



UNIVERSITY OF
OXFORD

Natural-XAI: Explainable AI with Natural Language Explanations



Oana-Maria Camburu

Postdoctoral Researcher

University of Oxford



Zeynep Akata

Professor of Computer Science

University of Tübingen



Scope

- This tutorial aims to give an overview of the research direction that we call *Natural-XAI*, i.e., AI systems with natural language explanations. We will *not* give a comprehensive overview of XAI in general, but there will be some introduction and discussion on general XAI.
- No pre-requirements (just basic deep learning knowledge).
- Designed for everyone: academia and industry, different modalities, and different applications.

Natural-XAI is an emerging direction, with high potential
and lots of open questions.

Part I

1. Introduction
2. The Puzzle of Natural-XAI
 - a. The Potentials
 - b. The Challenges
3. NLP Works
4. Live Q&A for Part I

Break

Part II

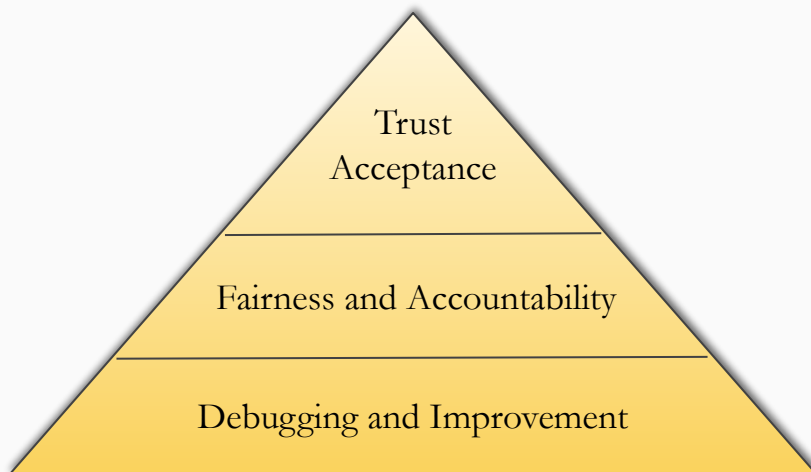
1. Explanations Advance Visual Learning
2. Computer Vision Applications
 - a. Fine-Grained Recognition
 - b. Zero-Shot Learning
 - c. Self-Driving Cars
 - d. Explanations as a means for effective communication
3. Summary and Open Questions
4. Live Q&A for Part II

Introduction

Deep neural networks have been responsible for SOTA in many areas, but are still typically black-boxes.

Even when they have high performance on test sets, they are notoriously prone to

- relying on spurious correlations in datasets (Chen et al., 2016; Gururangan et al., 2018; McCoy et al., 2019)
- adversarial attacks (Szegedy et al., 2014; Moosavi-Dezfooli et al., 2017; Jia and Liang, 2017)
- exacerbating discrimination (Bolukbasi et al., 2016; Buolamwini and Gebru, 2018)



<https://www.wired.com/2016/10/understanding-artificial-intelligence-decisions/>

- D. Chen et al., A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task, ACL, 2016.
T. McCoy et al., Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference, ACL, 2019.
S. Gururangan et al., Annotation Artifacts in Natural Language Inference Data, NAACL, 2019.
C. Szegedy et al., Intriguing Properties of Neural Networks, ICLR, 2014.
S. Moosavi-Dezfooli et al., Universal Adversarial Perturbations, CVPR, 2017.
R. Jia and P. Liang, Adversarial Examples for Evaluating Reading Comprehension Systems, EMNLP, 2017.
T. Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, NeurIPS, 2016.
J. Buolamwini and T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, FAT, 2018.

Introduction

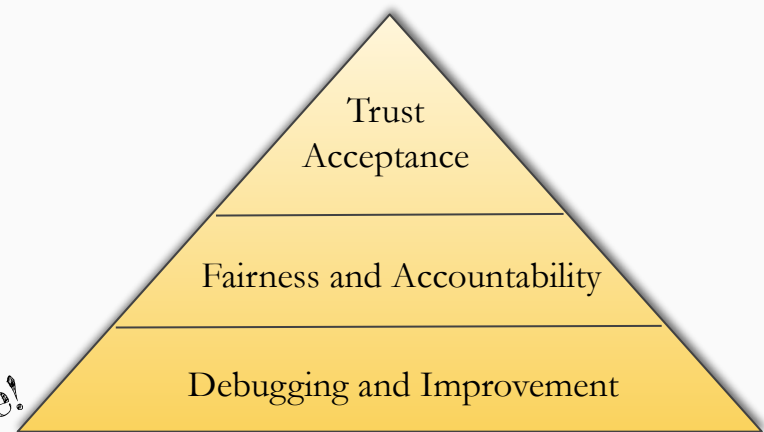
For XAI to achieve these goals, explanations should be *at least*

- audience-friendly
 - understandable
 - satisfactory
- aligned with the decision-making process of the system (faithful)

and ultimately

- allow for further interaction with the users
- lead to better AI
 - better performance
 - better decision-making process
- improve human decision-making

Not exhaustive!



Audience-friendly explanations

- Easy to understand by the target audience (e.g., lay users vs experts)
 - not all explanations in the current XAI literature are easy to understand, even for ML experts. Kaur et al. (2020): *“data scientists over-trust and misuse interpretability tools”* and *“few of our participants [197 data scientists] were able to accurately describe the visualizations output by these tools.”*
- Satisfactory: adhere to human desiderata
 - Miller (2019): *“people employ certain **biases** and **social expectations** when they generate and evaluate explanations”*. *“explanations are not just the presentation of associations and causes (causal attribution), they are **contextual**. While an event may have many causes, often the explainee cares only about a **small subset** (relevant to the context), the explainer selects a subset of this subset (based on several different criteria)”*
 - Graaf and Malle (2017): *“people will regard most autonomous intelligent systems as intentional agents and apply the conceptual framework and psychological mechanisms of human behavior explanation to them.”*

Faithfulness (alignment with the decision-making process of the system)

- Unfaithful explanations can lead to over-trusting or under-trusting a system
- Difficult to assess
- Plausibility \neq Faithfulness
 - plausibility is valuable when the explanations are used individually for assisting humans in making decisions
 - for models that generates their own explanations (the topic of this tutorial), plausibility may fairly lead to higher trustworthiness (Camburu et al., 2018)

Introduction

Interactive XAI

- Being able to interact and argue about a decision increases trust and can lead to better decisions. Wilkenfeld and Lombrozo (2015): “*explaining for the best inference*” vs “*inference to the best explanation*”, engaging in explanation even without arriving at a correct explanation can still improve one’s understanding.
- Druzdel (1996): “*The insight gained during the interaction is even more important than the actual recommendation.*”
- Arguably, a system that can interact and argue with users for the reasons behind a decision is indeed more trustworthy.

Better AI

- Humans do not learn just from labeled examples. Explanations are a valuable resource for us to understand a task and perform better at it. Heider (1958): people look for explanations to improve their understanding of someone or something so that they can derive a stable model that can be used for prediction and control.
- Explaining already trained AI systems may help us spot certain spurious correlations on which these systems rely, but there is no generic way to make the systems bypass these correlations, which is a difficult open question usually addressed via task-specific techniques (Belinkov et al., 2019).
- Can we develop models that learn from explanations for the ground-truth answers in order to arrive to correct decision-making processes?

Improve human decisions-making

- for cases where AIs are intended to assist humans in making decisions, if explanations do not help humans make better decisions then they are of little use
 - Alufaisan et al. (2020): *“any kind of AI prediction tends to improve user decision accuracy, but no conclusive evidence that explainable AI has a meaningful impact.”*; *“users were somewhat able to detect when the AI was correct versus incorrect, but this was not significantly affected by including an explanation”*.

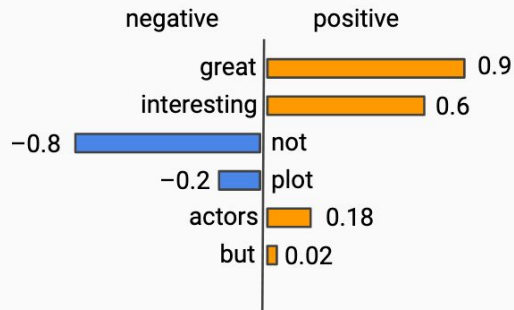
Introduction

Types of explanations

Types of explanations

1. Feature-based

“The plot was not interesting, but the actors were great.”

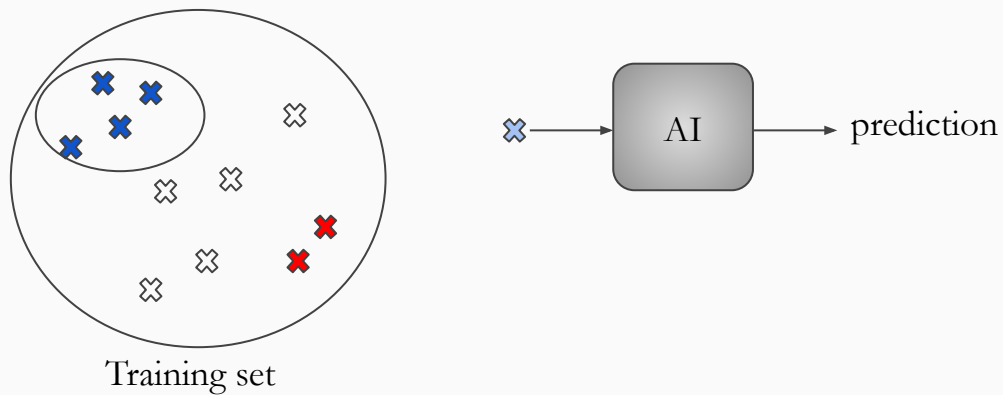


M. Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD, 2016.
S. Lundberg and S. Lee, A Unified Approach to Interpreting Model Predictions, NeurIPS, 2017.
M. Sundararajan, Axiomatic Attribution for Deep Networks, ICML, 2017.

