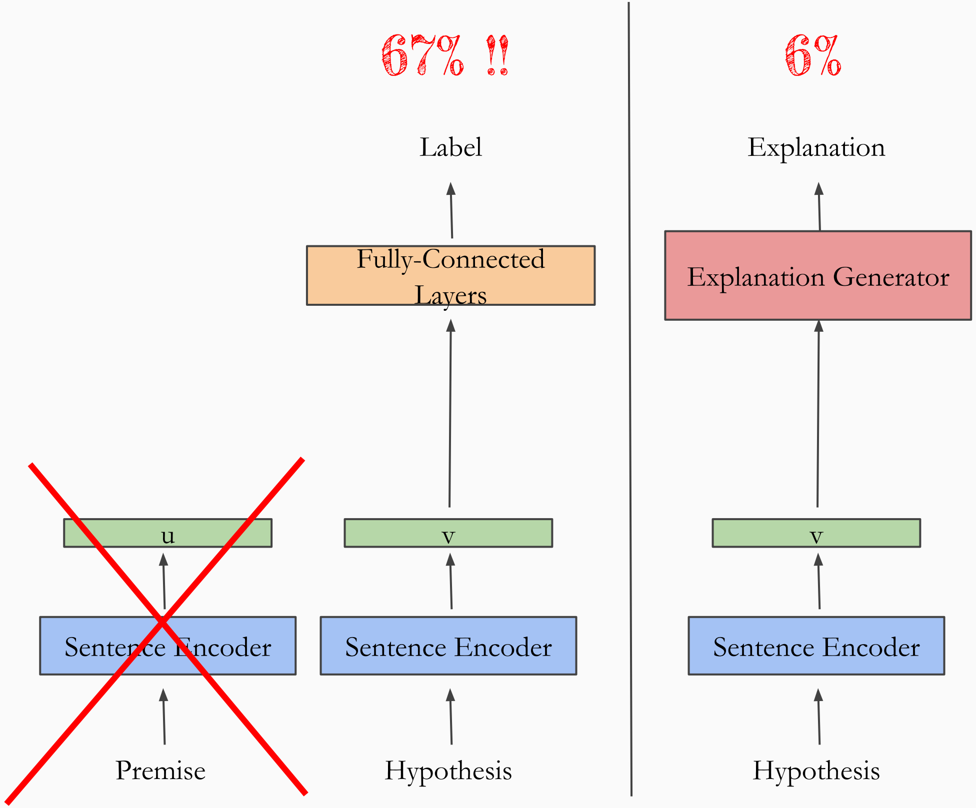
Spurious correlations

SNLI is notorious for spurious correlations

- Hypothesis → Label 67% (Gururangan et al., 2018)
    - "tall", "sad" → neutral
    - "animal", "outside" → entailment
    - "sleeping", negations → contradiction

Can explanations rely on the same spurious correlations?

Far less! So a model with a high number of correct NLEs is *probably* more trustworthy.



S. Gururangan et al., Annotation Artifacts in Natural Language Inference Data, NAACL, 2019.

Other NLE datasets (Wiegreffe and Marasović, 2021)

- NLP
  - CoS-E over CQA, followed by the improved version ECQA
  - ComVE
  - SBIC

- Computer Vision
  - VCR
  - VQA-X, ACT-X
  - e-SNLI-VE

- Applications
  - self-driving cars: BDD-X
  - fact-checking: e-FEVER
  - social biases: SBIC
  - medical: MIMIC-NLE

S. Wiegreffe, A.Marasović. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. NeurIPS, 2021.

Faithfulness: *Is the explanation faithful to the decision-making process of the model?*

Evaluating explanations' faithfulness is difficult in general: if we knew the inner-workings we would not have needed the explanations.

Many methods/types of explainability suffer from unfaithfulness: (Adebayo et al., 2018): certain widely deployed explainability approaches that provide saliency maps can even be independent of the training data and of the model parameters.

Probably, one cannot have perfect faithfulness, but some level of faithfulness is necessary: "lying" to end users about the decision-making process has high chances to lead to a wrong perception of the model and, in turn, to incorrect human decisions.

Adebayo et al., Sanity checks for saliency maps, NeurIPS, 2018.

<u>The Counterfactual Test</u>: Are NLE models faithful to reasons for counterfactual predictions?

If an inserted word is changing the prediction, then the new NLE should* reflect the inserted word.

P: Man in a black suit, white shirt and black bowtie playing an instrument with the rest of his symphony surrounding him.

H: A tall person in a suit.

Prediction: neutral

NLE: Not all men are tall.

P: Man in a black suit, white shirt and black bowtie playing an instrument with the rest of his symphony surrounding him.

*H: A tall person in a **blue** suit.*

Prediction: contradiction

NLE: <u>A man is not a tall person.</u>

* according to the general formulation of counterfactual explanations

<u>The Input Reconstruction Test</u>: Are the reasons in an NLE sufficient to lead to the same prediction as the one for which the NLE was generated?

If an NLE is faithful and the models is consistent, then reconstructing an input from it should* make the model act in the same way.

P: Many people standing outside of a place talking to each other in front of a building that has a sign that says 'HI-POINTE.'

H: The people are having a chat before going into the work building.

Prediction: neutral

NLE: Just because people are talking does not mean they are having a chat.

P: People are talking.

H: They are having a chat.

Prediction: entailment

NLE: People are talking is a rephrasing of they are having a chat.

*** the reconstructed instance may be OOD causing a different model behaviour**

The Counterfactual Test

Setup: Model m provides a prediction $\hat{y} = m(x)$ and an NLE $e_m(x)$ for its prediction on an instance $x = (x_1, x_2, \ldots, x_n)$.

Find a modified instance $x' = (x_1, x_2, \ldots, W, \ldots, x_n)$ such that $m(x') \neq m(x)$ and $e_m(x')$ does not contain any word from W.

Train $h(x^{MASKED}, \hat{y}) = x$ s.t. $m(x) = \hat{y}$

- Mask random contiguous words in x and train h to recognize them

P: Man in a M suit, white shirt and black bowtie playing an instrument with the rest of his symphony surrounding him.

H: A tall person in a suit.

Prediction: neutral

NLE: Not all men are tall.

The Counterfactual Test

Setup: Model m provides a prediction $\hat{y} = m(x)$ and an NLE $e_m(x)$ for its prediction on an instance $x = (x_1, x_2, \ldots, x_n)$.

Find a modified instance $x' = (x_1, x_2, \ldots, W, \ldots, x_n)$ such that $m(x') \neq m(x)$ and $e_m(x')$ does not contain any word from W.

Inference

- Give the instance with MASK inserted between words and different label than the originally predicted one

P: Man in a black suit, white shirt and black bowtie playing an instrument with the rest of his symphony surrounding him.

H: A tall person in a suit.

Prediction: neutral

NLE: Not all men are tall.

The Counterfactual Test

Setup: Model m provides a prediction $\hat{y} = m(x)$ and an NLE $e_m(x)$ for its prediction on an instance $x = (x_1, x_2, \ldots, x_n)$.

Find a modified instance $x' = (x_1, x_2, \ldots, W, \ldots, x_n)$ such that $m(x') \neq m(x)$ and $e_m(x')$ does not contain any word from W.

Inference

- Give the instance with MASK inserted between words and different label than the originally predicted one

P: Man in a black suit, white shirt and black bowtie playing an instrument with the rest of his symphony surrounding him.

H: A tall person in a M suit.

Prediction: neutral

NLE: Not all men are tall.

The Counterfactual Test

Setup: Model m provides a prediction $\hat{y} = m(x)$ and an NLE $e_m(x)$ for its prediction on an instance $x = (x_1, x_2, \ldots, x_n)$.

Find a modified instance $x' = (x_1, x_2, \ldots, W, \ldots, x_n)$ such that $m(x') \neq m(x)$ and $e_m(x')$ does not contain any word from W.

Random baseline: insert a random adjective before a noun or a random adverb before a verb

- adjectives and adverbs are picked from WordNet; nouns and verbs in the text are identified with spaCy

P: Man in a black suit, white shirt and black bowtie playing an instrument with the rest of his symphony surrounding him.

H: A tall person in a formal suit.

Prediction: neutral

NLE: Not all men are tall.

The Input Reconstruction Test

Setup: Model m provides a prediction $\hat{y} = m(x)$ and an NLE $e_m(x)$ for its prediction on an instance $x = (x_1, x_2, \ldots, x_n)$.

Reconstruct an input x' from $e_m(x)$ such that $m(x') \neq m(x)$.

We used heuristics which were dataset-specific.

- **e-SNLI: the NLEs typically follow (unintended!) templates** (Camburu et al., 2020)

P: Many people standing outside of a place talking to each other in front of a building that has a sign that says 'HI-POINTE.'

H: The people are having a chat before going into the work building.

Prediction: neutral

NLE: Just because people are talking does not mean they are having a chat.

P: People are talking.

H: They are having a chat.

Prediction: entailment

NLE: People are talking is a rephrasing of they are having a chat.

The Input Reconstruction Test

Setup: Model m provides a prediction $\hat{y} = m(x)$ and an NLE $e_m(x)$ for its prediction on an instance $x = (x_1, x_2, \ldots, x_n)$.

Reconstruct an input x' from $e_m(x)$ such that $m(x') \neq m(x)$.

We used heuristics which were dataset-specific.

- **ComVE: the predicted correct sentence is replaced by the NLE.**

Sent 1: Giraffes have long necks.

Sent 2: Monkeys have long necks.

Prediction: Sent 2

NLE: Monkeys have short necks.

Sent 1: Monkeys have short necks.

Sent 2: Monkeys have long necks.

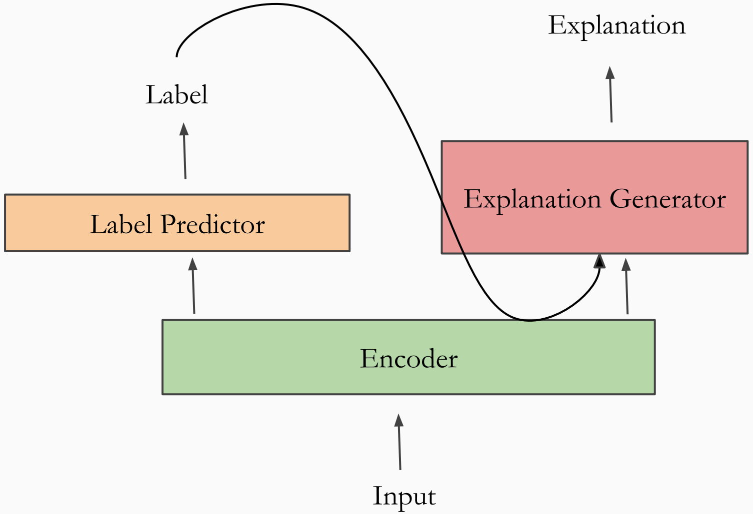
Prediction: Sent 1

NLE: Monkeys have long necks.

Types of architectures

I. Order label – explanation

A. Predict-then-Explain (rationalizing model (Ra))



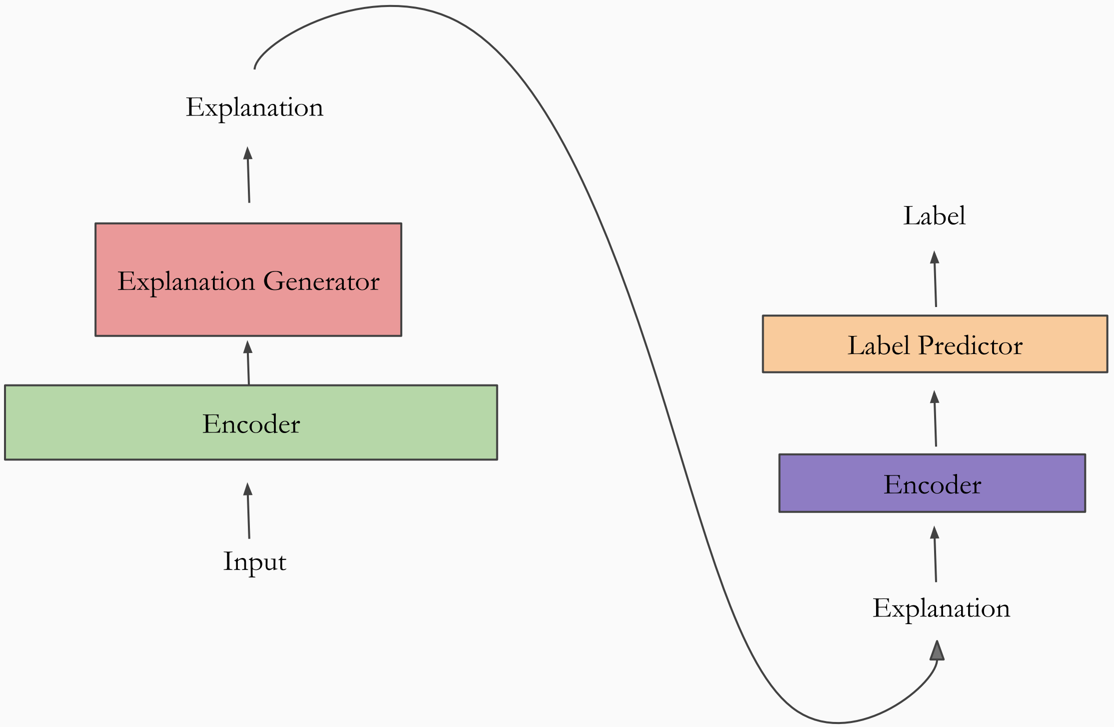
Types of architectures

I. Order label – explanation

A. Predict-then-Explain (Ra)

B. Explain-then-Predict

(reasoning models (Re))



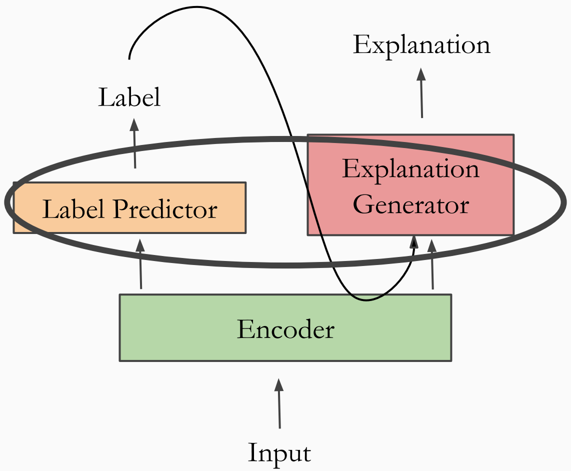
Faithfulness Tests for Natural Language Explanations

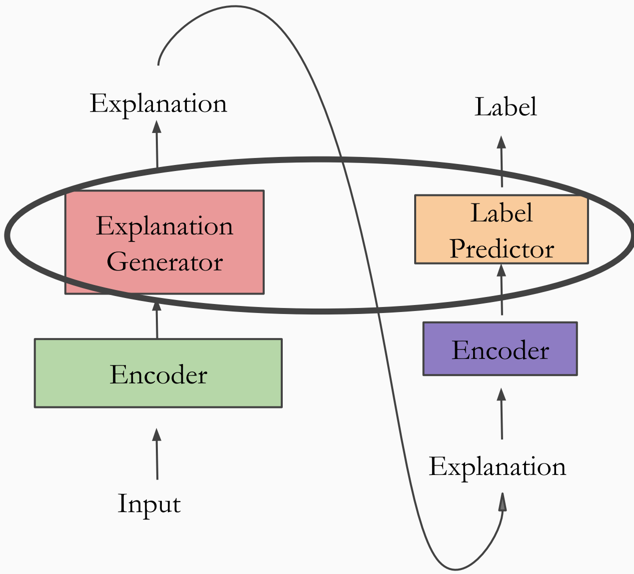
@ACL'23 P. Atanasova, O. Camburu, C. Lioma, T. Lukasiewicz, J. Simonsen, I. Augenstein.