Introduction

Types of explanations

1. Feature-based
2. Training-based



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

Introduction

Types of explanations

1. Feature-based
2. Training-based
3. Concept-based



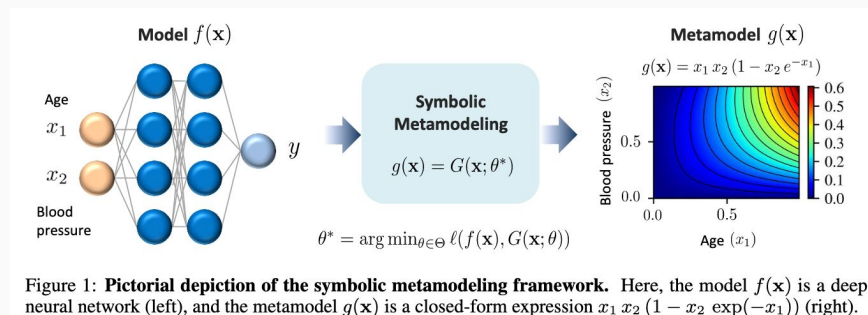
<https://medium.com/intuit-engineering/navigating-the-sea-of-explainability-f6cc4631f473>

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

Introduction

Types of explanations

1. Feature-based
2. Training-based
3. Concept-based
4. Surrogate models



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

Types of explanations

1. Feature-based
2. Training-based
3. Concept-based
4. Surrogate models
5. Natural language (In this tutorial!)
 -
 -
 -

Types of explanations

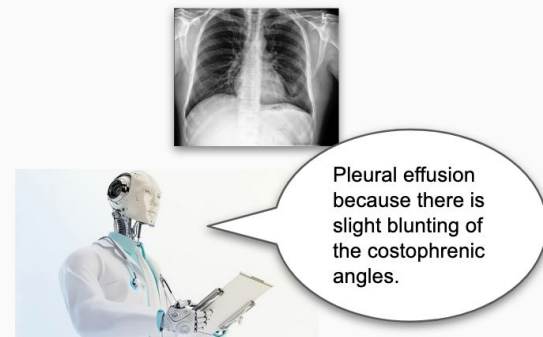
1. Feature-based
2. Training-based
3. Concept-based
4. Surrogate models
5. Natural language (In this tutorial!)
 -
 -
 -

Complementary!

AI models that

- **learn** from natural language explanations that justify the ground-truth labels
- **generate** natural language explanations for their predictions

Natural Language Explanations = NLEs



The Potential

1. Audience-friendly explanations
2. Better AI
3. Interactive XAI

Audience-friendly explanations

- NLEs have the potential to be **easy to understand by humans**.
 - Kaur et al. (2020): *“data scientists over-trust and misuse interpretability tools”* and *“few of our participants [197 data scientists] were able to accurately describe the visualizations output by these tools.”* (using feature-based explanations)
 - Alufaisan et al. (2020): *“any kind of AI prediction tends to improve user decision accuracy, but no conclusive evidence that explainable AI has a meaningful impact.”* (using feature-based explanations)
- NLEs collected from humans would, by default, encompass the **human desiderata** for explanations (**contextual**, **a small subset of arguments**, **social biases** -- Miller, 2019). Can be adapted to the **terminology** and **features** best suited to the target audience, can form a **narrative**, and express **uncertainty**.
 - Druzdzel (1996): qualitative explanation of reasoning leads to better user satisfaction and insight.

Replicate with NLEs

Better AI

- NLEs bring much more signal than a single label.
- Empirical evidence that NLEs can be a valuable signal for better model performance (Rajani et al., 2019; Atanasova et al., 2020)

Interactive XAI

- Interactive explainability could be possible with other forms of explanations, but having everything in natural language may facilitate the process



Passenger: Would you have stopped if there was no person crossing?

Car: No, because there is no traffic light at this crossover.

Passenger: OK, but would have slowed down?

Car: Yes, I always slow down before a crossover.

The Challenges

1. Faithfulness
2. Zero/Few-Shot Learning
3. Automatic Evaluation
4. Can we have NLEs for any task?

Faithfulness

- A model may learn to generate correct NLEs regardless of its inner-working for the final answer.
- Specific architectures to ensure faithfulness of the NLEs (Kumar and Talukdar, 2020).
- *Proxy* metrics for evaluating faithfulness
 - how well NLEs help an observer predict a model's output (Hase et al., 2020)
 - consistency of the NLEs (Camburu et al., 2020)

Zero/Few-Shot Learning

- NLEs are expensive and time-consuming to gather
 - although it can be done at the time of collecting labelled examples, and may even enhance the correctness of the datasets
- Novel zero/few-shot learning scenario
 - large amount of labelled examples but no/few NLEs
- Empirical evidence that zero/few-shot learning of NLEs is possible (Narang et al., 2020)

Automatic Evaluation

- Faithfulness
- Plausibility (correctness) of the generated NLEs
 - Can fairly enhance trustworthiness. Camburu et al. (2018): it is an order of magnitude more difficult for models to generate correct NLEs by relying on spurious correlations than to predict the correct labels.
 - Current automatic metrics for NLG are not reliable:
 - Camburu et al., (2018): BLEU on generated NLEs appeared better than BLEU on human-written NLEs
 - Kayser et al., (2021): comprehensive evaluation of automatic metrics vs human annotation and found little correlation. METEOR, BERTScore, and BLEURT correlate most with human scores

Can we have NLEs for any task?

- If we do not know the reasons behind a prediction, e.g., in knowledge discovery tasks, can we still get models to generate NLEs?

The Puzzle of Natural-XAI



The Puzzle of Natural-XAI



- e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)
- Make Up Your Mind! Adversarial Generation of Natural Language Explanations (Camburu et al., ACL'20)
- NILE: Natural Language Inference with Faithful Natural Language Explanations (Kumar and Talukdar, ACL'20)
- Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

e-SNLI = SNLI (Bowman et al., 2015) + human-written natural language explanations

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

e-SNLI = SNLI (Bowman et al., 2015) + human-written natural language explanations

SNLI: What is the relationship between the premise and the hypothesis? entailment, neutral, or contradiction

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

e-SNLI = SNLI (Bowman et al., 2015) + human-written natural language explanations

SNLI: What is the relationship between the premise and the hypothesis? entailment, neutral, or contradiction

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

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.
Hypothesis: A young mother is playing with her daughter in a swing.
Label: neutral
Explanation: Child does not imply daughter and woman does not imply mother.

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

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

e-SNLI

- train (~550K): 1 explanation per instance
- dev and test (~10K): 3 explanations per instance

e-SNLI: Natural Language Inference with Natural Language Explanations

(Camburu et al., NeurIPS'18)

e-SNLI

- train (~550K): 1 explanation per instance
- dev and test (~10K): 3 explanations per instance
- For quality control:
 - require annotators to highlight salient tokens
 - use the highlighted tokens in the explanation

Premise: An adult dressed in black **holds a stick**.

Hypothesis: An adult is walking away, **empty-handed**.

Label: contradiction

Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.

Hypothesis: A young **mother** is playing with her **daughter** in a swing.

Label: neutral

Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A **man** in an orange vest **leans over a pickup truck**.

Hypothesis: A man is **touching** a truck.

Label: entailment

Explanation: Man leans over a pickup truck implies that he is touching it.

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

e-SNLI

- train (~550K): 1 explanation per instance
- dev and test (~10K): 3 explanations per instance
- For quality control:
 - require annotators to highlight salient tokens
 - use the highlighted tokens in the explanation
 - in-browser checks
 - at least 3 tokens
 - not a copy of premise or hypothesis
 - highlighted at least one token
 - used at least half of highlighted tokens in the explanation
 - re-annotated trivial explanations such as *<premise> implies <hypothesis>*
 - manual annotation of 1000 samples showed ~9.6% of incorrect explanations

Premise: An adult dressed in black holds a stick.

Hypothesis: An adult is walking away, empty-handed.

Label: contradiction

Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.

Hypothesis: A young mother is playing with her daughter in a swing.

Label: neutral

Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A man in an orange vest leans over a pickup truck.

Hypothesis: A man is touching a truck.

Label: entailment

Explanation: Man leans over a pickup truck implies that he is touching it.

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

e-SNLI

- train (~550K): 1 explanation per instance
- dev and test (~10K): 3 explanations per instance
- For quality control:
 - require annotators to highlight salient tokens
 - use the highlighted tokens in the explanation
 - in-browser checks
 - at least 3 tokens
 - not a copy of premise or hypothesis
 - highlighted at least one token
 - used at least half of highlighted tokens in the explanation
 - re-annotated trivial explanations such as *<premise> implies <hypothesis>*
 - manual annotation of 1000 samples showed ~9.6% of incorrect explanations

Premise: An adult dressed in black holds a stick.

Hypothesis: An adult is walking away, empty-handed.

Label: contradiction

Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.

Hypothesis: A young mother is playing with her daughter in a swing.

Label: neutral

Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A man in an orange vest leans over a pickup truck.

Hypothesis: A man is touching a truck.

Label: entailment

Explanation: Man leans over a pickup truck implies that he is touching it.