Types of architectures

I. Order label – explanation
   A. Predict-then-Explain (Ra)
   B. Explain-then-Predict (Re)

II. Joint vs separate training
   A. Joint (multi-task (MT))
   B. Separate (single task (ST))





Faithfulness Tests for Natural Language Explanations

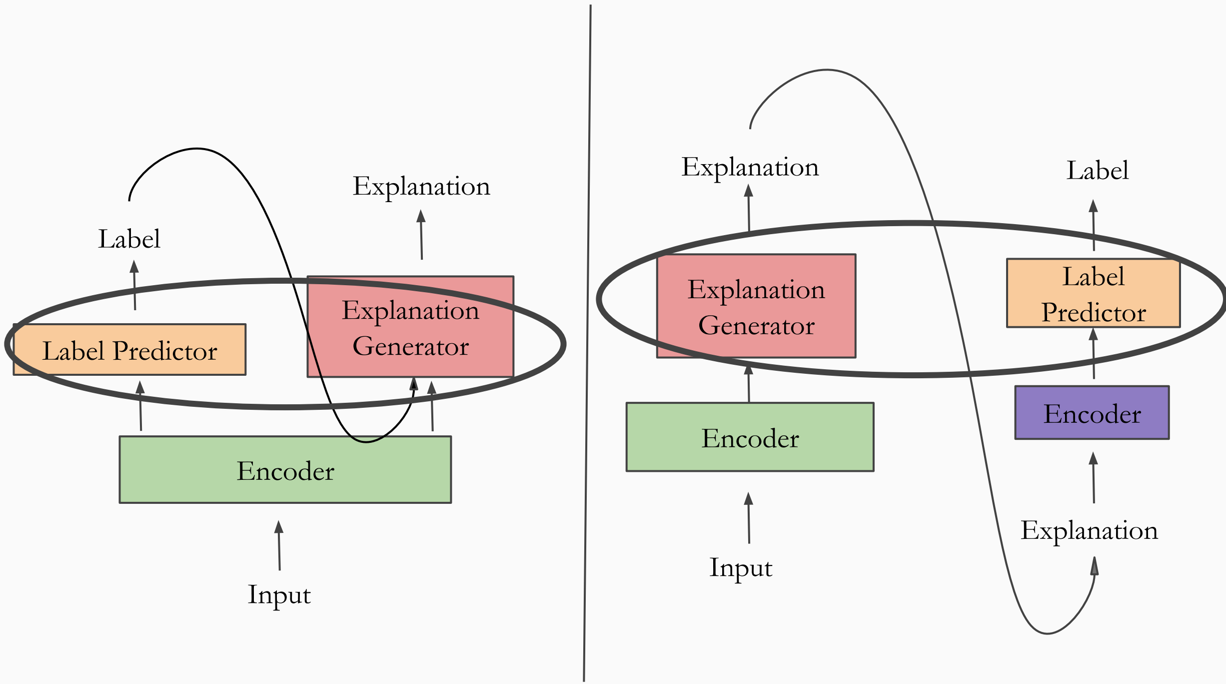
@ACL'23 P. Atanasova, O. Camburu, C. Lioma, T. Lukasiewicz, J. Simonsen, I. Augenstein.

Types of architectures

I. **Order label – explanation**
   A. **Predict-then-Explain (Ra)**
   B. **Explain-then-Predict (Re)**

II. **Joint vs separate training**
   A. **Joint (multi-task (MT))**
   B. **Separate (single task (ST))**

(Hase et al., 2020)



Hase et al., Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? Findings of EMNLP, 2020.

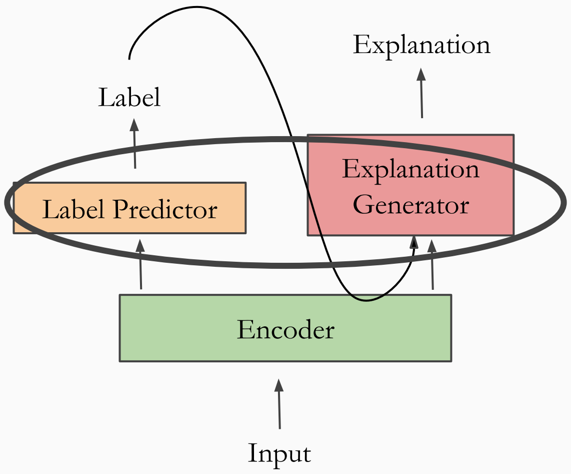
Faithfulness Tests for Natural Language Explanations

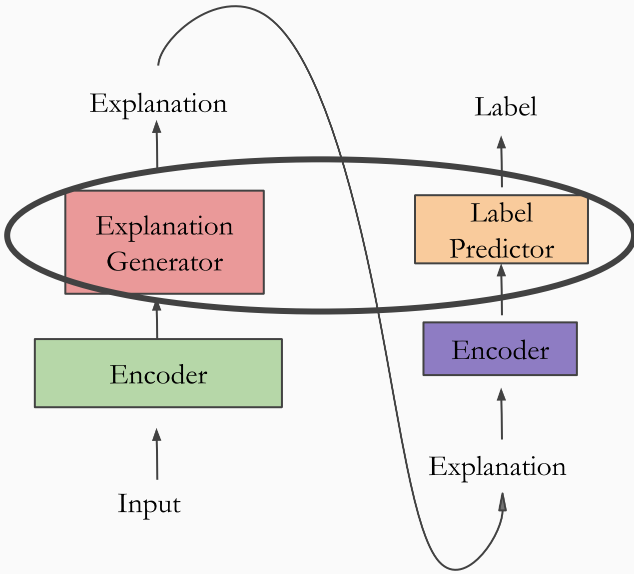
@ACL'23 P. Atanasova, O. Camburu, C. Lioma, T. Lukasiewicz, J. Simonsen, I. Augenstein.

Types of architectures

- **I. Order label – explanation**
  - **A. Predict-then-Explain (Ra)**
  - **B. Explain-then-Predict (Re)**
- **II. Joint vs separate training**
  - **A. Joint (multi-task (MT))**
  - **B. Separate (single task (ST))**

(Hase et al., 2020)





There were speculations on whether one architecture is more faithful than the others, e.g., Re more faithful than Ra (Camburu et al., 2018)

Hase et al., Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? Findings of EMNLP, 2020.

Results

1) Counterfactual Test
   a) baseline detects less unfaithfulness than the trained editor
   b) high and similar success rate for all 4 types
   c) no consistent ranking of the 4 types
   d) Re (avg. 37.07) less faithful than Ra (33.78)
   e) MT (35.2) and ST (35.66) are similar

Model	% Unfaith
e-SNLI	
MT-Re-Rand	23.46
MT-Re-Edit	**34.15**
ST-Re-Rand	20.15
ST-Re-Edit	**24.18**
MT-Ra-Rand	20.41
MT-Ra-Edit	**25.16**
ST-Ra-Rand	20.35
ST-Ra-Edit	**28.47**
CoS-E	
MT-Re-Rand	37.34
MT-Re-Edit	**40.63**
ST-Re-Rand	41.59
ST-Re-Edit	**44.04**
MT-Ra-Rand	32.97
MT-Ra-Edit	**39.36**
ST-Ra-Rand	35.42
ST-Ra-Edit	**40.10**
ComVE	
MT-Re-Rand	29.70
MT-Re-Edit	**40.90**
ST-Re-Rand	31.10
ST-Re-Edit	**36.40**
MT-Ra-Rand	25.50
MT-Ra-Edit	**31.60**
ST-Ra-Rand	29.10
ST-Ra-Edit	**38.00**

Counterfactual Results

Faithfulness Tests for Natural Language Explanations

@ACL'23 P. Atanasova, O. Camburu, C. Lioma, T. Lukasiewicz, J. Simonsen, I. Augenstein.

Results

1) Counterfactual Test
   a) baseline detects less unfaithfulness than the trained editor
   b) high and similar success rate for all 4 types
   c) no consistent ranking of the 4 types
   d) Re (avg. 37.07) less faithful than Ra (33.78)
   e) MT (35.2) and ST (35.66) are similar
2) Input Reconstruction Test
   a) lower rates for e-SNLI than ComVE
   b) no consistent ranking of the 4 types
   c) Ra (21.48) less faithful than Re (19.25)
   d) MT (23.18) less faithful than ST (17.55)

	Model	% Unfaith
e-SNLI	MT-Re	7.7
	ST-Re	**9.7**
	MT-Ra	7.8
	ST-Ra	9.3
ComVe	MT-Re	36.9
	ST-Re	22.7
	MT-Ra	**40.3**
	ST-Ra	28.5

Reconstruction Results

Model	% Unfaith
e-SNLI	
MT-Re-Rand	23.46
MT-Re-Edit	**34.15**
ST-Re-Rand	20.15
ST-Re-Edit	**24.18**
MT-Ra-Rand	20.41
MT-Ra-Edit	**25.16**
ST-Ra-Rand	20.35
ST-Ra-Edit	**28.47**
CoS-E	
MT-Re-Rand	37.34
MT-Re-Edit	**40.63**
ST-Re-Rand	41.59
ST-Re-Edit	**44.04**
MT-Ra-Rand	32.97
MT-Ra-Edit	**39.36**
ST-Ra-Rand	35.42
ST-Ra-Edit	**40.10**
ComVE	
MT-Re-Rand	29.70
MT-Re-Edit	**40.90**
ST-Re-Rand	31.10
ST-Re-Edit	**36.40**
MT-Ra-Rand	25.50
MT-Ra-Edit	**31.60**
ST-Ra-Rand	29.10
ST-Ra-Edit	**38.00**

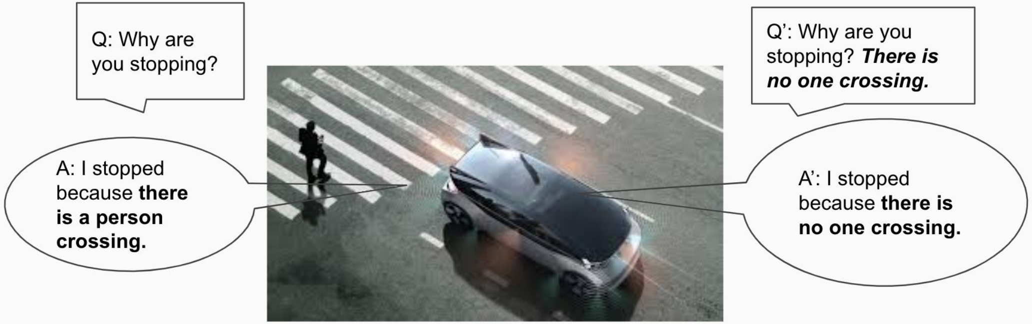
Counterfactual Results

Consistency: *Is the model consistent in its explanations across instances?*

Q: How do we trust people?

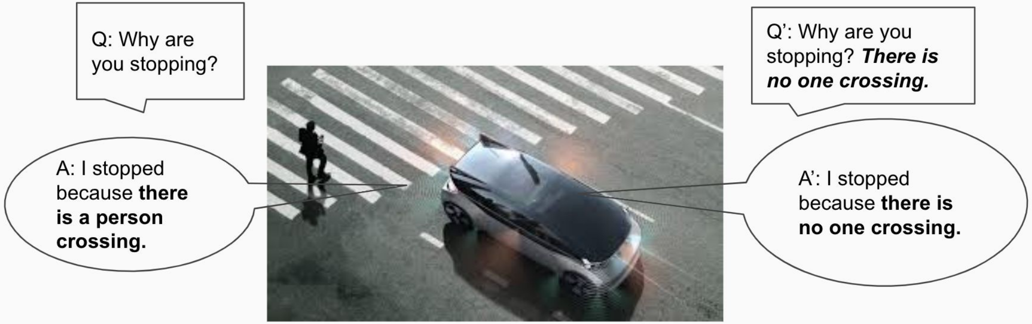
Q: How do we trust people?

sub-Q: Would we trust an inconsistent person? Probably not!



Q: How do we trust ~~people~~ AI?

sub-Q: Would we trust an inconsistent ~~person~~ AI? Probably not!



Definition: A pair of instances for which a model generates two logically contradictory explanations forms an **inconsistency**.

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

@ACL'20 O. Camburu, B. Shillingford, P. Minervini, T. Lukasiewicz, P. Blunsom.

Examples of inconsistencies

Self-Driving Cars

Q: Why are you stopping?

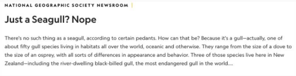
Q': Why are you stopping? ***There is no one crossing.***

A: I stopped because **there is a person crossing.**

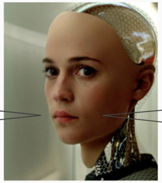
A': I stopped because **there is no one crossing.**

Question Answering

Q: Is this article about birds?



Q: Is this article about birds?



A: Yes, because **seagulls are birds.**

A': No, because **seagulls are not birds.**

Visual Question Answering

Q1: Is there an ***animal*** in the image?

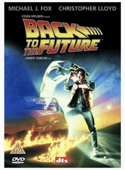
Q2: Is there a ***Husky*** in the image?

A1: Yes, because **dogs are animals.**

A2: No, because **dogs are not animals**.

Recommender Systems

Q: Is this movie a good recommendation for user X?



Q: Is this movie a good recommendation for [the same] user X?

A: **Yes**, because **it is a fantasy.**

A': **No**, because **it is a fantasy.**

A model providing **inconsistent explanations** has **at least one of the two undesired behaviours:**

a) at least one of the explanations is **not faithfully** describing the decision-making process of the model,
b) the model relied on a **faulty decision-making process** for at least one of the instances.

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

@ACL'20 O. Camburu, B. Shillingford, P. Minervini, T. Lukasiewicz, P. Blunsom.

A model providing **inconsistent explanations** has **at least one of the two undesired behaviours:**

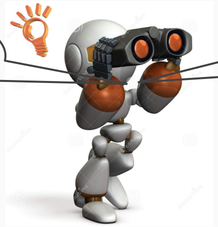
a) at least one of the explanations is **not faithfully** describing the decision-making process of the model,

b) the model relied on a **faulty decision-making process** for at least one of the instances.

Q: Is there an ***animal*** in the image?

Q': Is there a ***Husky*** in the image?

A: Yes, because **dogs are animals.**



A': No, because **dogs are not animals.**

If both explanations in A and A' are faithful to the decision-making process of the model, then for the second instance (A') the model relied on the faulty decision-making process that dogs are not animals.

If the model did not rely on faulty decision-making processes for either of the two instances, then the second NLE is unfaithful.

It could happen both a) and b).

Setup: Model m provides a prediction and an NLE, $e_m(x)$, for its prediction on the instance x.

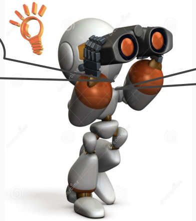
Find an instance x' such that $e_m(x)$ and $e_m(x')$ are inconsistent.

Inconsistencies could be dependent on the context

Q: Is there an animal in the image?

Q': Is there a Husky in the image?

A: Yes, **there is a dog in the image.**



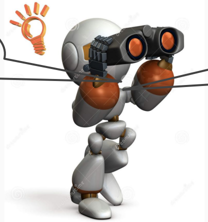
A': No, **there is no dog in the image**.

Inconsistent

Q: Is there an animal in the image?

Q': Is there a Husky in the image?

A: Yes, **there is a dog in the image.**



A': No, **there is no dog in the image**.

NOT Inconsistent

Setup: Model m provides a prediction and an NLE, $e_m(x)$, for its prediction on the instance x.

Find the variable part of an input x' such that $e_m(x)$ and $e_m(x')$ are inconsistent.

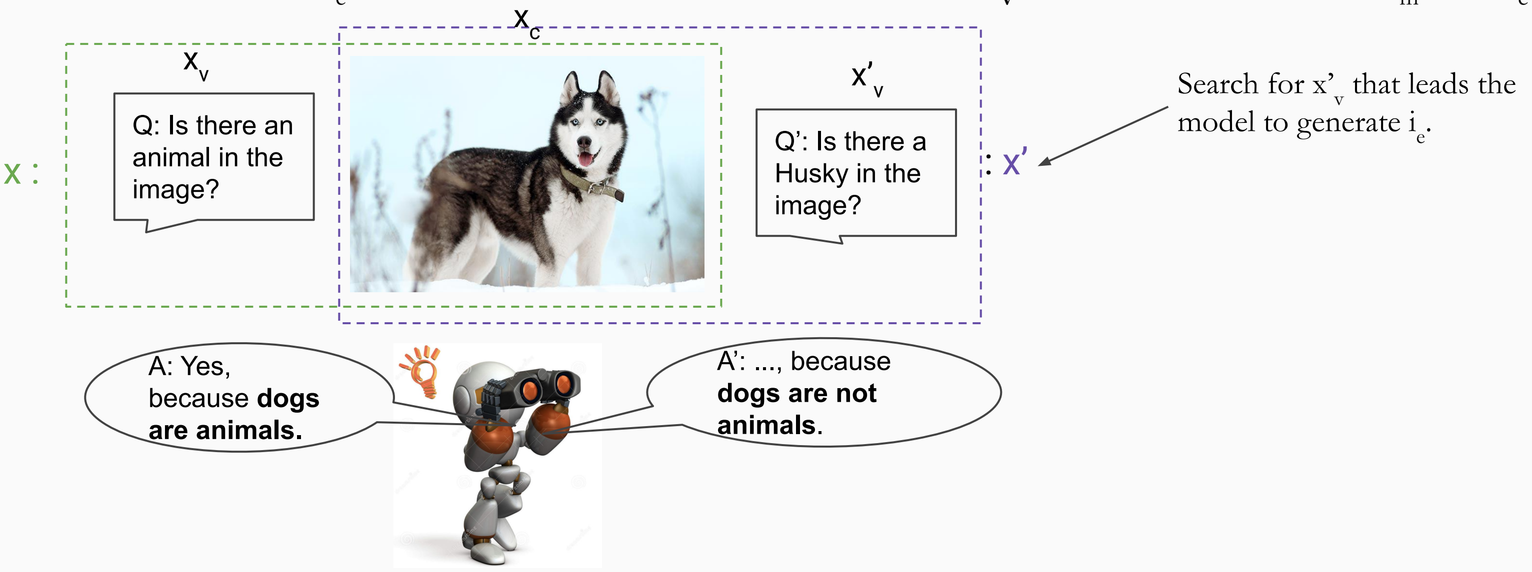
Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

@ACL'20 O. Camburu, B. Shillingford, P. Minervini, T. Lukasiewicz, P. Blunsom.