Publicly available:

<https://github.com/OanaMariaCamburu/e-SNLI>

Experiments

- I. Premise agnostic
- II. Full model
 - A. Predict then Explain
 - B. Explain then Predict
 - 1. Seq2Seq
 - 2. Attention
- III. Out-of-domain transfer

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

Premise agnostic

Gururangan et al. (2018): Hypothesis \rightarrow Label : 67% accuracy due to artifacts in SNLI

- correlations between tokens in hypotheses and labels:
 - “tall”, “sad” \rightarrow neutral, “animal”, “outside” \rightarrow entailment, “sleeping”, negations \rightarrow contradiction
- sentence length

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

Premise agnostic

Gururangan et al. (2018): Hypothesis \rightarrow Label : 67% accuracy due to artifacts in SNLI

- correlations between tokens in hypotheses and labels:
 - “tall”, “sad” \rightarrow neutral, “animal”, “outside” \rightarrow entailment, “sleeping”, negations \rightarrow contradiction
- sentence length

Our experiment

Hypothesis \rightarrow Label : 66% correct*

Hypothesis \rightarrow Explanation : 6% correct**

*in the first 100 instances in the test set **manual annotation over the first 100 instances in the test set

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

Premise agnostic

Gururangan et al. (2018): Hypothesis \rightarrow Label : 67% accuracy due to artifacts in SNLI

- correlations between tokens in hypotheses and labels:
 - “tall”, “sad” \rightarrow neutral, “animal”, “outside” \rightarrow entailment, “sleeping”, negations \rightarrow contradiction
- sentence length

Our experiment

Hypothesis \rightarrow Label : 66% correct*	} 10x more difficult to rely on spurious correlation to generate correct explanations than to produce correct labels
Hypothesis \rightarrow Explanation : 6% correct**	

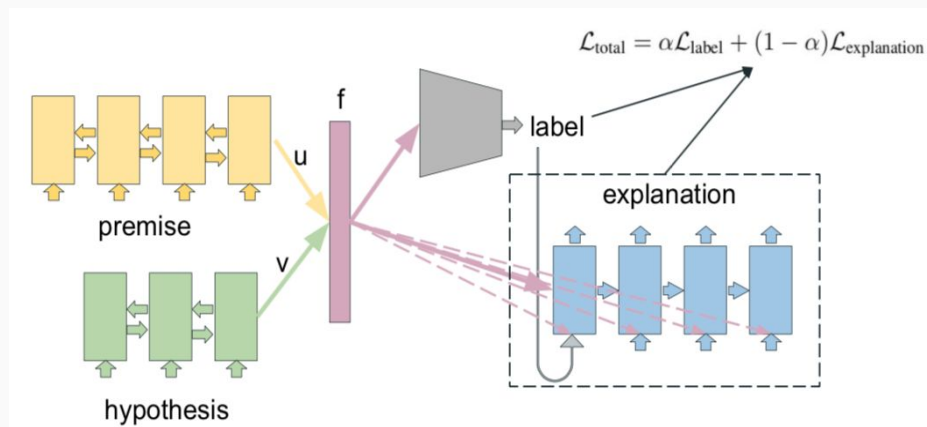
*in the first 100 instances in the test set **manual annotation over the first 100 instances in the test set

e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

Predict then Explain (BiLSTM-Max-PredExpl)

Generate the explanation conditioned on the predicted label

$$\mathbf{f} = [\mathbf{u}, \mathbf{v}, |\mathbf{u} - \mathbf{v}|, \mathbf{u} \odot \mathbf{v}]$$

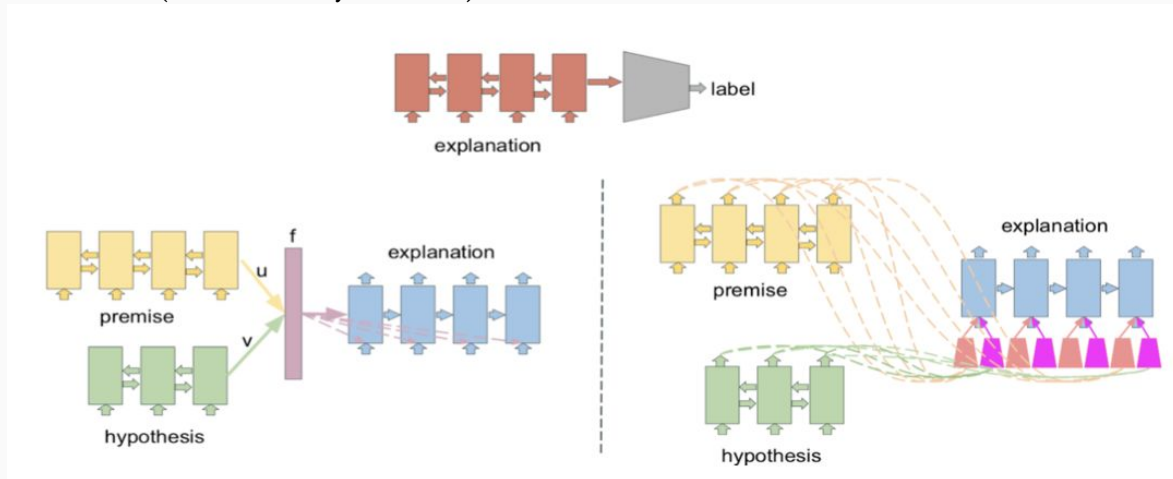


e-SNLI: Natural Language Inference with Natural Language Explanations

(Camburu et al., NeurIPS'18)

Explain then Predict (BiLSTM-Max-ExplPred)

- (premise, hypothesis) \rightarrow explanation
 - Seq2Seq (BiLSTM-Max-ExpPred-Seq2Seq)
 - Seq2Seq-Attention (BiLSTM-Max-ExplPred-Att)
- explanation \rightarrow label (test accuracy 96.83%)



e-SNLI: Natural Language Inference with Natural Language Explanations (Camburu et al., NeurIPS'18)

Model	Label Accuracy	Perplexity	BLEU	Expl@100
BiLSTM-MAX	84.01 (0.25)	-	-	-
BiLSTM-MAX-PREDEXPL	83.96 (0.26)	10.58 (0.40)	22.40 (0.70)	34.68
BiLSTM-MAX-EXPLPRED-SEQ2SEQ	81.59 (0.45)	8.95 (0.03)	24.14 (0.58)	49.8
BiLSTM-MAX-EXPLPRED-ATT	81.71 (0.36)	6.1 (0.00)	27.58 (0.47)	64.27

Inter-annotator BLEU: 22.51

(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) PREDICTED LABEL: neutral
EXPLANATION: Just because the men are in the middle of a street doesn't mean they are posing for a picture. [0]

(b) PREDICTED LABEL: entailment
EXPLANATION: three young men are people. [0.33]

(c) 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. [0]

(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) PREDICTED LABEL: contradiction
EXPLANATION: The firefighters can not be putting out a fire station and putting out a fire at the same time. [0]

(b) 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]

(c) 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) PREDICTED LABEL: contradiction
EXPLANATION: A man is not a woman. [1]

(b) PREDICTED LABEL: contradiction
EXPLANATION: One can not be looking and eating simultaneously. [0]

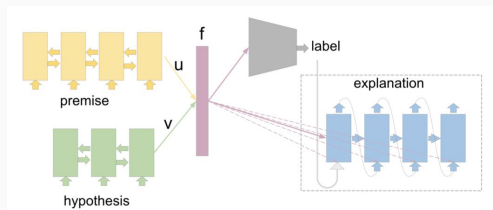
(c) PREDICTED LABEL: contradiction
EXPLANATION: A person can not be looking at a medical and a book at the same time. [0]

e-SNLI: Natural Language Inference with Natural Language Explanations

(Camburu et al., NeurIPS'18)

Out-of-domain transfer

- SICK-E (Marelli et al., 2014)
- MultiNLI (Williams et al., 2018)



Model	SICK-E acc/expl@100	MultiNLI acc/expl@100
BiLSTM-MAX	53.27 (1.65) / -	57 (0.41) / -
BiLSTM-MAX-AUTOENC	52.9 (1.77) / -	55.38 (0.9) / -
BiLSTM-MAX-PREDEXPL	53.54 (1.43) / 30.64	57.16 (0.51) / 1.92

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

(Camburu et al., ACL'20)

Are natural language self-generated explanations faithfully describing the decision-making processes of the model?

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

(Camburu et al., ACL'20)

Are natural language self-generated explanations faithfully describing the decision-making processes of the model?

As a **proxy** to answer this question, we can look at whether models generate inconsistent explanations.

Definition: *Two explanations are **inconsistent** if they provide logically contradictory arguments.*

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

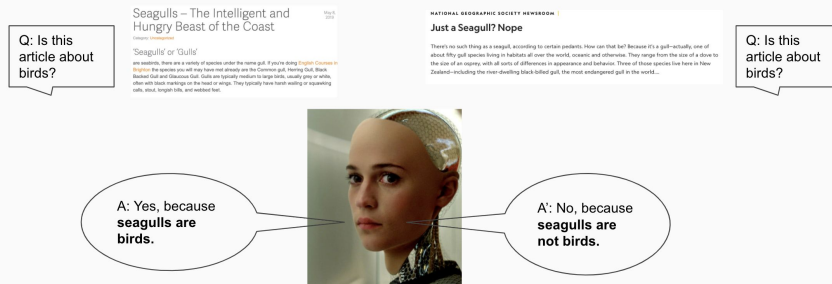
(Camburu et al., ACL'20)

Examples of inconsistent explanations

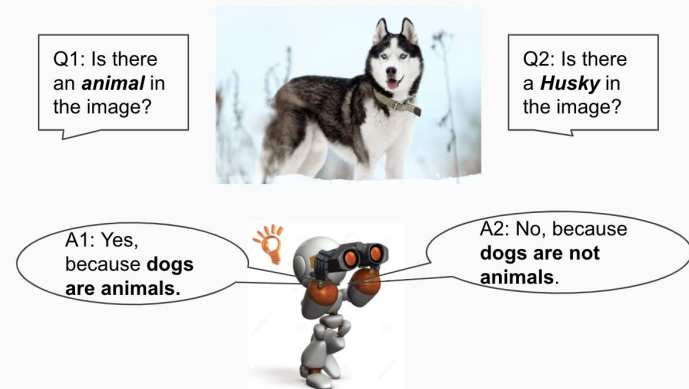
Self-Driving Cars



Question Answering



Visual Question Answering



Recommender Systems



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

(Camburu et al., ACL'20)

A model providing **inconsistent explanations** can have either 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.

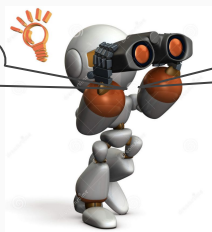
Q: Is there an **animal** in the image?



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

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

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



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

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

(Camburu et al., ACL'20)

Goal: Checking if models are robust against generating inconsistent natural language explanations.