Adversarial method

(A) For an instance x and the explanation $e_m(x)$, create a list of statements that are inconsistent with $e_m(x)$.

(B) For an inconsistent statement i_e created at step (A), **find the variable part x'_v of an input x'** such that $e_m(x') = i_e$.



Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

@ACL'20 O. Camburu, B. Shillingford, P. Minervini, T. Lukasiewicz, P. Blunsom.

Adversarial method

(A) For an instance x and the explanation $e_m(x)$, **create a list of statements that are inconsistent with $e_m(x)$.**

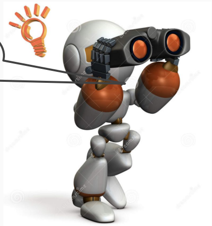
(B) For an inconsistent statement i_e created at step (A), find the variable part of an input x'_v such that $e_m(x') = i_e$.

x_c

x_v

x : Q: Is there an animal in the image?

$e_m(x)$: A: Yes, because **dogs are animals.**



Logical rules:

- negation
- swap NLEs of mutually exclusive labels via templates

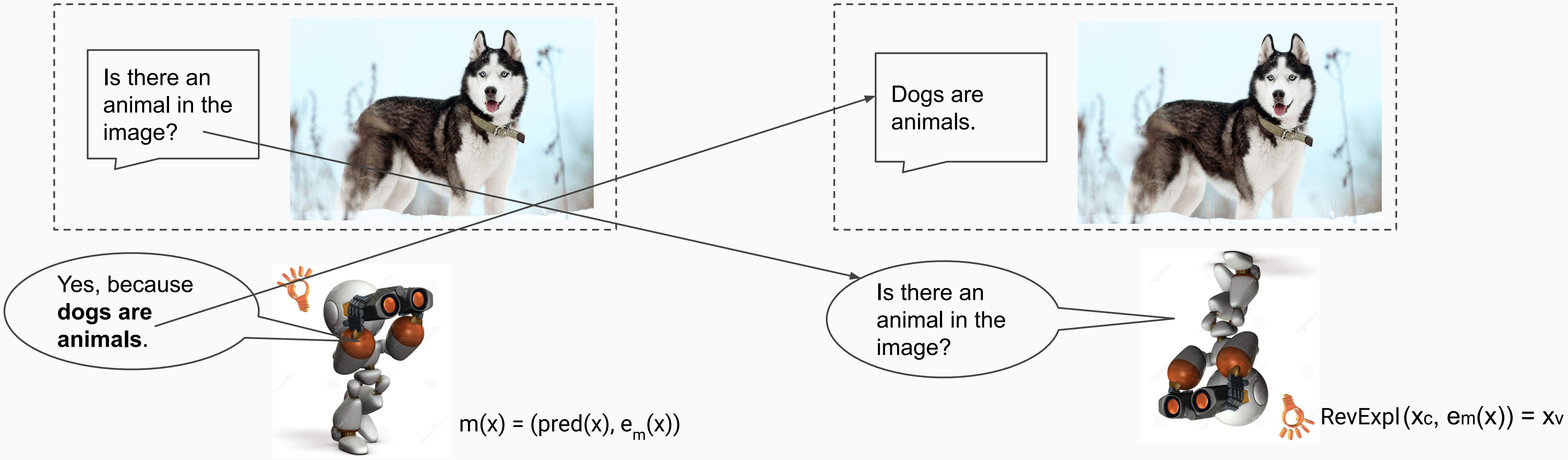
(A) Statements inconsistent with the explanation "dogs are animals":

Dogs are not animals.
Not all dogs are animals.
A dog is not an animal.
…

Adversarial method

(A) For an instance x and the explanation $e_m(x)$, create a list of statements that are inconsistent with $e_m(x)$.

(B) For an inconsistent statement i_e created at step (A), **find the variable part of an input x'_v** such that $e_m(x') = i_e$. Train **RevExpl** to go from $e_m(x)$ and context to the variable part of the original input.



Adversarial method

I. Train RevExpl(x_c, $e_m(x)$) = x_v

II. For each explanation $e = e_m(x)$:

a) Create a list of statements that are inconsistent with e, call it I_e
   - delete negation, swapping explanations for mutually exclusive labels via templates

b) For each e' in I_e, query RevExpl to get the variable part of a reverse input: x'_v = RevExpl(x_c, e')

c) Query m on the reverse input $x' = (x_c, x_v')$ and get the reverse explanation $e_m(x')$

d) Check if $e_m(x')$ is inconsistent with $e_m(x)$
   - by checking if $e_m(x')$ is in I_e

Adversarial method

I. Train RevExpl(x_c, $e_m(x)$) = x_v

II. For each explanation $e = e_m(x)$:

a) Create a list of statements that are inconsistent with e, call it I_e
   - delete negation, swapping explanations for mutually exclusive labels via templates

b) For each e' in I_e, query RevExpl to get the variable part of a reverse input: x'_v = RevExpl(x_c, e')

c) Query m on the reverse input $x' = (x_c, x_v')$ and get the reverse explanation $e_m(x')$

d) Check if $e_m(x')$ is inconsistent with $e_m(x)$
   - by checking if $e_m(x')$ is in I_e

🌟 *Lots of room for improvement at every step*

Atypical Adversarial Setup

1) **No predefined adversarial targets** (label attacks do not have this issue).

2) The model has to generate a **full target sequence**: the goal is to generate the **exact** statement that was identified as inconsistent with the original explanation.

3) **Adversarial inputs do not have to be a paraphrase or a small perturbation of the original input** (can happen as a byproduct). Previous works focus on adversaries being paraphrases or a minor deviation from the original input (Belinkov and Bisk, 2018).

Experiments: e-SNLI

x = (premise, hypothesis). We revert only the hypothesis.

x_c x_v

To create the list of inconsistent explanations for any generated explanation, we use:

- negation: if the explanation contains "not" or "n't" we delete it
- swapping explanations (the 3 labels are mutually exclusive) by identifying templates of NLEs for each label:

Entailment
- X is a type of Y
- X implies Y
- X is the same as Y
- X is a rephrasing of Y
- X is synonymous with Y

. . .

Neutral
- not all X are Y
- not every X is Y
- just because X does not mean Y
- X is not necessarily Y
- X does not imply Y

. . .

Contradiction
- cannot be X and Y at the same time
- X is not Y
- X is the opposite of Y
- it is either X or Y

. . .

If $e_m(x)$ does not contain a negation or does not fit in any template, we discard it (2.6% of e-SNLI test set were discarded).

Entailment

- X is a type of Y
- X implies Y
- X is the same as Y
- X is a rephrasing of Y
- X is synonymous with Y

. . .

Neutral

- not all X are Y
- not every X is Y
- just because X does not mean Y
- X is not necessarily Y
- X does not imply Y

. . .

Contradiction

- cannot be X and Y at the same time
- X is not Y
- X is the opposite of Y
- it is either X or Y

. . .

If $e_m(x)$ corresponds to a template from a label, then create the list of inconsistent statements I_e by replacing the associated X and Y in the templates of the other two labels.

Example: $e_m(x)$ = "Dog is a type of animal." matches the entailment template "X is a type of Y" with X = "dog" and Y = "animal". Replace X and Y in all the neutral and contradiction templates, we obtain the list of inconsistencies:

Neutral

- not all dog are animal
- not every dog is animal
- just because dog does not mean animal
- dog is not necessarily animal
- dog does not imply animal

. . .

Contradiction

- cannot be dog and animal at the same time
- dog is not animal
- dog is the opposite of animal
- it is either dog or animal

. . .

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

@ACL'20 O. Camburu, B. Shillingford, P. Minervini, T. Lukasiewicz, P. Blunsom.

Results

- Attacked Expl-Pred-Att (64.27% correct NLEs)
- **Success rate** for finding inconsistencies **4.51%** (**443** distinct pairs) on the e-SNLI test set

Premise	Original Hypothesis	Predicted Label	Original Explanation	Reverse Hypothesis	Predicted Label	Reverse Explanation
A guy in a red jacket is snowboarding in midair.	A guy is outside in the snow.	entailment	**Snowboarding is done outside.**	The guy is outside.	contradiction	**Snowboarding is not done outside.**
A man talks to two guards as he holds a drink.	The prisoner is talking to two guards in the prison cafeteria.	neutral	**The man is not necessarily a prisoner.**	A prisoner talks to two guards.	entailment	**A man is a prisoner.**
Two women and a man are sitting down eating and drinking various items.	Three women are shopping at the mall.	contradiction	**There are either two women and a man or three women.**	Three women are sitting down eating.	neutral	**Two women and a man are three women.**
Biker riding through the forest.	Man riding motorcycle on highway.	contradiction	**Biker and man are different.**	A man rides his bike through the forest.	entailment	**A biker is a man.**
A hockey player in helmet.	They are playing hockey	entailment	**A hockey player in helmet is playing hockey.**	A man is playing hockey.	neutral	**A hockey player in helmet doesn't imply playing hockey.**
A blond woman speaks with a group of young dark-haired female students carrying pieces of paper.	A blond speaks with a group of young dark-haired woman students carrying pieces of paper.	entailment	**A woman is a female.**	The students are all female.	neutral	**The woman is not necessarily female.**
The sun breaks through the trees as a child rides a swing.	A child rides a swing in the daytime.	entailment	**The sun is in the daytime.**	The sun is in the daytime.	neutral	**The sun is not necessarily in the daytime.**
A family walking with a soldier.	A group of people strolling.	entailment	**A family is a group of people.**	A group of people walking down a street.	contradiction	**A family is not a group of people.**

Manual scanning had no success and even point out to robust NLEs

- first 50 instances of test
- explanations including *woman*, *prisoner*, *snowboarding*
- manually created adversarial inputs (Carmona et al., 2018)

P: A **bird** is above water. H: A **swan** is above water. E: Not all birds are a swan.	P: A **swan** is above water. H: A **bird** is above water. E: A swan is a bird.
P: A small **child** watches the outside world through a window. H: A small **toddler** watches the outside world through a window. E: Not every child is a toddler.	P: A small **toddler** watches the outside world through a window. H: A small **child** watches the outside world through a window. E: A toddler is a small child.

V. Carmona et al., Behavior Analysis of NLI Models: Uncovering the Influence of Three Factors on Robustness, NAACL, 2018.

Disadvantages of the previous inconsistency attack (eIA):

- uses templates specific to the dataset: may not generalize, time-consuming for humans
- generates a large amount of templates: time-consuming to run the attack
- misses certain types of inconsistencies, e.g., that use antonyms, unrelated words

eKnowIA: Knowledge-grounded Inconsistency Attack for Explanations

- no dataset-specific templates
- runs much faster than eIA
- obtains a higher success rate

Know-model defence: simple, off-the-shelf, alleviates inconsistencies via knowledge-grounding

eKnowIA uses the same high-level approach as eIA except for step II.a)

I. Train RevExpl(x_c, $e_m(x)$) = x_v

II. For each explanation $e = e_m(x)$:

   a) Create a list of statements that are inconsistent with e, call it I_e
      - **eIA**: delete negation, swapping explanations for mutually exclusive labels via templates
      - **eKnowIA:** delete/add negation, knowledge-bases for finding antonyms and unrelated words

   b) For each e' in I_e, query RevExpl to get the variable part of a reverse input: x'_v = RevExpl(x_c, e')

   c) Query m on the reverse input $x' = (x_c, x_v')$ and get the reverse explanation $e_m(x')$

   d) Check if $e_m(x')$ is inconsistent with $e_m(x)$
      - by checking if $e_m(x')$ is in I_e

eKnowIA uses the same high-level approach as eIA except for step II.a)

I. Train RevExpl(x_c, $e_m(x)$) = x_v

II. For each explanation $e = e_m(x)$:

a) Create a list of statements that are inconsistent with e, call it I_e

- **eIA**: delete negation, swapping explanations for mutually exclusive labels via templates
- **eKnowIA:** delete/add negation, knowledge-bases for finding antonyms and unrelated words

b) For each e' in I_e, query RevExpl to get the variable part of a reverse input: x'_v = RevExpl(x_c, e')

c) Query m on the reverse input $x' = (x_c, x_v')$ and get the reverse explanation $e_m(x')$

d) Check if $e_m(x')$ is inconsistent with $e_m(x)$

- by checking if $e_m(x')$ is in I_e

🌟 *Lots of room for improvement at every step*

The Know- Defence

1. Find all entities in the input.
2. Find all knowledge triplets that contain each entity.
3. For each entity, rank the triplets according to the algorithm in (Xu et al., 2021).
4. For each entity, extract the triplet with the highest rank.

Xu et al., Fusing Context Into Knowledge Graph for Commonsense Question Answering, Findings of ACL, 2021.

KNOW How to Make Up Your Mind! Adversarially Detecting and Alleviating Inconsistencies in NLEs

@ACL'23 M. Jang, B. Majumder, J. McAuley, T. Lukasiewicz, O. Camburu.

Experiments

e-SNLI

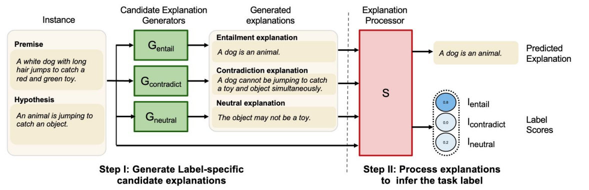


Figure 1: *Overview of NILE*: A Premise and Hypothesis pair is input to label-specific Candidate Explanation Generators G which generate natural language explanations supporting the corresponding label. The generated explanations are then fed to the Explanation Processor S, which generates label scores using the evidence present in these explanations (see Figure 3 for the architectures used in this work). In addition to the explanations, NILE also utilizes the premise and hypothesis pair (See Section 4.4.2 for a discussion on the challenges in building such a system). Please see Section 4 for details.

(Kumar and Talukdar, 2020)

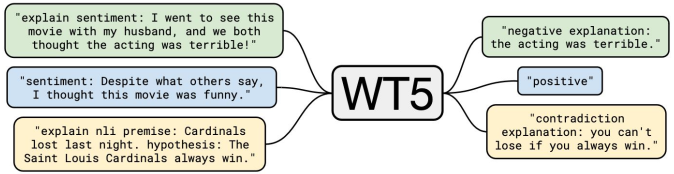


Figure 2: Diagram of our method for training a text-to-text model to explain its predictions. We train the model to generate an explanation when the text "explain" is prepended to the input. The model can still be trained for classification (without an explanation) simply by omitting the "explain" keyword. This approach is readily applicable to sentiment analysis, natural language inference (NLI), and other text tasks.

(Narang et al., 2020)

CoS-E

(a) One time-step of training a CAGE language model to generate explanations from CoS-E. It is conditioned on the question tokens Q concatenated with the answer choice tokens A_1, A_2, A_3 and previously generated tokens $E_1, \ldots, E_{i-1}$. It is trained to generate token E_i.

(b) A trained CAGE language model is used to generate explanations for a downstream commonsense reasoning model (CSRM), which itself predicts one of the answer choices.

Figure 1: An overview of CAGE trained on CoS-E and CQA.

(Rajani et al., 2019)

Kumar and Talukdar, NILE : Natural Language Inference with Faithful Natural Language Explanations, ACL, 2020.
Rajani et al., Explain Yourself! Leveraging Language Models for Commonsense Reasoning, ACL, 2019.
Narang et al., WT5?! Training Text-to-Text Models to Explain their Predictions, arxiv, 2020.

Results

Dataset	Method	Time	$\mathcal{S}_r$	$\mathcal{H}_r$
e-SNLI	eIA	10 days	2.19	384/24M
	eKnowIA	**40 min**	**12.88**	**1,494/88K**
Cos-E	eIA	2.5 days	0.32	5/5M
	eKnowIA	**5 min**	**0.95**	**13/11K**

Table 2: Comparison between eIA and eKnowIA on WT5-base. The best results are in bold; $\mathcal{S}_r$ is given in %; $\mathcal{H}_r$ values are in fractions to emphasise the high denominators of the eIA.

Results

Model	e-SNLI Acc.	e-SNLI $\mathcal{S}_r$	e-SNLI $\mathcal{H}_r$	e-SNLI e-ViL	Cos-E Acc.	Cos-E $\mathcal{S}_r$	Cos-E $\mathcal{H}_r$	Cos-E e-ViL
NILE	90.7	3.13	2.27	0.80	-	-	-	-
KnowNILE	**90.9**	**2.42**†	**1.99**†	**0.82**	-	-	-	-
CAGE	-	-	-	-	61.4	0.42	0.06	0.43
KnowCAGE	-	-	-	-	**62.6**	**0.11**†	**0.01**†	**0.44**
WT5-base	90.6	12.88	1.70	0.76	65.1	0.95	0.12	0.55
KnowWT5-base	**90.9**	**11.45**	**1.19**†	**0.80**†	**65.5**	**0.84**†	**0.09**†	**0.56**

Table 1: Results of our eKnowIA attack and our method for mitigating IN-NLEs. The best results for each pair of (model, Know-model) are in bold; $\mathcal{S}_r$ and $\mathcal{H}_r$ are given in %; † indicates that Know-models showed statistically significant difference with p-value < 0.05 (†) using the t-test.

Results

Model	e-SNLI				Cos-E			
	Acc.	$\mathcal{S}_r$	$\mathcal{H}_r$	e-ViL	Acc.	$\mathcal{S}_r$	$\mathcal{H}_r$	e-ViL
NILE	90.7	3.13	2.27	0.80	-	-	-	-
KnowNILE	**90.9**	**2.42**†	**1.99**†	**0.82**	-	-	-	-
CAGE	-	-	-	-	61.4	0.42	0.06	0.43
KnowCAGE	-	-	-	-	**62.6**	**0.11**†	**0.01**†	**0.44**
WT5-base	90.6	12.88	1.70	0.76	65.1	0.95	0.12	0.55
KnowWT5-base	**90.9**	**11.45**	**1.19**†	**0.80**†	**65.5**	**0.84**†	**0.09**†	**0.56**

Table 1: Results of our eKnowIA attack and our method for mitigating IN-NLEs. The best results for each pair of (model, Know-model) are in bold; $\mathcal{S}_r$ and $\mathcal{H}_r$ are given in %; † indicates that Know-models showed statistically significant difference with p-value < 0.05 (†) using the t-test.

Better NLE correctness does not guarantee fewer inconsistencies.

KNOW How to Make Up Your Mind! Adversarially Detecting and Alleviating Inconsistencies in NLEs

@ACL'23 M. Jang, B. Majumder, J. McAuley, T. Lukasiewicz, O. Camburu.

Results

PREMISE: A man is riding his dirt bike through the air in the desert.	
HYPOTHESIS: A man is on a motorbike PREDICTED LABEL: entailment EXPLANATION: A dirt bike is a motorbike.	HYPOTHESIS: The man is riding a motorbike. PREDICTED LABEL: contradiction EXPLANATION: A dirt bike is not a motorbike.
QUESTION: What is a person who is good at sports considered?	
CHOICES: talented, affluent, reproduce PREDICTED LABEL: talented EXPLANATION: a person who is good at sports is considered talented.	CHOICES: talented, untalented, good at PREDICTED LABEL: untalented EXPLANATION: a person who is good at sports is considered untalented

Table 3: Examples of inconsistent NLEs detected by eKnowIA for WT5 on e-SNLI and NILE on Cos-E. The first column shows the original variable part, and the second column shows the adversarial one.

PREMISE: A man is riding his dirt bike through the air in the desert.	
HYPOTHESIS: A man is on a motorbike PREDICTED LABEL: entailment EXTRACTED KNOWLEDGE: {dirt bike, IsA, motorcycle}, {desert, MannerOf, leave}, {air, HasA, oxygen} EXPLANATION: A dirt bike is a motorbike.	HYPOTHESIS: The man is riding a motorbike. PREDICTED LABEL: entailment EXTRACTED KNOWLEDGE: {dirt bike, IsA, motorcycle}, {desert, MannerOf, leave}, {air, HasA, oxygen} EXPLANATION: A dirt bike is a motorbike.
QUESTION: What is a person who is good at sports considered? CHOICES: talented, untalented, good at	
CHOICES: talented, affluent, reproduce PREDICTED LABEL: talented EXTRACTED KNOWLEDGE: {talent, RelatedTo, sports} EXPLANATION: a person who is good at sports is considered talented.	CHOICES: talented, untalented, good at PREDICTED LABEL: talented EXTRACTED KNOWLEDGE: {talent, RelatedTo, sports} EXPLANATION: a person who is good at sports is considered talented.

Table 4: Examples of successfully defended instances by KnowWT5 on e-SNLI and KnowNILE on Cos-E. This table should be read together with Table 3 to appreciate the defence.

A medical application

Explaining Chest X-ray Pathologies in Natural Language

@MICCAI'22 M. Kayser, C. Emde, B. Papiez, O. Camburu, G. Parsons, T. Lukasiewicz.

MIMIC-NLE: the first dataset of NLEs for a medical task (~45k instances)

Extract **diagnoses** and **NLEs for the diagnoses** from the radiology reports in MIMIC-CXR (Johnson et al., 2019)

LABELS: Atelectasis (Positive)

Natural Language Explanations for ***Atelectasis:***

	Clinical Evaluation:
Ground-Truth: *Opacification at the right base again is consistent with collapse of the right middle and lower lobes.*	5
RATCHET: *There is a new opacity at the right lung base which may represent atelectasis.*	4
DPT: *Bibasilar opacities likely represent atelectasis.*	1
TieNet: *Retrocardiac opacity likely reflects atelectasis.*	1

Johnson et al., MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019.

Explaining Chest X-ray Pathologies in Natural Language

@MICCAI'22 M. Kayser, C. Emde, B. Papiez, O. Camburu, G. Parsons, T. Lukasiewicz.

MIMIC-NLE: the first dataset of NLEs for a medical task (~45k instances)

Extract **diagnoses** and **NLEs for the diagnoses** from the radiology reports in MIMIC-CXR (Johnson et al., 2019)

LABELS: Atelectasis (Positive)

Natural Language Explanations for ***Atelectasis:***

	Clinical Evaluation:
Ground-Truth: *Opacification at the right base again is consistent with collapse of the right middle and lower lobes.*	5
RATCHET: *There is a new opacity at the right lung base which may represent atelectasis.*	4
DPT: *Bibasilar opacities likely represent atelectasis.*	1
TieNet: *Retrocardiac opacity likely reflects atelectasis.*	1

Improved dataset and models very soon!

Johnson et al., MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019.

Open Questions

Open Questions

Current benchmarks do not *yet* include explainability, even if it is a desiderata.

Liang et al., 2022

36 models

AI21 Labs / J1-Jumbo v1 (178B)
AI21 Labs / J1-Large v1 (7.5B)
AI21 Labs / J1-Grande v1 (17B)
AI21 Labs / J1-Grande v2 beta (17B)
Aleph Alpha / Luminous Base (13B)
Aleph Alpha / Luminous Extended (30B)
Aleph Alpha / Luminous Supreme (70B)
Anthropic / Anthropic-LM v4-s3 (52B)
BigScience / BLOOM (176B)
BigScience / BLOOMZ (176B)
BigScience / T0pp (11B)
BigCode / SantaCoder (1.1B)
Cohere / Cohere xlarge v20220609 (52.4B)
Cohere / Cohere large v20220720 (13.1B)
Cohere / Cohere medium v20220720 (6.1B)
Cohere / Cohere small v20220720 (410M)
Cohere / Cohere xlarge v20221108 (52.4B)
Cohere / Cohere medium v20221108 (6.1B)
Cohere / Cohere command nightly (6.1B)
Cohere / Cohere command nightly (52.4B)
DeepMind / Gopher (280B)
DeepMind / Chinchilla (70B)
EleutherAI / GPT-J (6B)
EleutherAI / GPT-NeoX (20B)
Google / T5 (11B)
Google / UL2 (20B)
Google / Flan-T5 (11B)
Google / PaLM (540B)
HazyResearch / H3 (2.7B)
Meta / OPT-IML (175B)
Meta / OPT-IML (30B)
Meta / OPT (175B)
Meta / OPT (66B)
Meta / Galactica (120B)
Meta / Galactica (30B)

42 scenarios

Question answering
- MMLU
- BoolQ
- NarrativeQA
- NaturalQuestions (closed-book)
- NaturalQuestions (open-book)
- QuAC
- HellaSwag
- OpenbookQA
- TruthfulQA

Information retrieval
- MS MARCO (regular)
- MS MARCO (TREC)

Summarization
- CNN/DailyMail
- XSUM

Sentiment analysis
- IMDB

Toxicity detection
- CivilComments

Text classification
- RAFT

Aspirational scenarios
- Data-to-text generation
- Fact verification
- Copywriting
- Story generation
- Biomedical scenarios
- Clinical scenarios
- Financial scenarios
- Customer services scenarios

57 metrics

Accuracy
- none
- Quasi-exact match
- F1
- Exact match
- RR@10
- NDCG@10
- ROUGE-2
- Bits/byte
- Exact match (up to specified indicator)
- Absolute difference
- F1 (set match)
- Equivalent
- Equivalent (chain of thought)
- pass@1

Calibration
- Max prob
- 1-bin expected calibration error
- 10-bin expected calibration error
- Selective coverage-accuracy area
- Accuracy at 10% coverage
- 1-bin expected calibration error (after Platt s
- 10-bin Expected Calibration Error (after Plat
- Platt Scaling Coefficient
- Platt Scaling Intercept

Robustness
- Quasi-exact match (perturbation: typos)
- F1 (perturbation: typos)
- Exact match (perturbation: typos)
- RR@10 (perturbation: typos)
- NDCG@10 (perturbation: typos)
- Quasi-exact match (perturbation: synonyms
- F1 (perturbation: synonyms)
- Exact match (perturbation: synonyms)

Venue	Desiderata
ACL, EMNLP, NAACL, LREC …	accuracy, bias, environmental impact, explainability, fairness, interpretability, linguistic plausibility, robustness sample efficiency, toxicity, training efficiency
SIGIR	accuracy, bias, explainability, fairness, inference efficiency, privacy, security, user experience/interaction
NeurIPS, ICML, ICLR, …	accuracy, fairness, interpretability, privacy, robustness, sample efficiency, theoretical guarantees, training efficiency uncertainty/calibration, user experience/interaction
AAAI	accountability, accuracy, bias, causality, creativity, emotional intelligence, explainability, fairness, interpretability memory efficiency, morality, privacy, robustness, sample efficiency, security, theoretical guarantees, transparency trustworthiness, uncertainty/calibration, user experience/interaction
COLT, UAI, AISTATS	accuracy, causality, fairness, memory efficiency, privacy, sample efficiency, theoretical guarantees, training efficiency
The Web Conference (WWW), ICWSM	accessibility, accountability, accuracy, bias, credibility/provenance, fairness, inference efficiency, legality, privacy, reliability robustness, security, transparency, trustworthiness, user experience/interaction
FAccT	causality, explainability, fairness, interpretability, legality, oversight, participatory design, privacy, security transparency, user experience/interaction
WSDM	accountability, accuracy, credibility/provenance, explainability, fairness, inference efficiency, interpretability privacy, robustness, toxicity, transparency, trustworthiness, user experience/interaction
KDD	accuracy, explainability, fairness, inference efficiency, interpretability, maintainability, memory efficiency, privacy robustness, training efficiency
Union	accessibility, accountability, accuracy, bias, causality, creativity, credibility/provenance, emotional intelligence environmental impact, explainability, fairness, inference efficiency, interpretability, legality linguistic plausibility, maintainability, memory efficiency, morality, oversight, participatory design, privacy reliability, robustness, sample efficiency, security, theoretical guarantees, toxicity, training efficiency transparency, trustworthiness, uncertainty/calibration, user experience/interaction

Table 2. **Enumeration of desiderata.** To enumerate the space of desiderata, we first compile a list of venues from https://aideadlin.es/. For each venue, we enumerate desiderata that are well-studied in that community.

Category	Desiderata
Requires knowledge of how model was created	causality, environmental impact, linguistic plausibility, memory efficiency, participatory design, privacy sample efficiency, training efficiency, theoretical guarantees
Requires the model have specific structure	credibility/provenance, explainability
Requires more than blackbox access	interpretability
Require knowledge about the broader system	maintainability, reliability, security, transparency
Requires knowledge about the broader social context	accessibility, accountability, creativity, emotional intelligence, legality, morality, oversight trustworthiness, user experience/interaction
Satisfies our conditions (i.e. none of the above)	accuracy, bias, fairness, inference efficiency, robustness, toxicity, uncertainty/calibration

Table 3. **Taxonomy of desiderata.** To taxonomize the space of desiderata, we categorize each desideratum based on the requirements needed to properly measure it.

Holistic Evaluation of Language Models

Percy Liang† Rishi Bommasani† Tony Lee†[1]
Dimitris Tsipras* Dilara Soylu* Michihiro Yasunaga* Yian Zhang* Deepak Narayanan* Yuhuai Wu*[2]

Ananya Kumar Benjamin Newman Binhang Yuan Bobby Yan Ce Zhang
Christian Cosgrove Christopher D. Manning Christopher Ré Diana Acosta-Navas
Drew A. Hudson Eric Zelikman Esin Durmus Faisal Ladhak Frieda Rong Hongyu Ren
Huaxiu Yao Jue Wang Keshav Santhanam Laurel Orr Lucia Zheng Mert Yuksekgonul
Mirac Suzgun Nathan Kim Neel Guha Niladri Chatterji Omar Khattab Peter Henderson
Qian Huang Ryan Chi Sang Michael Xie Shibani Santurkar Surya Ganguli
Tatsunori Hashimoto Thomas Icard Tianyi Zhang Vishrav Chaudhary William Wang
Xuechen Li Yifan Mai Yuhui Zhang Yuta Koreeda

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

Language models (LMs) are becoming the foundation for almost all major language technologies, but their capabilities, limitations, and risks are not well understood. We present Holistic Evaluation of Language Models (HELM) to improve the transparency of language models. First, we taxonomize the vast space of potential scenarios (i.e. use cases) and metrics (i.e. desiderata) that are of interest for LMs. Then we select a broad subset based on coverage and feasibility, noting what's missing or underrepresented (e.g. question answering for neglected English dialects, metrics for trustworthiness). Second, we adopt a multi-metric approach: We measure 7 metrics (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency) for each of 16 core scenarios to the extent possible (87.5% of the time), ensuring that metrics beyond accuracy don't fall to the wayside, and that trade-offs across models and metrics are clearly exposed. We also perform 7 targeted evaluations, based on 26 targeted scenarios, to more deeply analyze specific aspects (e.g. knowledge, reasoning, memorization/copyright

Liang et al., Holistic Evaluation of Language Models. 2022

Open Questions

- **XAI (e.g., NLEs) to be part of benchmarks (e.g., HELM)**

Open Questions

- XAI (e.g., NLEs) to be part of benchmarks (e.g., HELM)
- Metrics for NLEs: faithfulness, correctness, consistency, …

Open Questions

- XAI (e.g., NLEs) to be part of benchmarks (e.g., HELM)
- Metrics for NLEs: faithfulness, correctness, consistency, …
- Enhance faithfulness, correctness, consistency, …

Open Questions

- XAI (e.g., NLEs) to be part of benchmarks (e.g., HELM)
- Metrics for NLEs: faithfulness, correctness, consistency, …
- Enhance faithfulness, correctness, consistency, …
- Usefulness (user studies)

Open Questions

- XAI (e.g., NLEs) to be part of benchmarks (e.g., HELM)
- Metrics for NLEs: faithfulness, correctness, consistency, …
- Enhance faithfulness, correctness, consistency, …
- Usefulness (user studies)
- Enhancing other aspects: robustness, performance

…

Thank you!

Questions?

@oanacamb

Visual-Textual Understanding

SNLI

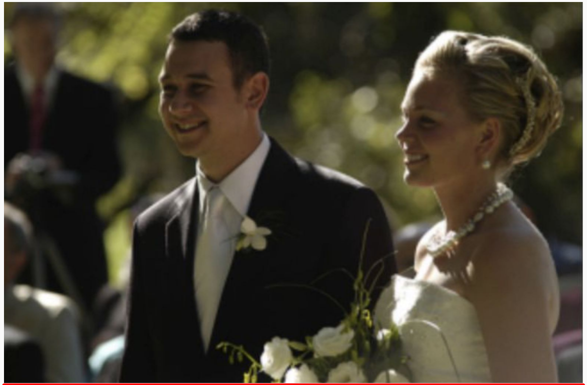
Premise:
A man and woman getting married.

Hypothesis:
A man and a woman inside a church.

Label:
Neutral

(Xie et al., 2019)

Flickr30k



Caption:
A man and woman getting married.

Xie. et al., A novel task for fine-grained image understanding, 2019

e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks

@ICCV'21 M. Kayser, O. Camburu, L. Salewski, C. Emde, V. Do, Z. Akata, T. Lukasiewicz.