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

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

High-level Approach

- (A) For an instance x and the explanations $e_m(x)$, create a list of explanations that are inconsistent with $e_m(x)$.
- (B) For an inconsistent explanation i_e created at step (A) find an input x' such that $e_m(x') = i_e$.

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

(Camburu et al., ACL'20)

Context-free vs. Context-dependent Inconsistencies

Context-free: inconsistency no matter what input, e.g., explanations formed by pure background knowledge.

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**.

Inconsistent

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

(Camburu et al., ACL'20)

Context-free vs. Context-dependent Inconsistencies

Context-free: inconsistency no matter what input, e.g., explanations formed by pure background knowledge.

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**.

Inconsistent

Context-dependent: inconsistency depends on parts of the input.

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

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

(Camburu et al., ACL'20)

Context-free vs. Context-dependent Inconsistencies

Context-free: inconsistency no matter what input, e.g., explanations formed by pure background knowledge.

Context-dependent: inconsistency depends on parts of the input.

Q: Is there an animal in the image?



Q': Is there a Husky in the image?

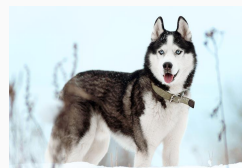
A: Yes, because **dogs are animals**.



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

Inconsistent

Q: Is there an animal in the image?



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



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

NOT Inconsistent



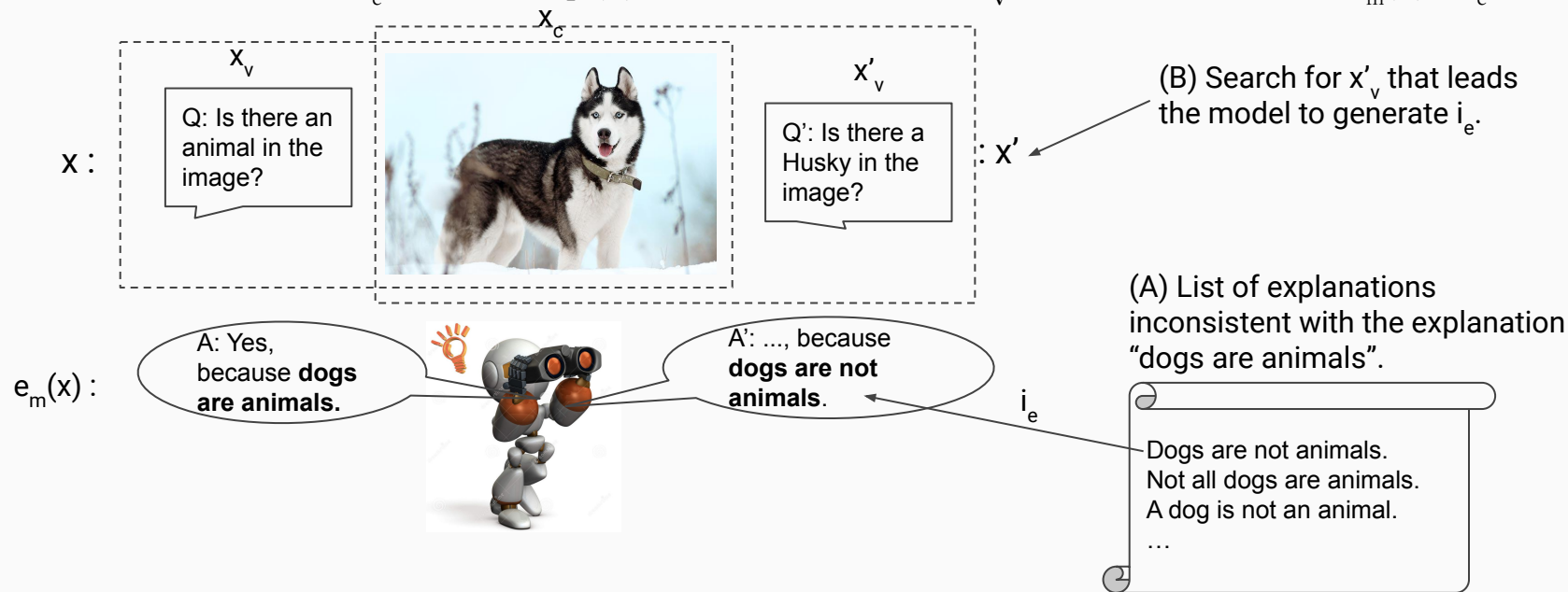
Q': Is there a Husky in the image?

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

(Camburu et al., ACL'20)

High-level Approach

- (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 $e_m(x') = i_e$.

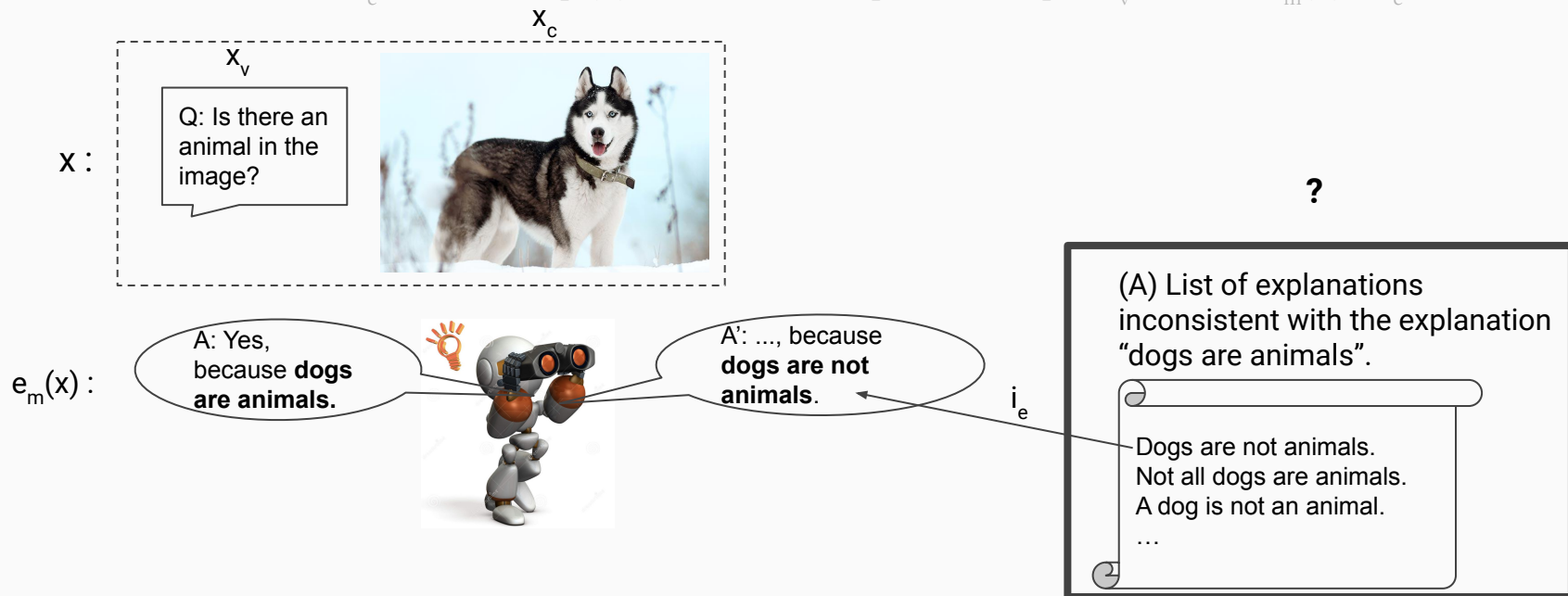


Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

(Camburu et al., ACL'20)

High-level Approach

- (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 $e_m(x') = i_e$.



Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)

High-level Approach

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

For a given task, one may define **a set of logical rules** to transform an explanation into an inconsistent counterpart:

1. **Negation:** “*A dog is an animal.*” \iff “*A dog is not an animal.*”
2. **Task-specific antonyms:** “*The car continues because it is green light.*” \iff “*The car continues because it is red light.*”
3. **Swap explanations of mutually exclusive labels:**

Recommender(movie X, user U) = **No** because “*X is a horror.*” \iff Recommender(movie Z, user U) = **No** because “*Z is a comedy.*”

Recommender(movie Y, user U) = **Yes** because “*Z is a comedy.*” \iff Recommender(movie K, user U) = **Yes** because “*K is a horror.*”

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

(Camburu et al., ACL'20)

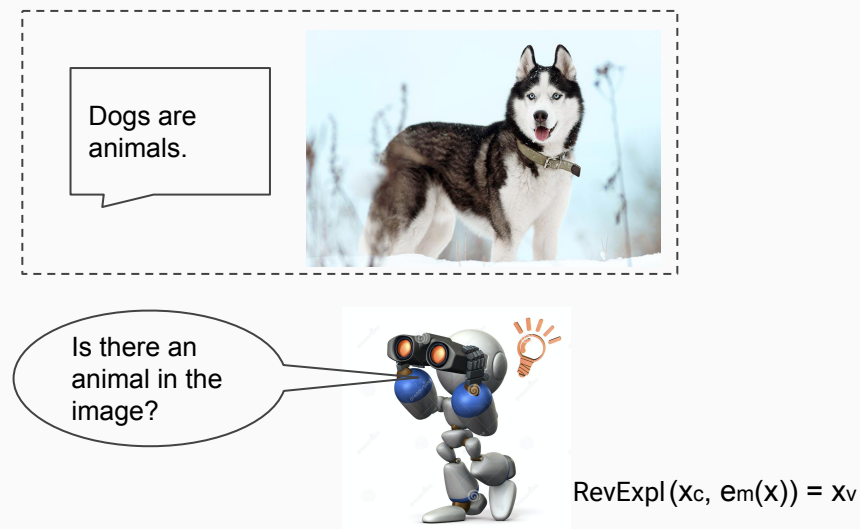
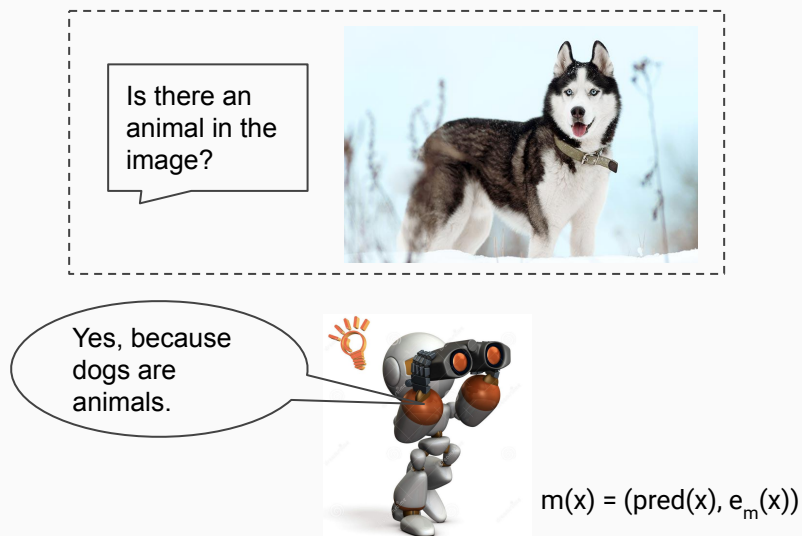
High-level Approach

- (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 $e_m(x') = i_e$.