SNLI-VE (Xie et al., 2019)

Premise:

Hypothesis:
Two women are holding food in their hands.

Label:
Entailment

Premise:

Hypothesis:
A man is driving down a lonely road.

Label:
Contradiction

Premise:

Hypothesis:
A man is repainting a garage

Label:
Neutral

Xie. et al., A novel task for fine-grained image understanding, 2019

e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks

@ICCV'21 M. Kayser, O. Camburu, L. Salewski, C. Emde, V. Do, Z. Akata, T. Lukasiewicz.

e-SNLI-VE = SNLI-VE + e-SNLI + Corrections → large dataset (400k, 14k, 14k)

Premise:

Hypothesis:
Two women are holding food in their hands.

Label:
Entailment

Explanation: Holding to go packages implies that there is food in it.

Premise:

Hypothesis:
A man is driving down a lonely road.

Label:
Contradiction

Explanation: A road can't be lonely if there is a crowd of people.

Premise:

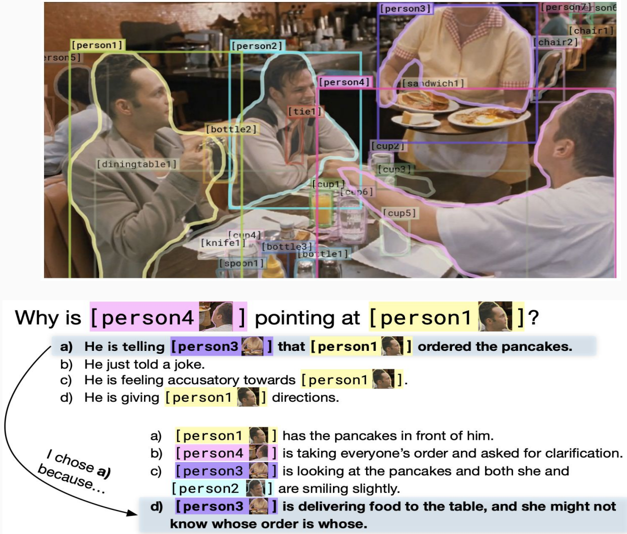
Hypothesis:
A man is repainting a garage

Label:
~~Neutral~~ Contradiction

Explanation: The man is just staying in front of the garage with no signs of repairing being done.

Other Datasets with NLEs

VCR (Zellers et al., 2019)
(~240k instances)



VQA-X (Park et al., 2018)
(~33k instances)

Q: What is the person doing?
A: Snowboarding.
Because... they are on a snowboard in snowboarding outfit.

Park et al., Multimodal explanations: Justifying decisions and pointing to the evidence. In CVPR, 2018.
Zellers et al., From recognition to cognition: Visual commonsense reasoning. In CVPR, 2019.

📏 How do we evaluate NLEs?

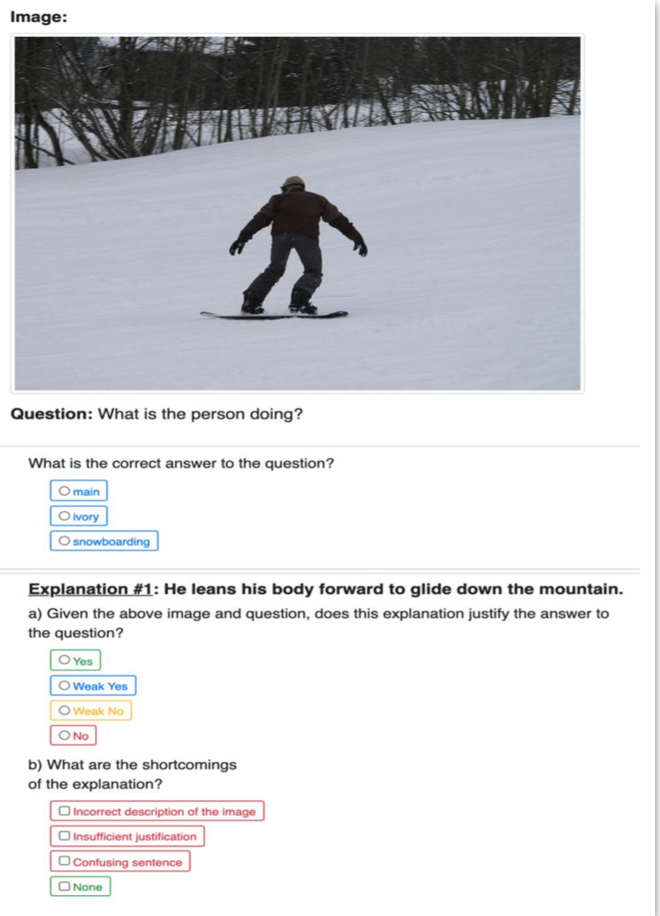
❌ Lack of **unified** evaluation framework

- Different automatic metrics
- Different human evaluation
  - correct/incorrect
  - scale (1 to 5)
  - better/same/worse than ground-truth

e-ViL: NLEs’ Correctness Metric

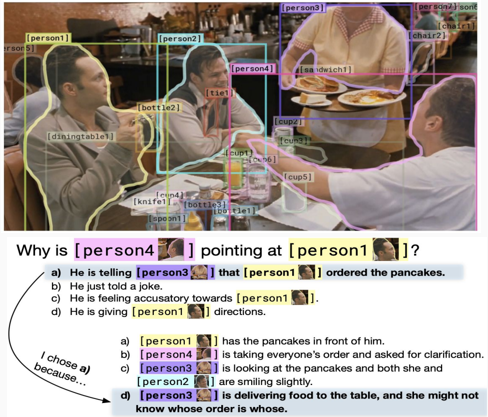
A **human evaluation** framework for NLEs

- One model at a time to **avoid potential anchoring effects among models**
- For every generated NLE, **ground-truth is also evaluated** for uniform anchoring and comparison
- Given the image and question, **does the explanation justify the answer?**
  - No / Weak_No / Weak_Yes / Yes
  - **e-ViL score = #Yes + ⅔ #Weak_Yes + ⅓ #Weak_No**
- Collect potential **shortcomings**
  - incorrect description of the image
  - insufficient justification
  - nonsensical



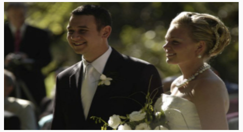
e-ViL: The Datasets

VCR (Zellers et al., 2019)



e-SNLI-VE

Premise:



Hypothesis:
The man and woman are about to go on a honeymoon.

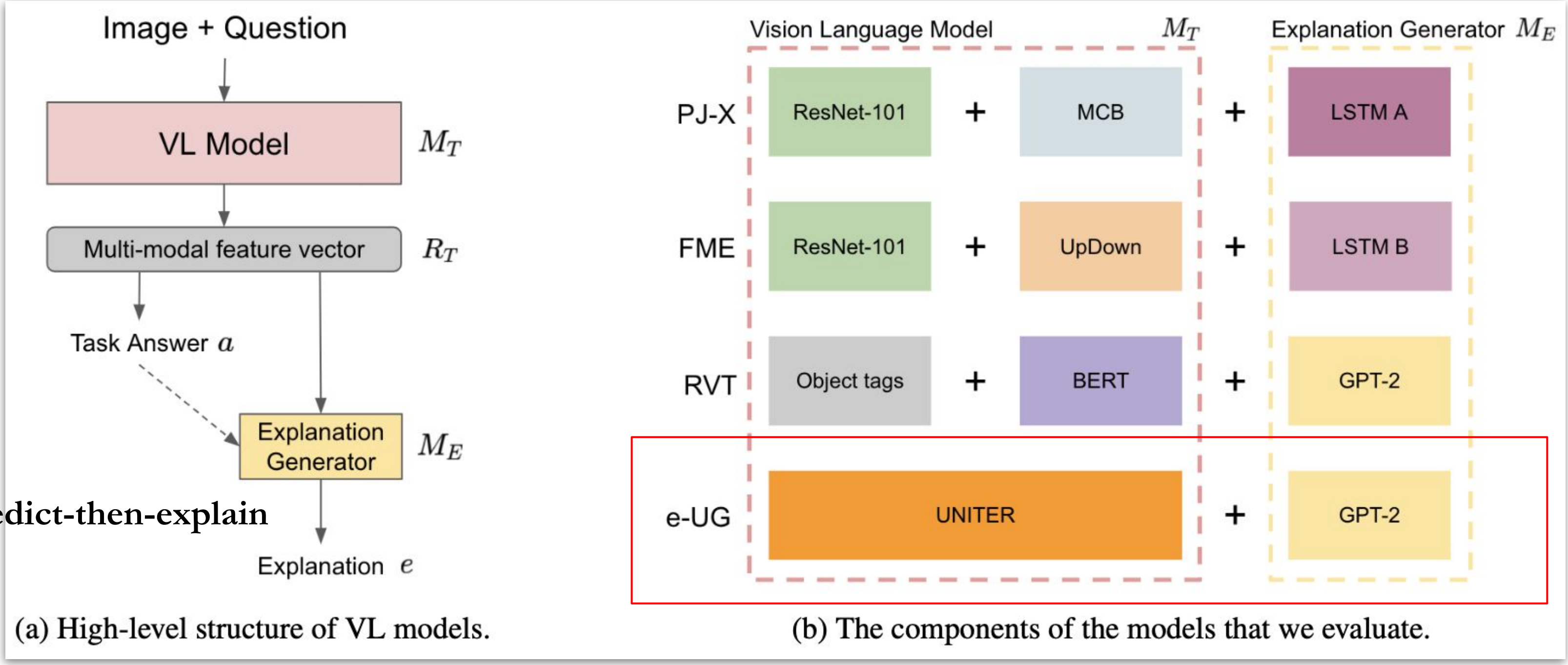
Label: Neutral

Explanation:
Not all couples go on a honeymoon right after getting married.

VQA-X (Park et al., 2018)

Park et al., Multimodal explanations: Justifying decisions and pointing to the evidence. In CVPR, 2018.
Zellers et al., From recognition to cognition: Visual commonsense reasoning. In CVPR, 2019.

📏 e-ViL: The Models



Predict-then-explain

(a) High-level structure of VL models.

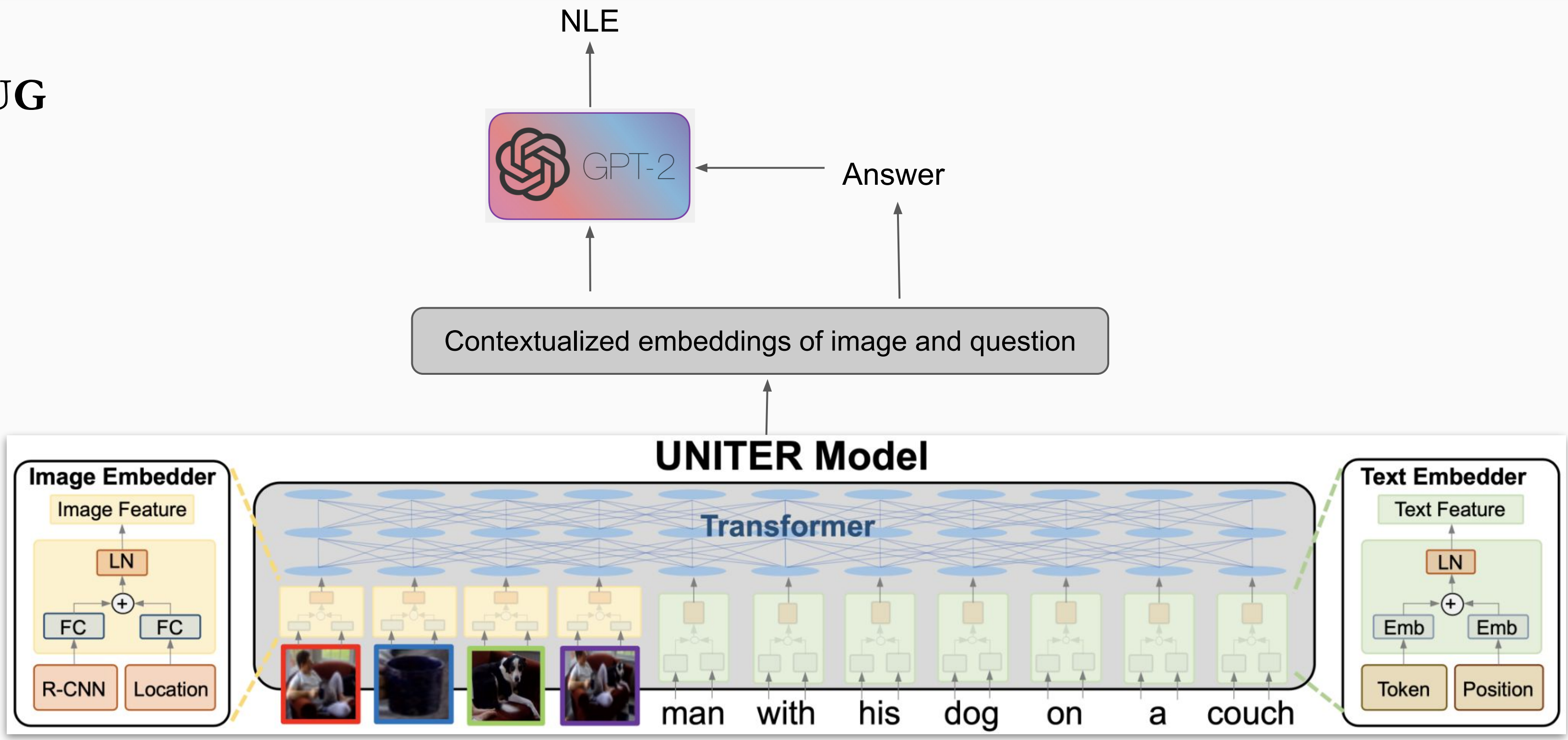
(b) The components of the models that we evaluate.

Park et al., Multimodal explanations: Justifying decisions and pointing to the evidence. CVPR 2018.
Wu and Mooney, Faithful multimodal explanation for visual question answering. BlackboxNLP 2019.
Marasović et al., Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. EMNLP Findings 2020.
Chen et al., UNITER: Universal image-text representation learning. ECCV 2020.

e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks

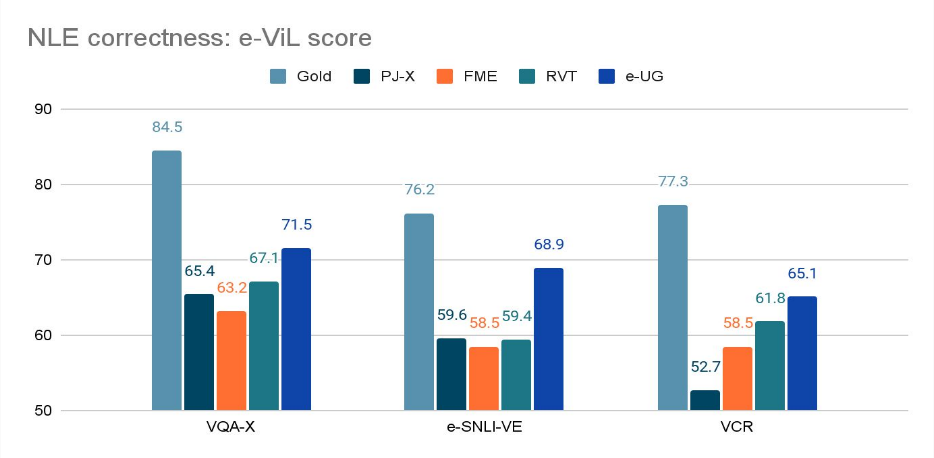
@ICCV'21 M. Kayser, O. Camburu, L. Salewski, C. Emde, V. Do, Z. Akata, T. Lukasiewicz.

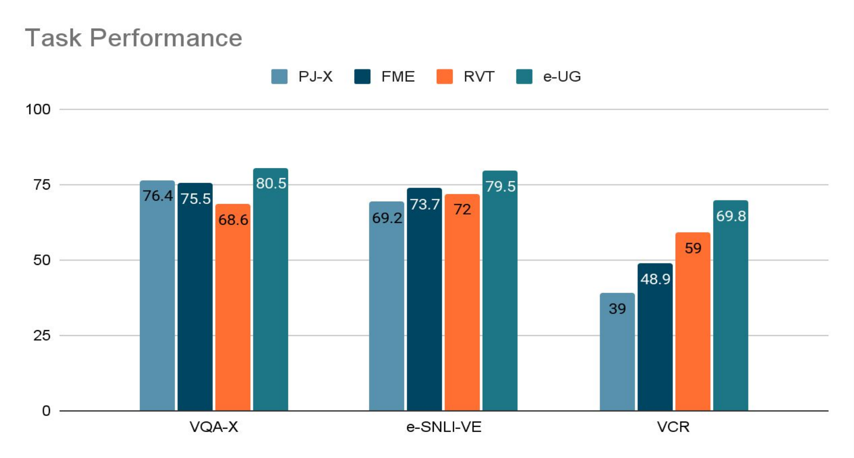
e-UG



Chen et al., UNITER: Universal image-text representation learning. ECCV 2020.