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)

High-level Approach

- (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 a model, RevExpl, to go from an explanation $e_m(x)$ to the input that caused m to generate the explanation.



Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)

Approach

- I. Train $\text{RevExpl}(x_c, e_m(x)) = x_v$
- II. For each explanation $e = e_m(x)$:
 - a) Create a list of statements that are inconsistent with e , call it I_e
 - by using logic rules: negation, task-specific antonyms, and swapping between explanations for mutually exclusive labels
 - b) For each e' in I_e , query RevExpl to get the variable part of a reverse input: $x'_v = \text{RevExpl}(x_c, e')$
 - 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

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)

High-level Approach

- (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 an input x' such that $e_m(x') = i_e$.

Novel Adversarial Setup

- 1) No predefined adversarial targets (label attacks do not have this issue).
- 2) At step (B), the model has to generate a **full target sequence**: the goal is to generate the exact explanation that was identified at step (A) as inconsistent with the explanation $e_m(x)$. Current attacks focus on the presence/absence of a very small number of tokens in the target sequence (Cheng et al., 2020, Zhao et al., 2018).
- 3) Adversarial inputs x' do not have to be a paraphrase or a small perturbation of the original input (can happen as a byproduct). Current works focus on adversaries being paraphrases or a minor deviation from the original input (Belinkov and Bisk, 2018).

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)

e-SNLI

$x = (\text{premise}, \text{hypothesis})$. 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 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).

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)

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: $e_m(x) = \text{“Dog is a type of animal.”}$ matches the entailment template “ X is a type of Y ” with $X = \text{“dog”}$ and $Y = \text{“animal”}$. 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

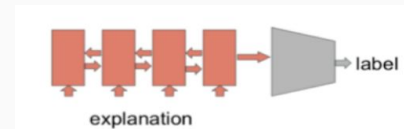
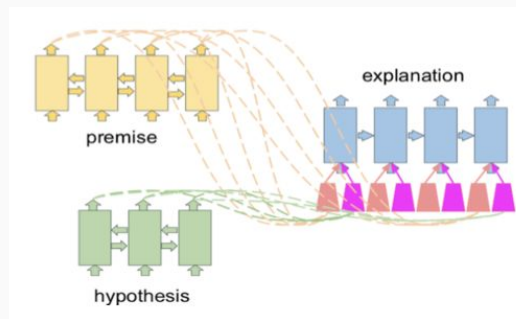
- *cannot be dog and animal at the same time*
- *dog is not animal*
- *dog is the opposite of animal*
- *it is either dog or animal*
- ...

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

(Camburu et al., ACL'20)

BiLSTM-Max-ExplPred-Att model

- 64.27% correct explanations



- $\text{RevExpl}(\text{premise}, \text{explanation}) = \text{hypothesis}$
 - same architecture as ExplainThenPredict-Att
 - 32.78% test accuracy (exact string match for the generated hypothesis)
- Manual annotation of 100 random reverse hypothesis gives 82% to be realistic
 - majority of unrealistic are due to repetition of a token
- Success rate of our adversarial method for finding inconsistencies $\sim 4.51\%$ on the e-SNLI test set
 - ~ 443 distinct pairs of inconsistent explanations

Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)


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

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

(Camburu et al., ACL'20)


Manual scanning had no success

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



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.



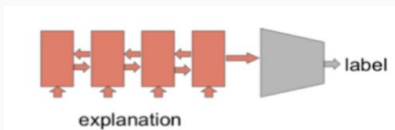
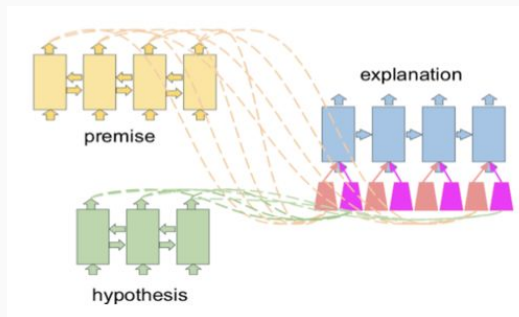
P: A small **child** watches the outside world through a window.
H: A small **toddler** watches the outside world through a window.
E: Not every child is a toddler.

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.

NILE : Natural Language Inference with Faithful Natural Language Explanations

(Kumar and Talukdar, ACL'20)

Can we build systems for which we can probe the faithfulness of the generated NLEs?



- The form of the explanation is enough to get predict the label, likely undermining faithfulness.
- How can we probe faithfulness?

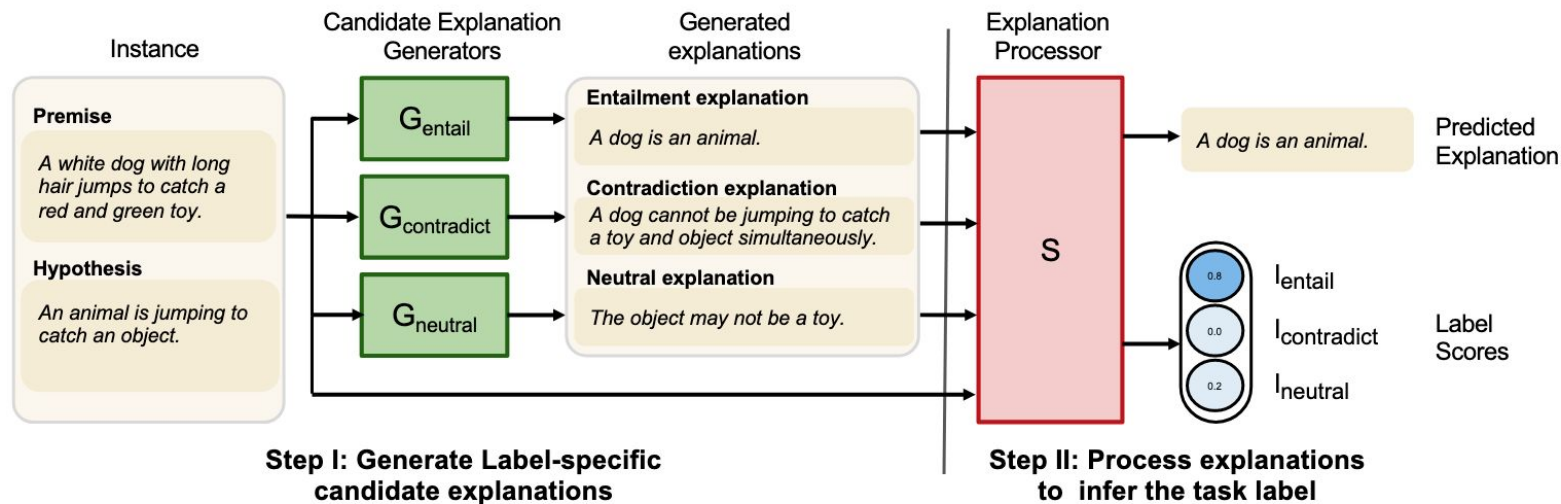
NILE : Natural Language Inference with Faithful Natural Language Explanations (Kumar and Talukdar, ACL'20)

Can we build systems for which we can probe the faithfulness of the generated NLEs?

NILE : Natural Language Inference with Faithful Natural Language Explanations

(Kumar and Talukdar, ACL'20)

Can we build systems for which we can probe the faithfulness of the generated NLEs?



NILE : Natural Language Inference with Faithful Natural Language Explanations

(Kumar and Talukdar, ACL'20)

- Measuring faithfulness by perturbing the input to the explanation processor
 - comprehensiveness (what happens when we remove the explanation from the input)
 - sufficiency (what happens if we keep only the explanations)
 - shuffling (explanation is replaced by a randomly selected explanation of the same label)
- NILE-NS: negative explanations for an instance, of the same form as the correct label

Model		I+ Exp	I only	Exp only
NILE-NS	Independent	91.6	33.8	69.4
	Aggregate	91.6	33.8	74.5
	Append	91.7	91.2	72.9
NILE	Independent	91.3	33.8	46.1
	Aggregate	91.2	33.8	40.7

Table 3: *Estimating the sensitivity of the system's predictions to input explanations through erasure.*

Model		Dev Set	Shuffled Dev Set
NILE-NS	Independent	91.6	88.1
	Aggregate	91.6	89.6
	Append	91.7	88.5
NILE	Independent	91.3	35.3
	Aggregate	91.2	31.6

Table 4: *Probing the sensitivity of the system's predictions by shuffling instance-explanation pairs.*

Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

How can we tackle the lack of commonsense knowledge in current AIs generating NLEs?

Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

How can we tackle the lack of commonsense knowledge in current AIs generating NLEs?

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.**

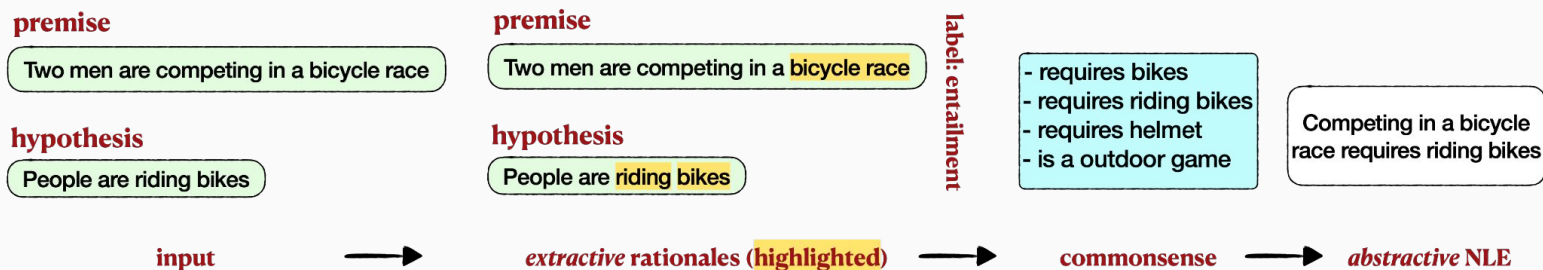
Camburu et al., 2020



Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

How can we tackle the lack of commonsense knowledge in current AIs generating NLEs?



Rationale-Inspired Natural Language **Ex**planations with **C**ommonsense

REXC

Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

RExC

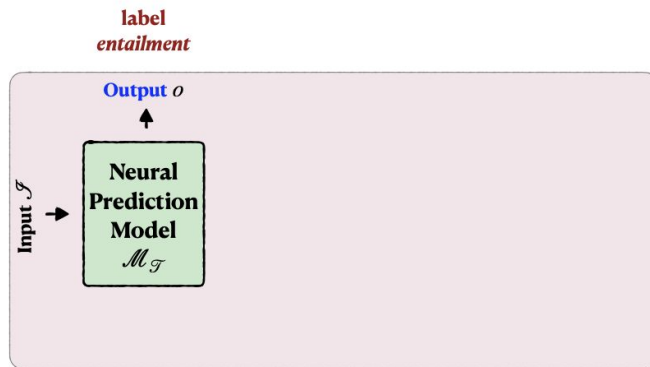
Input \mathcal{I} is passed
to Neural Prediction
Model $\mathcal{M}_{\mathcal{I}}$, to
obtain output o

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes



Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

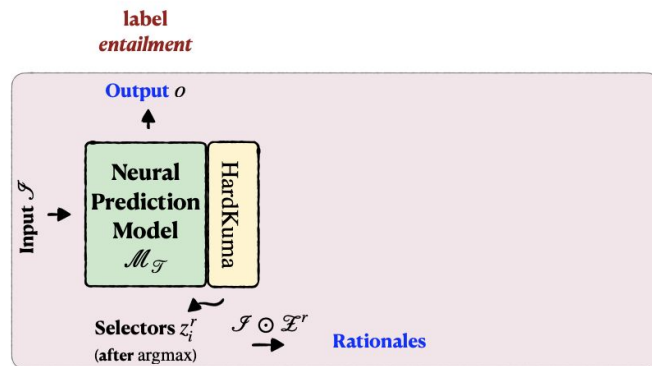
RExC

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes



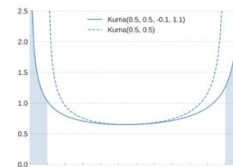
premise

Two men are competing in a **bicycle race**

hypothesis

People are **riding bikes**

A series of binary latent variables z_i^r are used to discretely select parts of the input as *rationales*



Bastings et al., 2020

L_1
regularization
for sparsity

X

Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

RExC

- requires bikes
- requires riding bikes
- requires helmet
- is a outdoor game

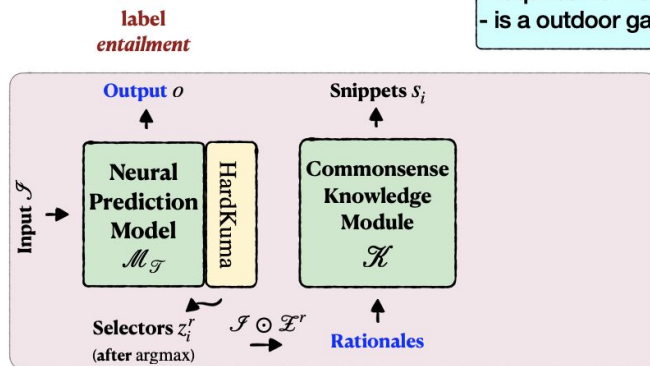
Each lexical unit from rationales are sent to the commonsense module \mathcal{K} , that result in knowledge snippets s_i