Results





Model	Untrue to Image	Lack of Justification	Non-sensical Sentence
PJ-X	25.0%	26.4%	8.9%
RVT	20.4%	24.2%	12.0%
FME	21.8%	**23.1%**	13.7%
e-UG	**15.9%**	25.0%	**7.4%**

Results

	Human Evaluation:
Hypothesis: A dog is playing with a cat. **Relation:** Contradiction **GT Explanation:** A man running and a dog playing with a cat are two very distinct activities.	
PJ-X: a dog is not a cat	0.00
FME: a dog is not a cat	0.17
RVT: A cat is not a dog.	0.00
e-UG: A dog is not a football player.	0.56

(a) e-SNLI-VE.

	Human Evaluation:
Hypothesis: The lady is the owner of the store. **Relation:** Neutral **GT Explanation:** We cannot tell from this picture if the lady is the owner of the store.	
PJ-X: a woman looking at a microscope does not imply that she is looking for the store	0.56
FME: a woman can be a man or a woman	0.17
RVT: Just because a lady is holding a book does not mean she is the owner of the store.	0.67
e-UG: Just because a lady is working at a store does not mean she is the owner.	1

(b) e-SNLI-VE.

Results

⚖️ **Automatic metrics for correctness**

- Mostly weak correlations
- **Recommended metrics: BERTScore, METEOR, and BLEURT**

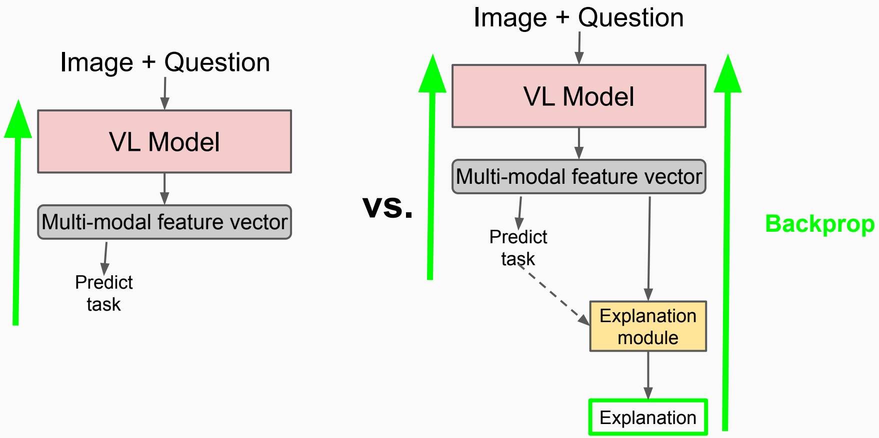
🌟 *Open Question:* How to automatically evaluate the correctness of NLEs?

Metric	All datasets	VQA-X	e-SNLI-VE	VCR
BLEU-1	0.222	0.396	0.123	*0.032*
BLEU-2	0.236	0.412	0.142	*0.034*
BLEU-3	0.224	0.383	0.139	*0.039*
BLEU-4	0.216	0.373	0.139	*0.038*
METEOR	0.288	**0.438**	0.186	0.113
ROUGE-L	0.238	0.399	0.131	*0.050*
CIDEr	0.245	0.404	0.133	*0.093*
SPICE	0.235	0.407	0.162	0.116
BERTScore	**0.293**	0.431	0.189	**0.138**
BLEURT	0.248	0.338	**0.208**	0.128

Table 6: Correlation between human evaluation and automatic NLG metrics on NLEs. All values, except those in *italic*, have p-values < 0.001.

Results

Can NLEs increase task performance?

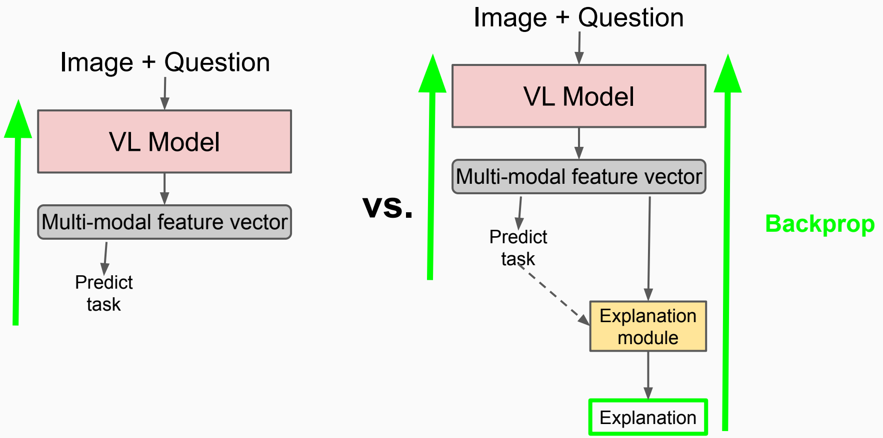


		VQA-X		SNLI-VE		VCR	
Model	M_T model	M_T only	Joint	M_T only	Joint	M_T only	Joint
PJ-X	MCB [18]	N.A.	N.A.	69.7	69.2	38.5	39.0
FME	UpDown [3]	N.A.	N.A.	71.4	73.7	35.7	48.9
e-UG	UNITER [15]	80.0	80.5	79.4	79.5	69.3	69.8

Table 4: Comparison of task scores S_T (e.g., accuracies) when the models are trained only on task T vs. when trained jointly on tasks T and E. Scores are underlined if their difference is greater than 0.5.

Results

Can NLEs increase task performance?



🌟 *Open Question:* How can we take more advantage of the rich signal in the NLEs to improve performance?

		VQA-X		SNLI-VE		VCR	
Model	M_T model	M_T only	Joint	M_T only	Joint	M_T only	Joint
PJ-X	MCB [18]	N.A.	N.A.	69.7	69.2	38.5	39.0
FME	UpDown [3]	N.A.	N.A.	71.4	<u>73.7</u>	35.7	<u>48.9</u>
e-UG	UNITER [15]	80.0	80.5	79.4	79.5	69.3	69.8

Table 4: Comparison of task scores S_T (e.g., accuracies) when the models are trained only on task T vs. when trained jointly on tasks T and E. Scores are underlined if their difference is greater than 0.5.

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

Goal: **knowledge grounding** for NLEs-generating models

Model	Untrue to Image	Lack of Justification	Non-sensical Sentence
PJ-X	25.0%	26.4%	8.9%
RVT	20.4%	24.2%	12.0%
FME	21.8%	**23.1%**	13.7%
e-UG	**15.9%**	25.0%	**7.4%**

Table 5: Main shortcomings of the generated explanations, by models and by datasets. Human judges could choose multiple shortcomings per explanation. The best model results are in bold.

(Kayser et al., 2021)

PREMISE: A guy in a red jacket is snowboarding in midair.

ORIGINAL HYPOTHESIS: A guy is outside in the snow.
PREDICTED LABEL: entailment
ORIGINAL EXPLANATION: **Snowboarding is done outside.**

REVERSE HYPOTHESIS: The guy is outside.
PREDICTED LABEL: contradiction
REVERSE EXPLANATION: **Snowboarding is not done outside.**

PREMISE: The sun breaks through the trees as a child rides a swing.

ORIGINAL HYPOTHESIS: A child rides a swing in the daytime.
PREDICTED LABEL: entailment
ORIGINAL EXPLANATION: **The sun is in the daytime.**

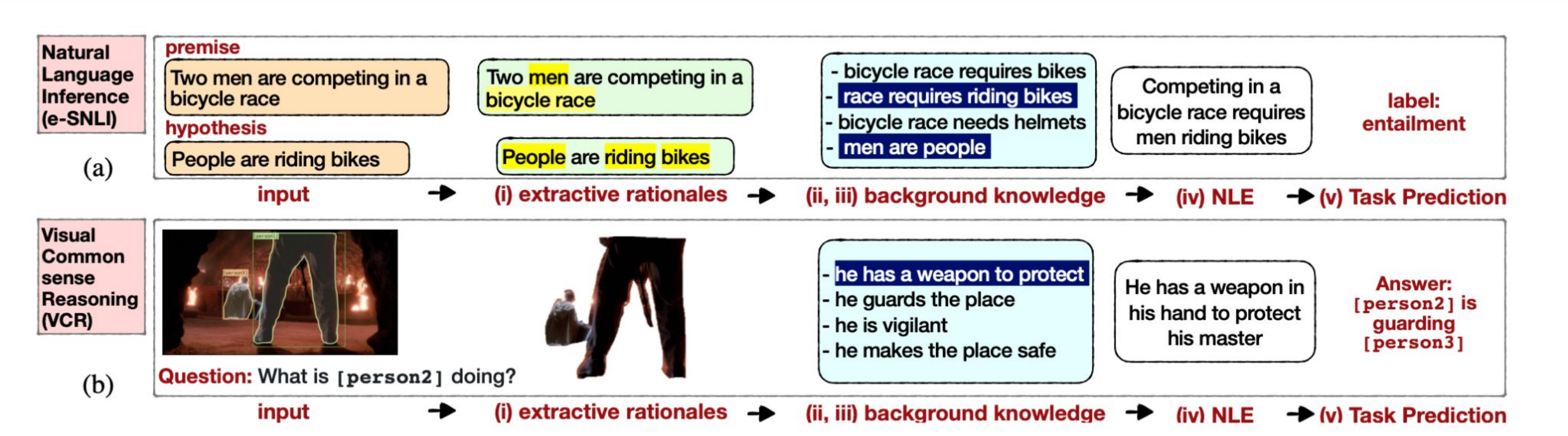
REVERSE HYPOTHESIS: The sun is in the daytime.
PREDICTED LABEL: neutral
REVERSE EXPLANATION: **The sun is not necessarily in the daytime.**

(Camburu et al., 2020)

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

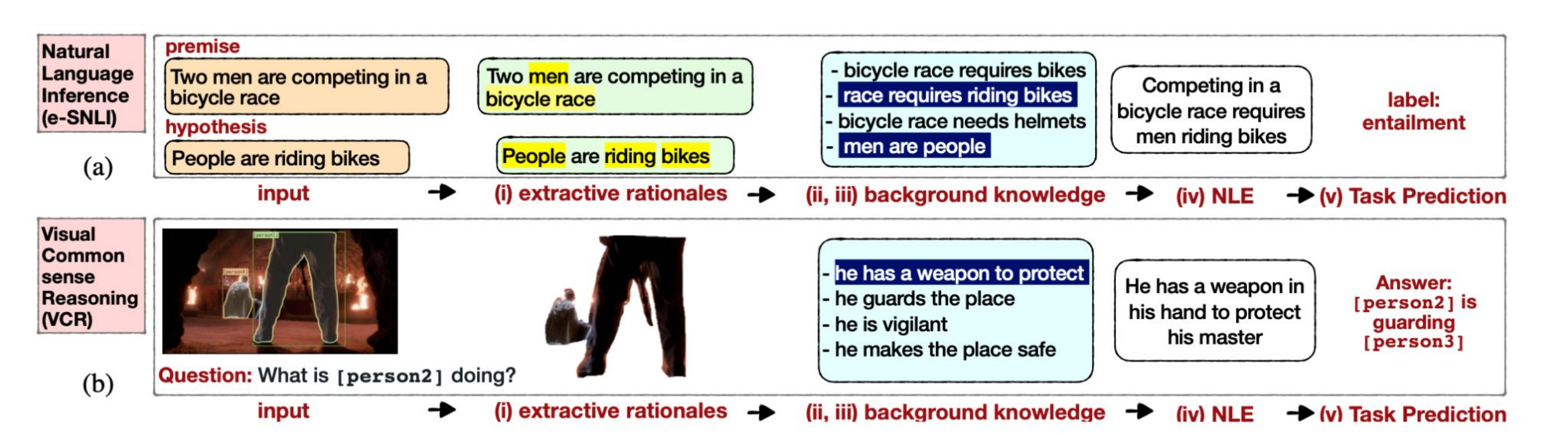
RExC: Extractive **R**ationales, Natural Language **Ex**planations, and **C**ommonsense

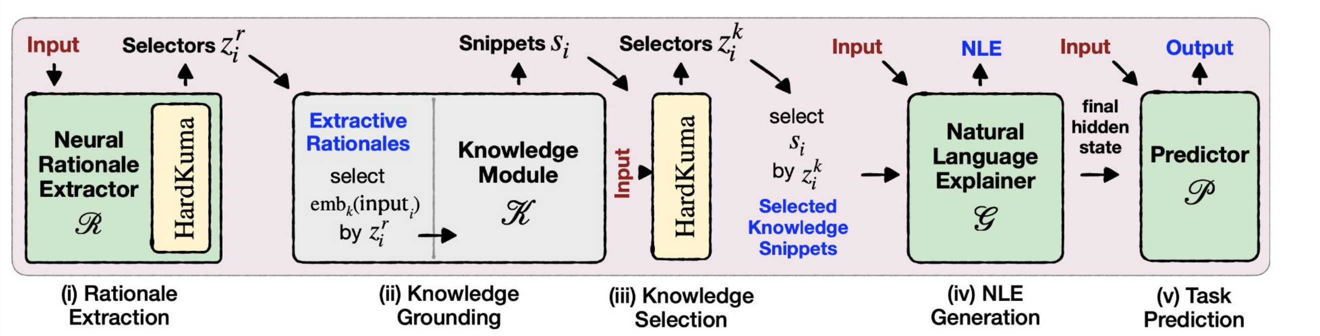


Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

RExC: Extractive **R**ationales, Natural Language **Ex**planations, and **C**ommonsense



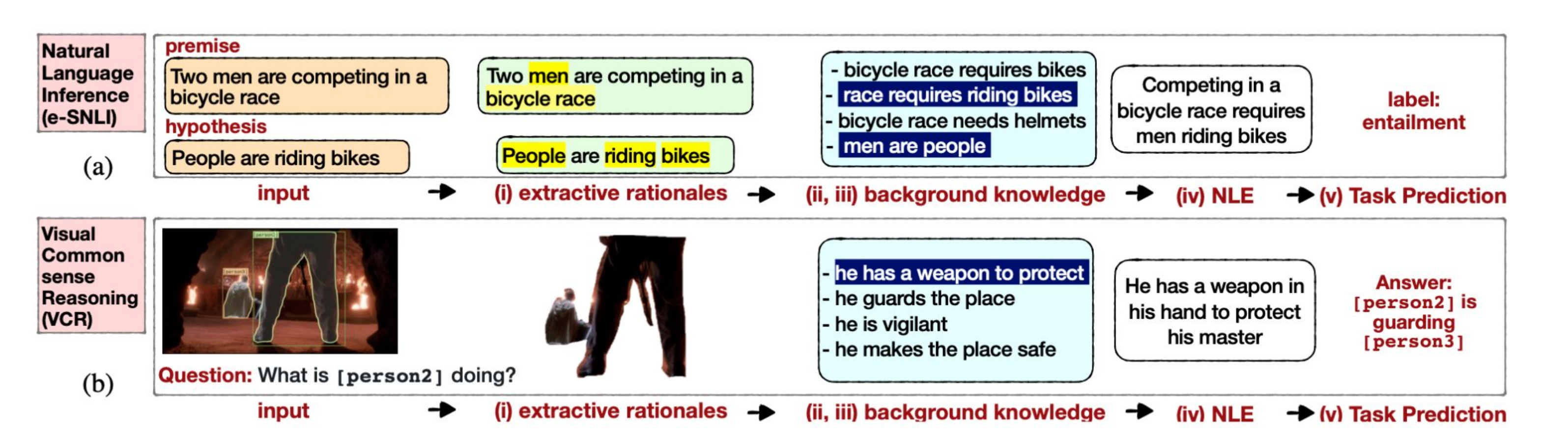


For HardKuma: J. Bastings et al., Interpretable Neural Predictions with Differentiable Binary Variables, ACL 2019.

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

Advantages of RExC



(1) **knowledge-grounded** self-rationalization model

(2) SOTA in both **extractive rationales** (ERs) and **natural language explanations** (NLEs)

(3) **"white-layer"/"peephole" architecture might** give better faithfulness

(4) **self-explainable** model that also obtains **SOTA task-performance**

(5) **replaceable modules: could use ChatGPT as the knowledge module**

(6) strong **zero-shot NLE** performance

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

Experiments

Natural Language Tasks

Natural Language Inference

premise: Two men are competing in a bicycle race

hypothesis: People are riding bikes

label *entailment*

e-SNLI (Camburu et al., 2018)

Commonsense Validation

A: Coffee stimulates people

B: Coffee depresses people

label *B is invalid*

ComVE (Wang et al., 2019)

Commonsense QA

Q: Where does a wild bird usually live?

A: a) cage, b) sky, c) countryside, d) desert, e) windowsill

label *sky*

CoS-E (Rajani et al., 2019)

Vision Language Tasks

Visual Entailment

Hypothesis: Some tennis players pose

label *entailment*

e-SNLI-VE (Kayser et al., 2021)

Visual Commonsense Reasoning

Q: What is the place?

label *They are in a hospital room*

VCR (Zellers et al., 2019)

C. Wang et al., Does it make sense? And why? A pilot study for sense making and explanation. ACL, 2019.
N. Rajani et al., Explain Yourself! Leveraging Language Models for Commonsense Reasoning, ACL, 2019.
M. Kayser et al., e-ViL: A Dataset and Benchmark for Natural Language Explanations inVision-Language Tasks, 2021.
R. Zellers et al., From recognition to cognition: Visual commonsense reasoning. CVPR, 2019.

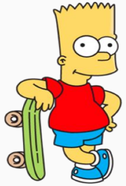
Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

NLP

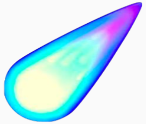
BART: a Seq2Seq pretrained transformer with a MLP prediction head

(Lewis et al., 2020)



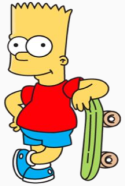
COMET: Commonsense Transformer trained on ConceptNet

(Bosselut et al., 2019)



BART: a Seq2Seq pretrained transformer with a Language Model head

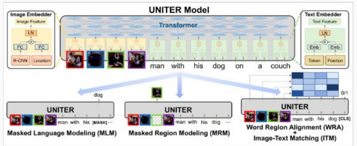
(Lewis et al., 2020)



VL

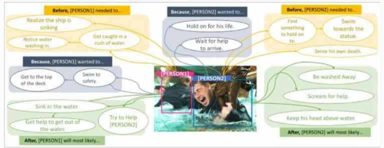
UNITER: a Seq2Seq pretrained transformer for text and images with a MLP prediction head

(Chen et al., 2020)



Visual-COMET: Commonsense Transformer trained on Visual Commonsense Graph

(Park et al., 2020)



GPT2: a pretrained transformer-based Language Model

(Radford et al., 2020)

GPT-2

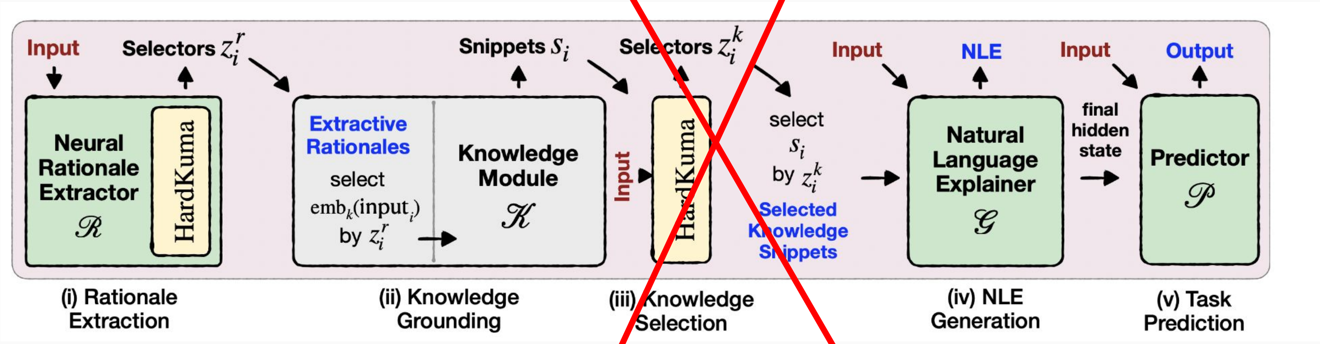
Generative knowledge modules to avoid no-hit issue of indexed KBs

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

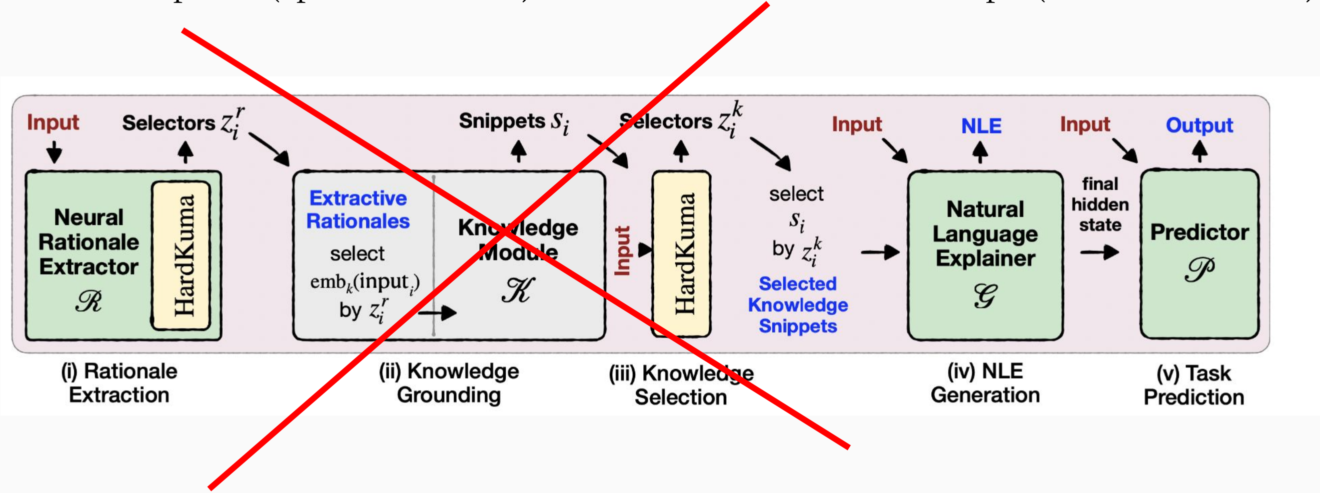
Ablations

- knowledge selection (w/o KN-Sel)
- ER and knowledge selectors (w/o KN & ER)
- NLE generator (RExC-ZS) – supervision only from the output and selected knowledge snippets as NLEs
- generative knowledge module replaced with a retrieval-based knowledge source (RExC-RB)
    - ConceptNet (Speer et al., 2017) and Visual Commonsense Graph (Zellers et al., 2019)



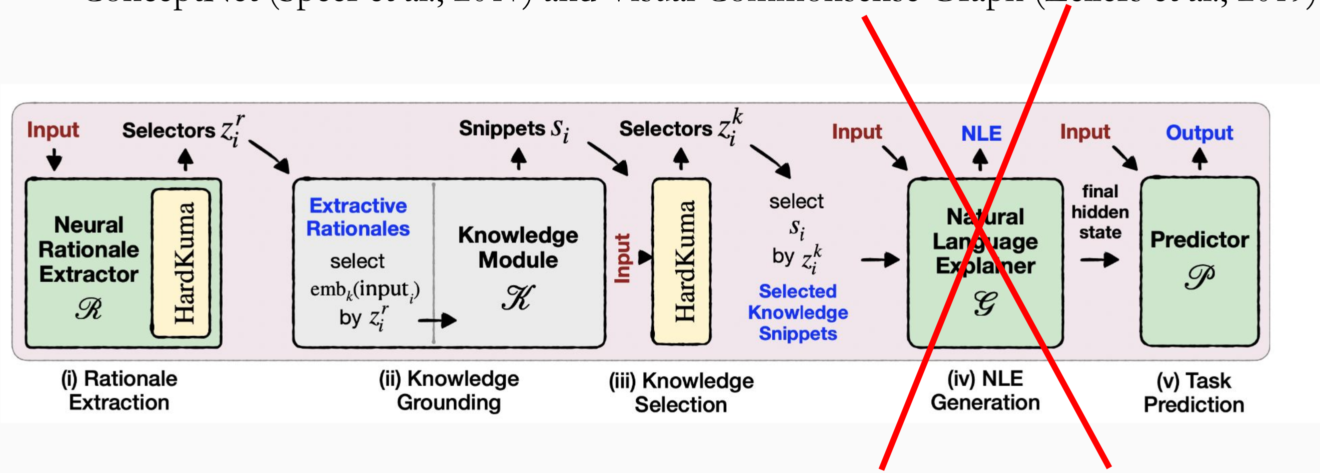
Ablations

- knowledge selection (w/o KN-Sel)
- ER and knowledge selectors (w/o KN & ER)
- NLE generator (RExC-ZS) – supervision only from the output and selected knowledge snippets as NLEs
- generative knowledge module replaced with a retrieval-based knowledge source (RExC-RB)
    - ConceptNet (Speer et al., 2017) and Visual Commonsense Graph (Zellers et al., 2019)



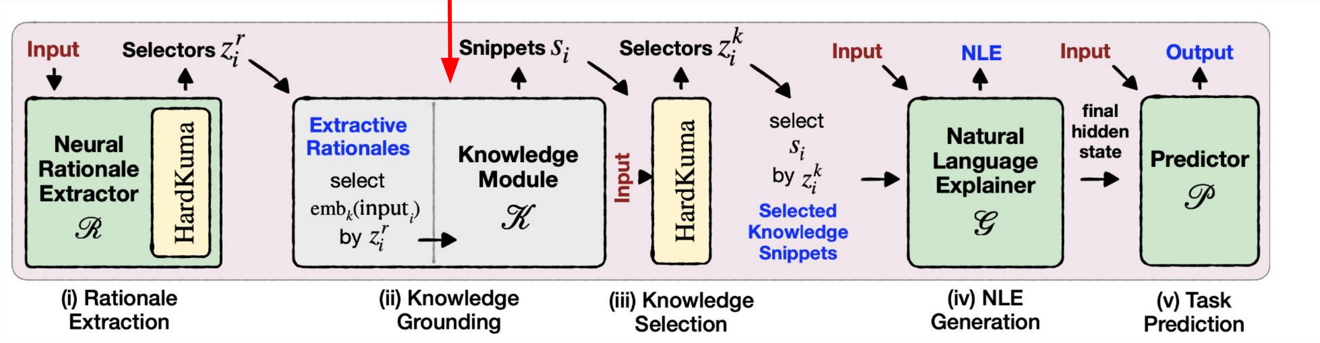
Ablations

- knowledge selection (w/o KN-Sel)
- ER and knowledge selectors (w/o KN & ER)
- NLE generator (RExC-ZS) – supervision only from the output and selected knowledge snippets as NLEs
- generative knowledge module replaced with a retrieval-based knowledge source (RExC-RB)
    - ConceptNet (Speer et al., 2017) and Visual Commonsense Graph (Zellers et al., 2019)



Ablations

- knowledge selection (w/o KN-Sel)
- ER and knowledge selectors (w/o KN & ER)
- NLE generator (RExC-ZS) – supervision only from the output and selected knowledge snippets as NLEs
- generative knowledge module replaced with a retrieval-based knowledge source (RExC-RB)
    - ConceptNet (Speer et al., 2017) and Visual Commonsense Graph (Zellers et al., 2019)



Human evaluation of NLE quality

NLE score

Yes → 1

Weak Yes → ⅔

Weak No → ⅓

No → 0

Image:

Question: how does [person2] feel about what [person1] is telling him?

1. What is the correct answer?

- ○ He is enjoying it.
- ○ He doesn't like what [person1] is saying.
- ◉ He is concerned and a little upset.
- ○ [person6] is upset that [person1] is ridiculing his plan.

Given the image and the question, do the explanations below justify the answer to the question?

Explanation #1: He is in shock thinking something bad is about to happen.

- ◉ Yes
- ○ Weak Yes
- ○ Weak No
- ○ No

What are the shortcomings of Explanation #1?

- ☐ Contradicts commonsense
- ☐ Insufficient justification
- ☐ Irrelevant to the inout image and question
- ☐ Too verbose or repititive
- ☐ Too trivial
- ☑ None

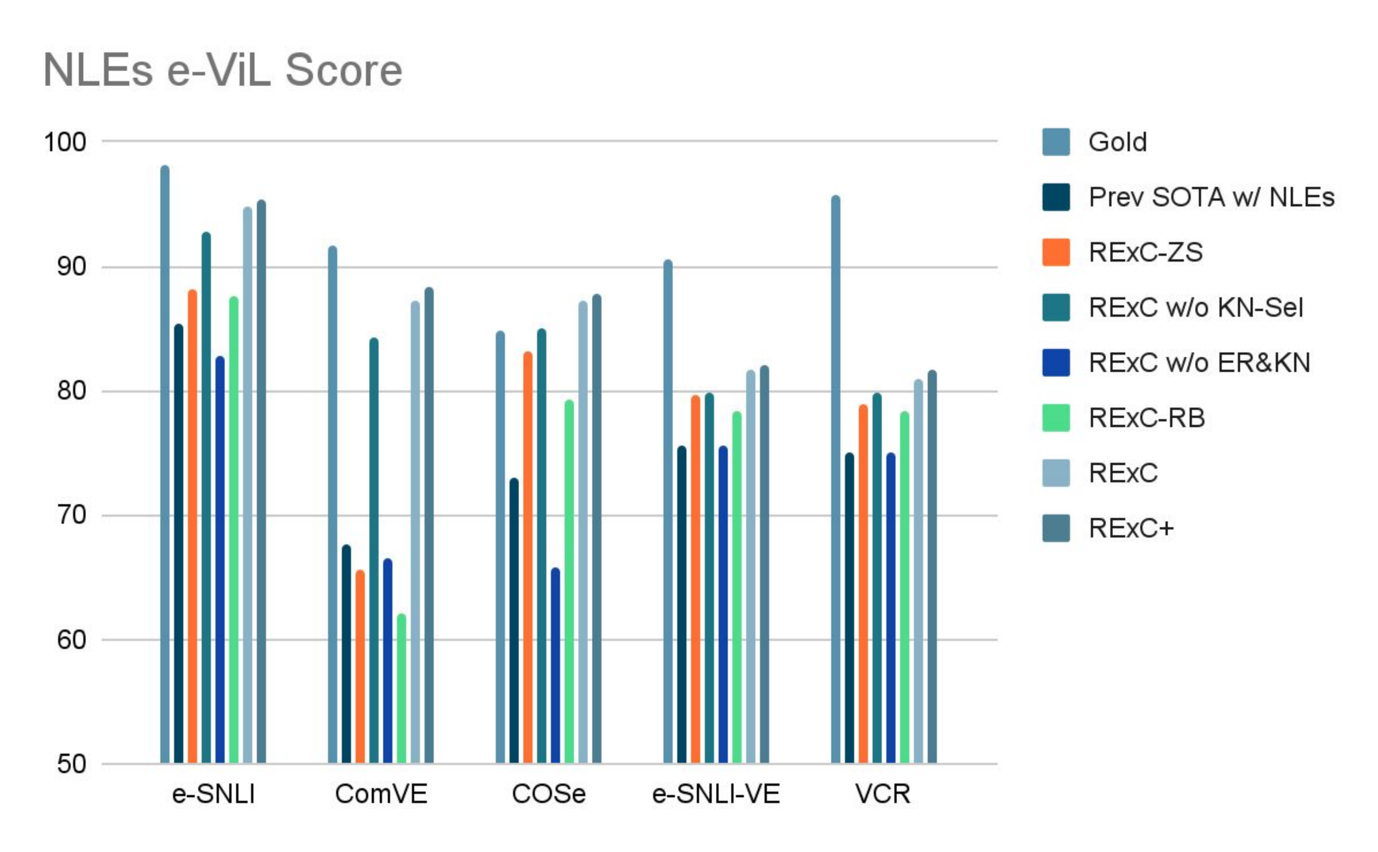
It's a good explanation.

Figure 10. Snapshot of our human evaluation with a list of possible shortcomings.