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes



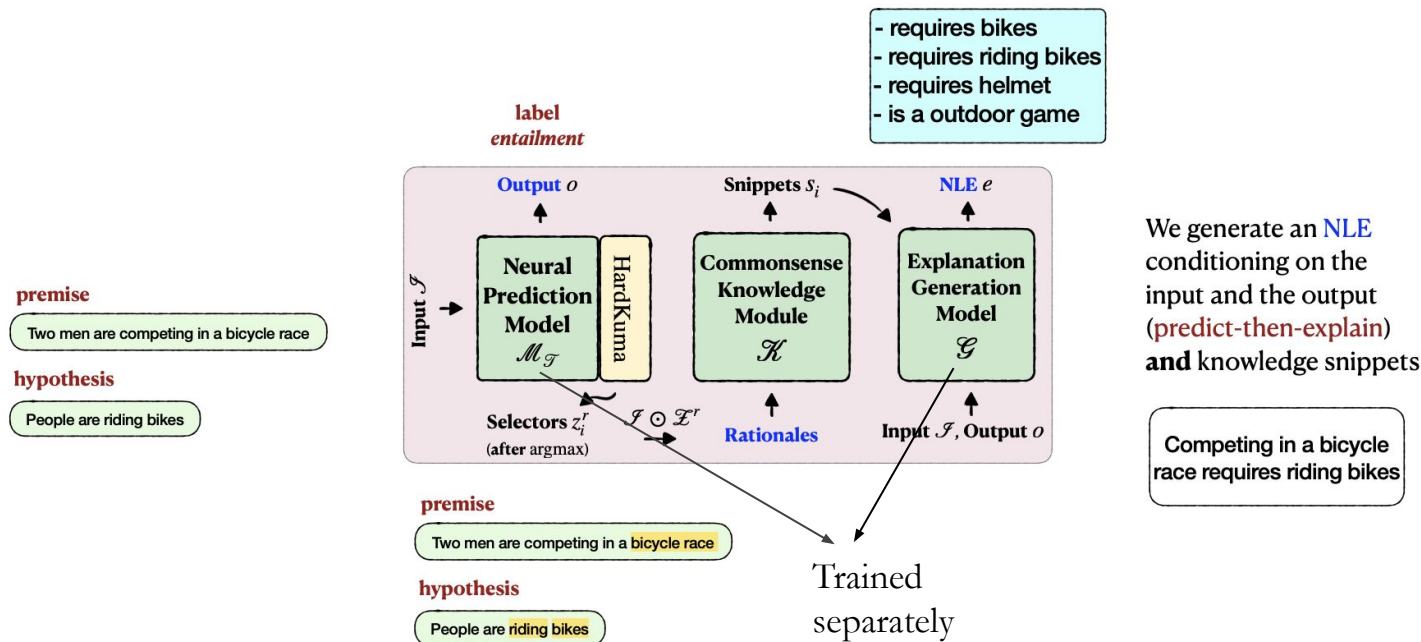
premise

Two men are competing in a bicycle race

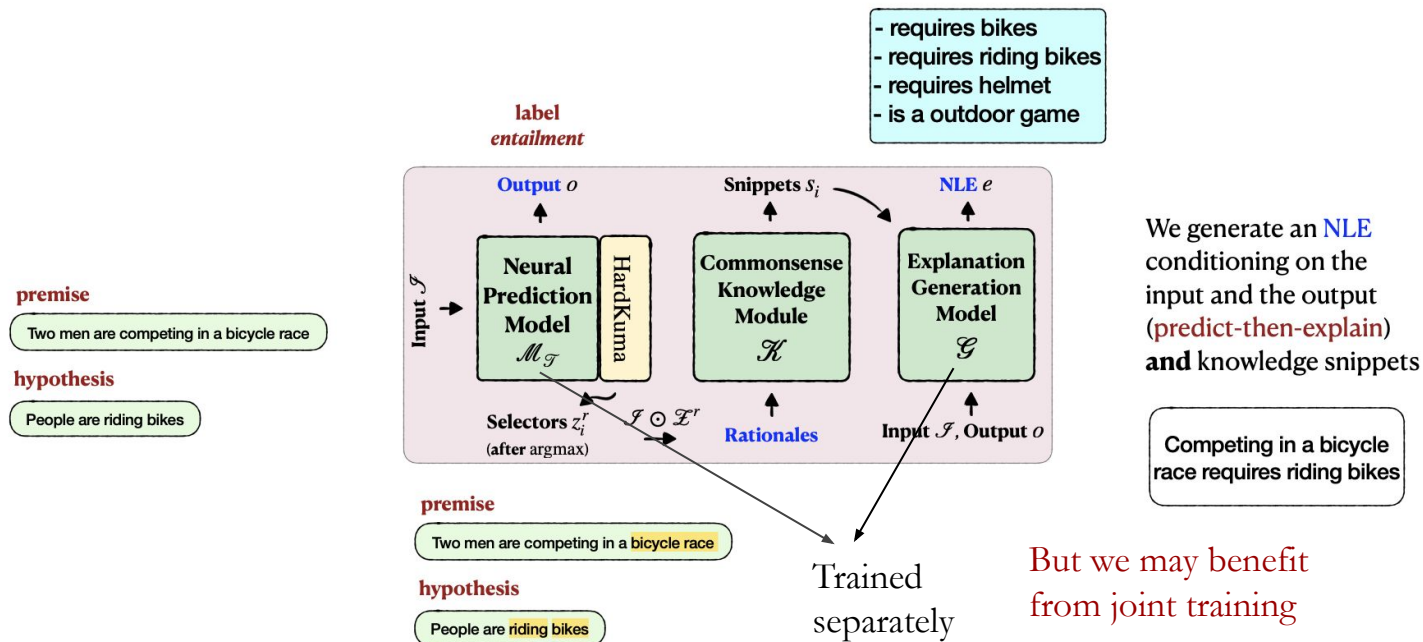
hypothesis

People are riding bikes

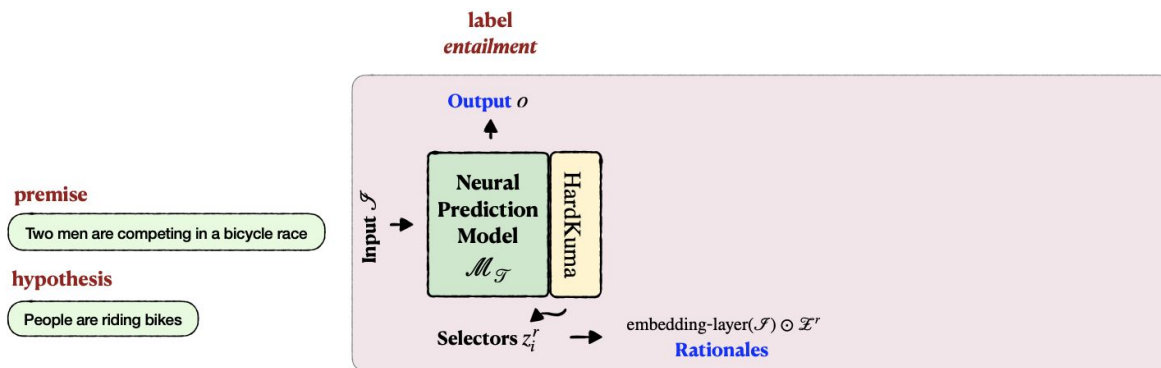
RExC — Modular



RExC — Modular



RExC

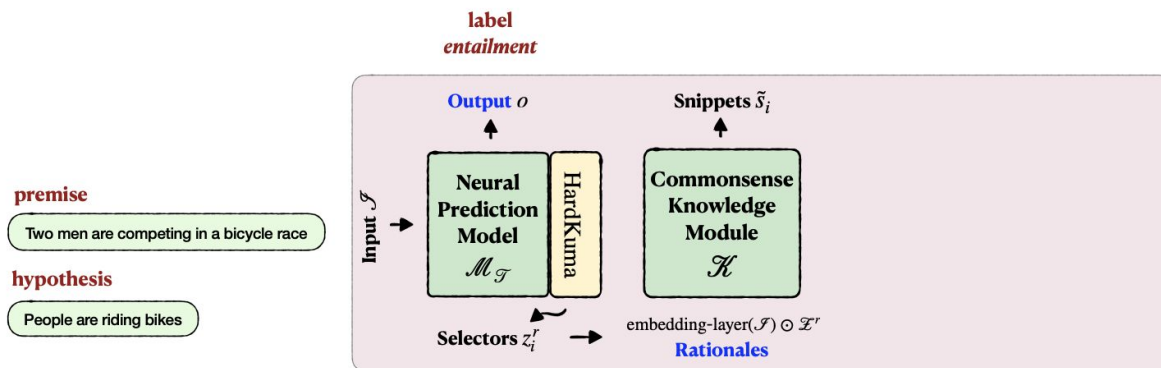


The series of binary latent variables z_i^r are used as masks on the embedded input

Rationale-Inspired Natural Language Explanations with Commonsense

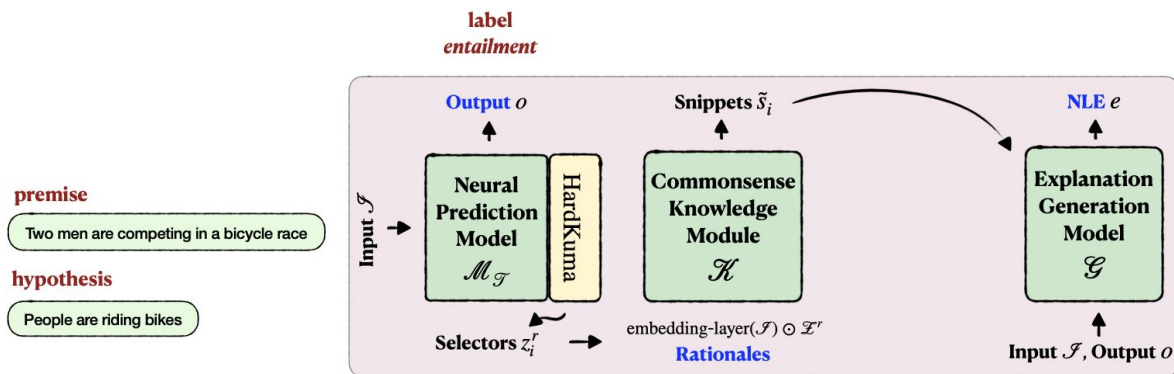
(Majumder et al., 2021)

RExC



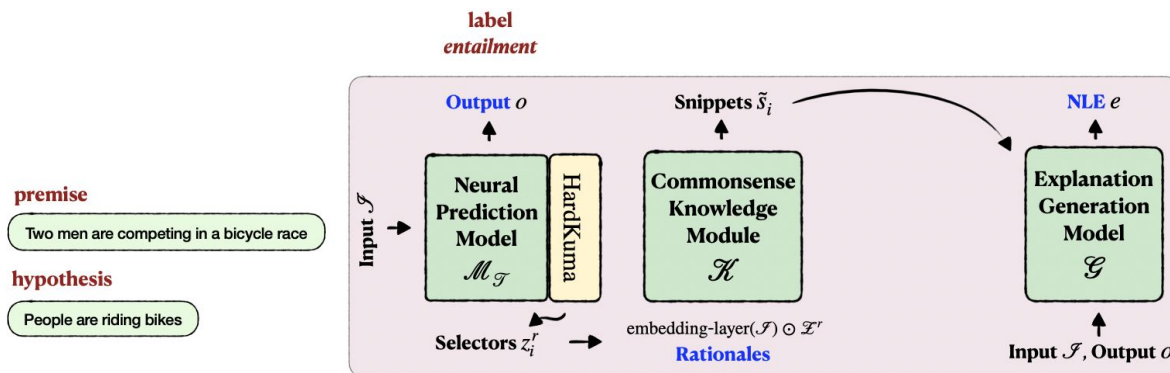
... and directly sent to a generative commonsense module \mathcal{K} , mirroring the modular approach

RExC — E2E



... and directly sent to a
generative commonsense
module \mathcal{K} , mirroring the
modular approach

RExC — E2E



... and directly sent to a
generative commonsense
module \mathcal{K} , mirroring the
modular approach

But we may benefit
from doing a selection
of the snippets

Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

RExC

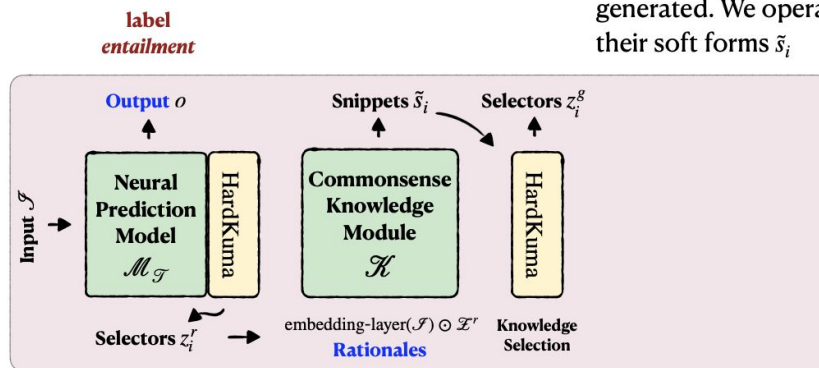
Another series of HardKuma variables are used to sample from all knowledge snippets generated. We operate on their soft forms \tilde{s}_i