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

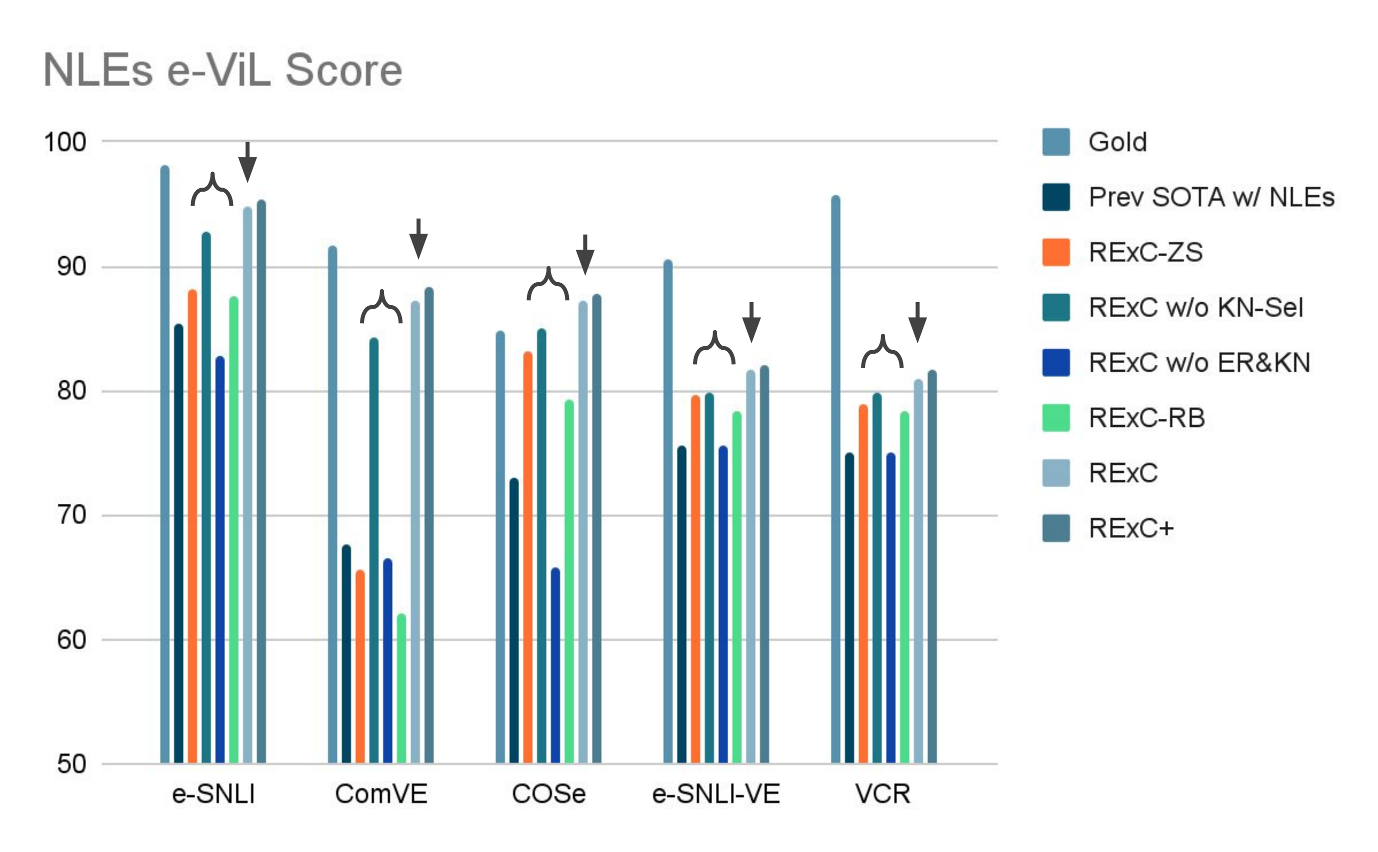
Results



Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

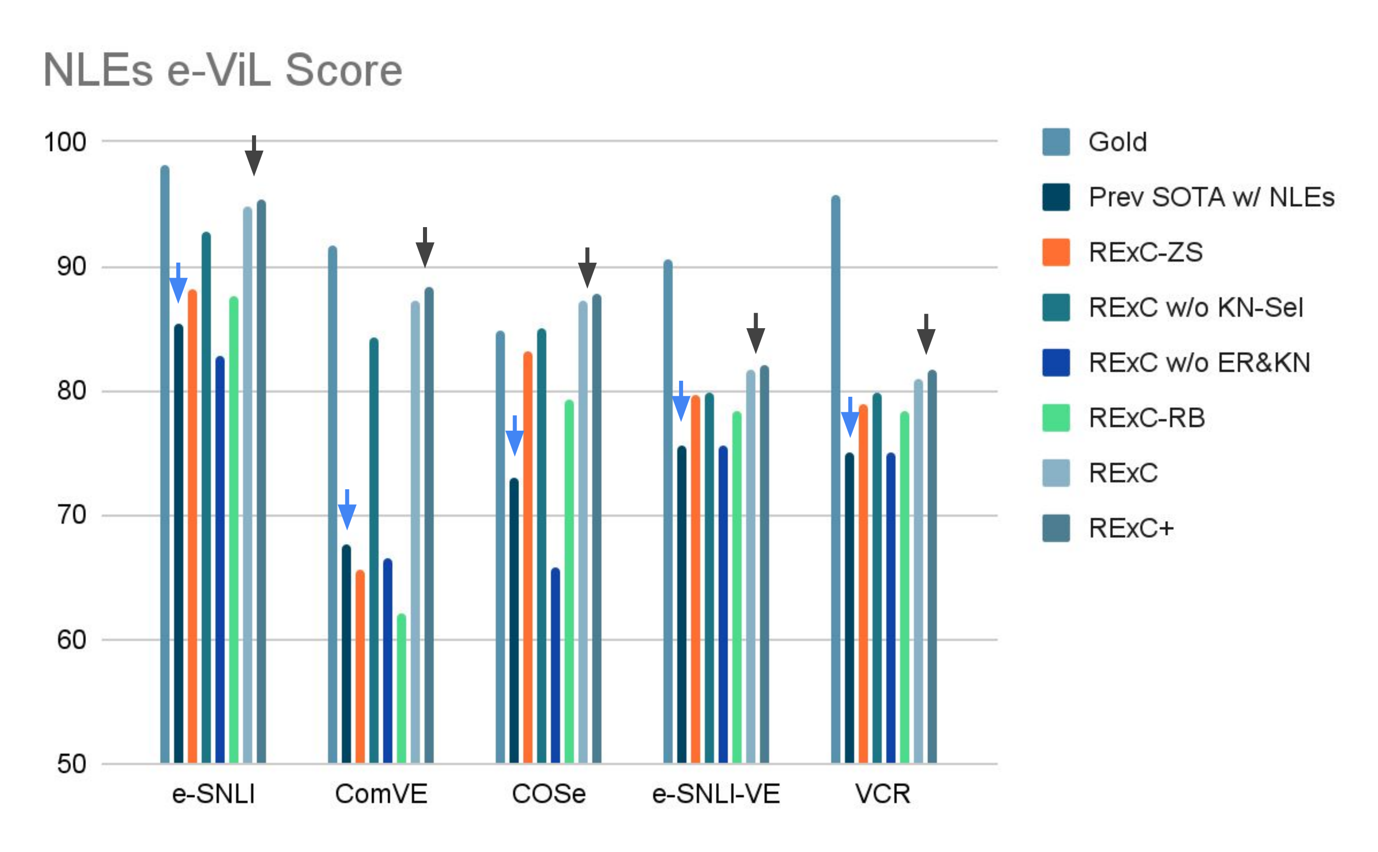
Results



Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

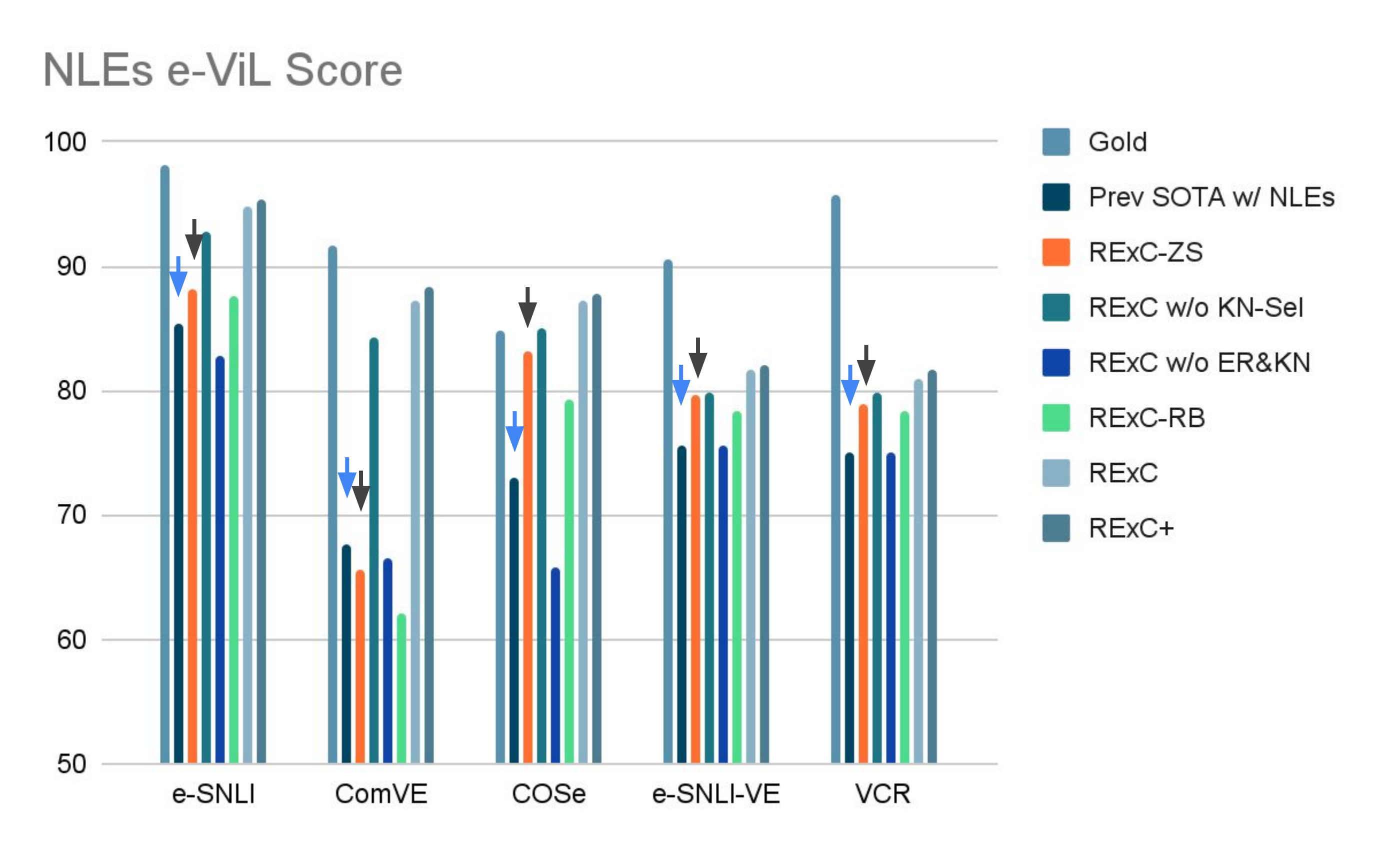
Results



Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

Results



Results

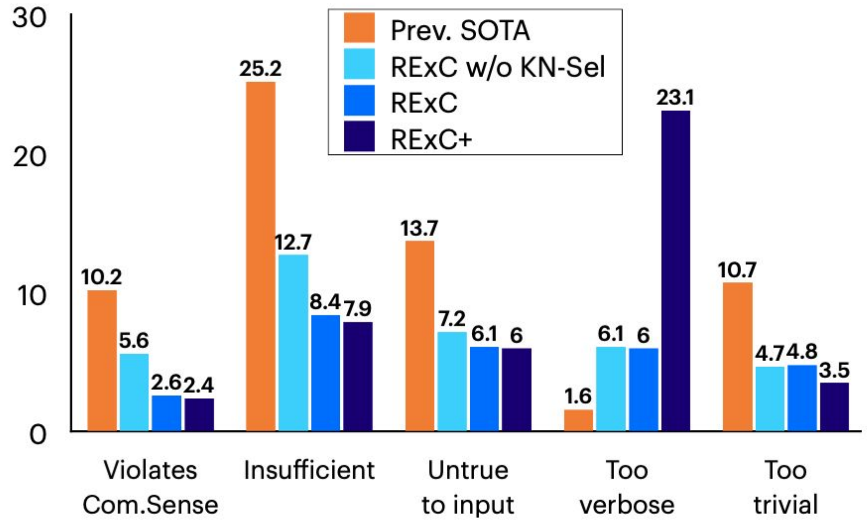
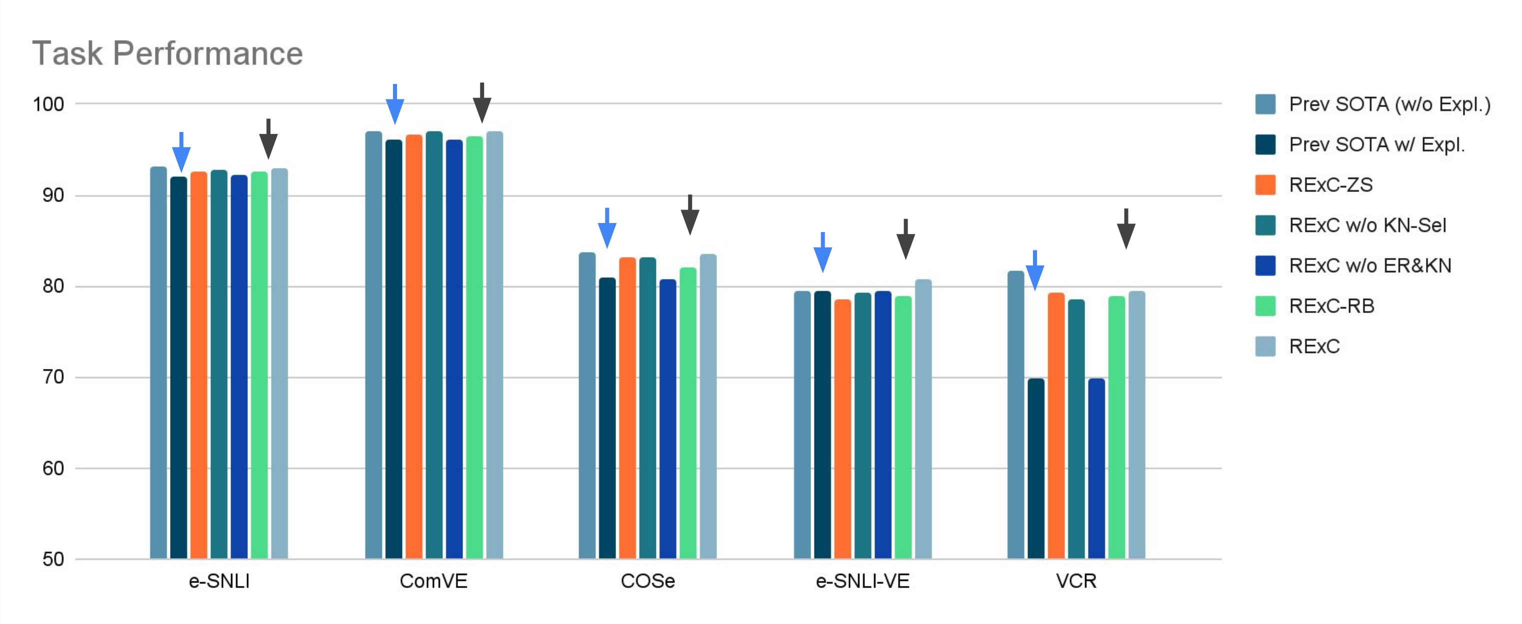


Figure 6. **Main limitations of the generated NLEs** obtained from user study. All numbers are in % and are averaged by systems and datasets for both NL and VL tasks. Human annotators could choose multiple limitations for an NLE.

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

Results



Results

RExC also outperforms the previous SOTA for extractive rationales

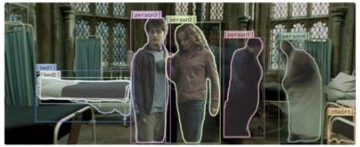
Table 3. **ER quality.** Comparison of previous SOTA models (DeYoung et al., 2020) for rationale extraction vs. RExC for ER quality. Best numbers are in **bold**.

	e-SNLI			COSe		
System	**Acc.**	**IOU**	**Tok.**	**Acc.**	**IOU**	**Tok.**
SOTA	73.3	70.4	70.1	34.4	38.9	51.9
RExC	**78.3**	**72.8**	**73.5**	**39.2**	**41.6**	**56.2**
w/o KN-Sel.	77.8	72.3	73.1	38.6	40.5	55.6

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

Results

	Input	ER	Knowledge Snippets	NLE	Prev. SOTA NLE	Prediction
COSe	Q: People do many things to alleviate boredom. If you can't get out of the house you might decide to do what? A: a) play cards, b) skateboard, c) meet interesting people, d) listen to music	boredom, house, music	1. Music alleviates boredom 2. Music is listened at home 3. Boredom can lead to mental health problems 4. Music is relaxing	Music can alleviate boredom when you are alone at home	People listen to music	listen to music
VCR	 Q: Where are [person3] and [person2] right now? A: a) They are in a hospital room, b) They are in an empty office building, c) They are at a party, d) [person1] and [person2] are attending a formal dance	 [person2], [person3]	1. Hospital room has hospital beds 2. Hospital has nurses 3. Nurses care the patients 4. Hospital provides critical care to patients	There are hospital beds and nurses in the room	They are patients in the room	They are in a hospital room