premise

Two men are competing in a bicycle race

hypothesis

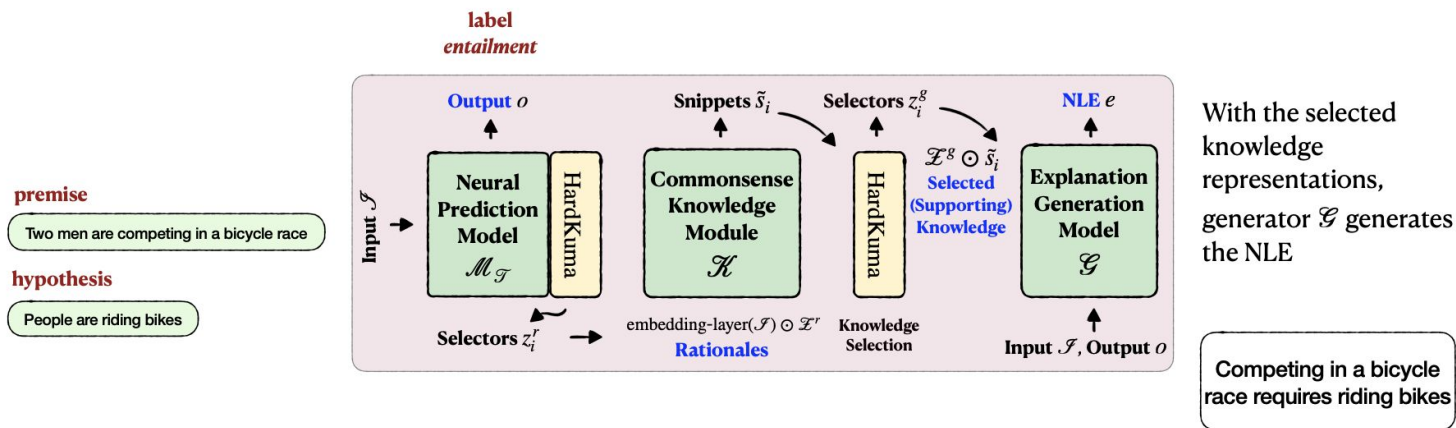
People are riding bikes



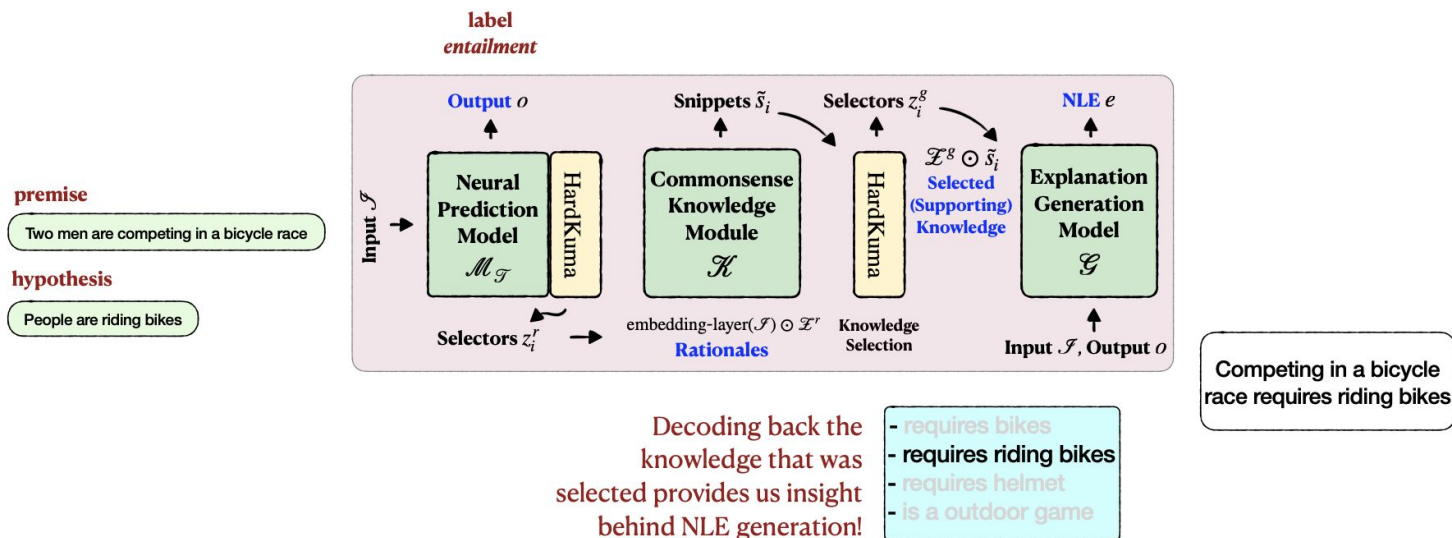
Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

RExC

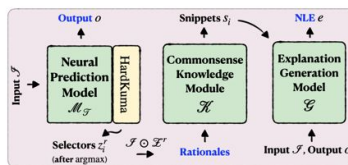


RExC — KS



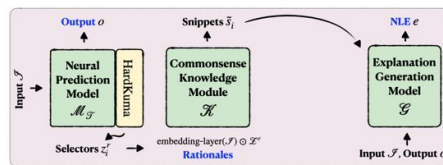
Variants of RExC

RExC-Mod



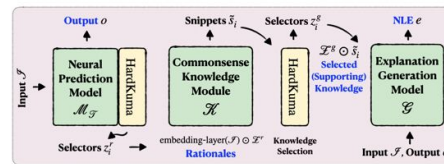
Modular, **separate**
training for rationales
and NLEs

RExC-E2E



End-to-end, **joint**
training for rationales
and NLEs

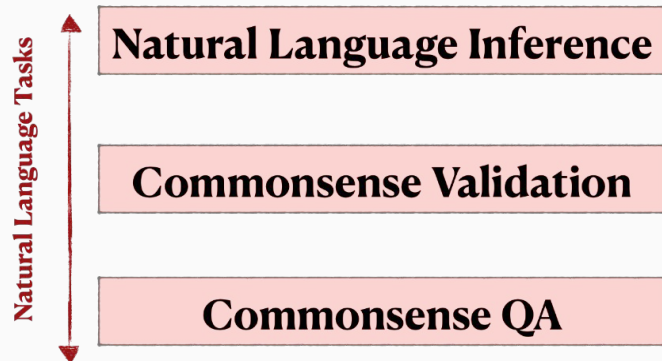
RExC-KS



RExC-KS+

Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)



premise Two men are competing in a bicycle race

hypothesis People are riding bikes

A: Coffee stimulates people
B: Coffee depresses people

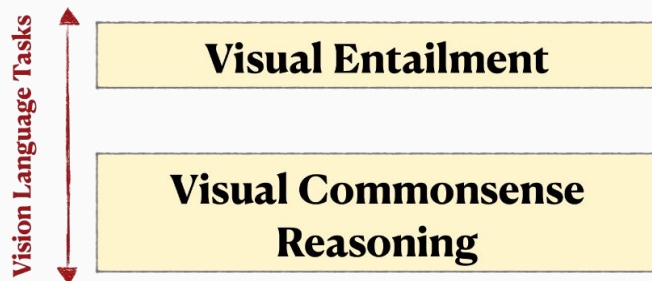
Q: Where does a wild bird usually live?

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

label
entailment e-SNLI
(Camburu et al., 2018)

label
B is invalid ComVE
(Wang et al., 2019)

label
sky CoS-E
(Rajani et al., 2019)



Hypothesis:
Some tennis players pose

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



Q: What is the place?

label
They are in a hospital room VCR
(Zellers et al., 2019)

Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

NLP Tasks

 $\mathcal{M}_{\mathcal{T}}$

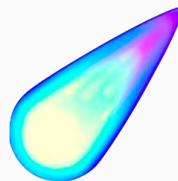
BART: a Seq2Seq
pretrained transformer with
a MLP prediction head

(Lewis et al., 2020)

 \mathcal{K}

COMET: Commonsense
Transformer trained on
ConceptNet

(Bosselut et al., 2019)

 \mathcal{G}

BART: a Seq2Seq
pretrained transformer with
a Language Model head



Rationale-Inspired Natural Language Explanations with Commonsense

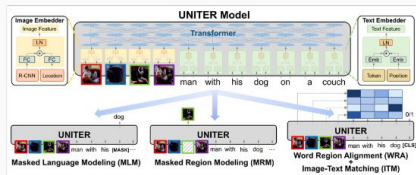
(Majumder et al., 2021)

Vision-Language Tasks

 $\mathcal{M}_{\mathcal{T}}$

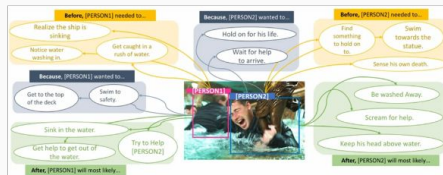
UNITER: a Seq2Seq
pretrained transformer for
text and images with a MLP
prediction head

(Chen et al., 2020)

 \mathcal{K}

Visual-COMET:
Commonsense Transformer
trained on Visual
Commonsense Graph

(Park et al., 2020)

 \mathcal{G}

GPT2: a pretrained
transformer-based
Language Model

(Radford et al., 2020)



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

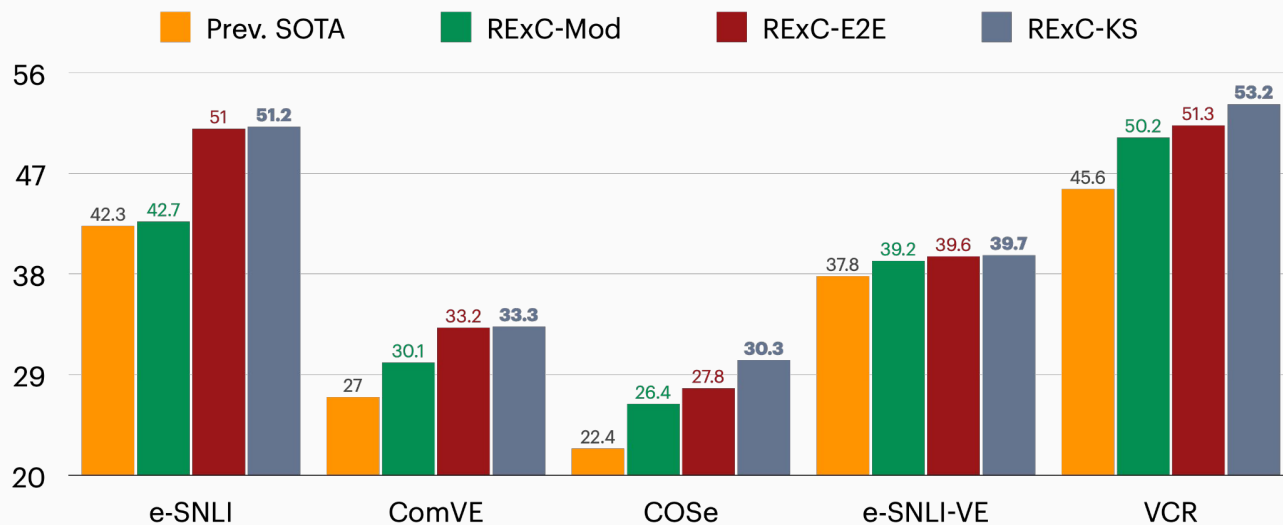
J. Park et al., VisualCOMET: Reasoning about the dynamic context of a still image. ECCV, 2020.

A. Radford et al., Language Models are Unsupervised Multitask Learners, 2019.

Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

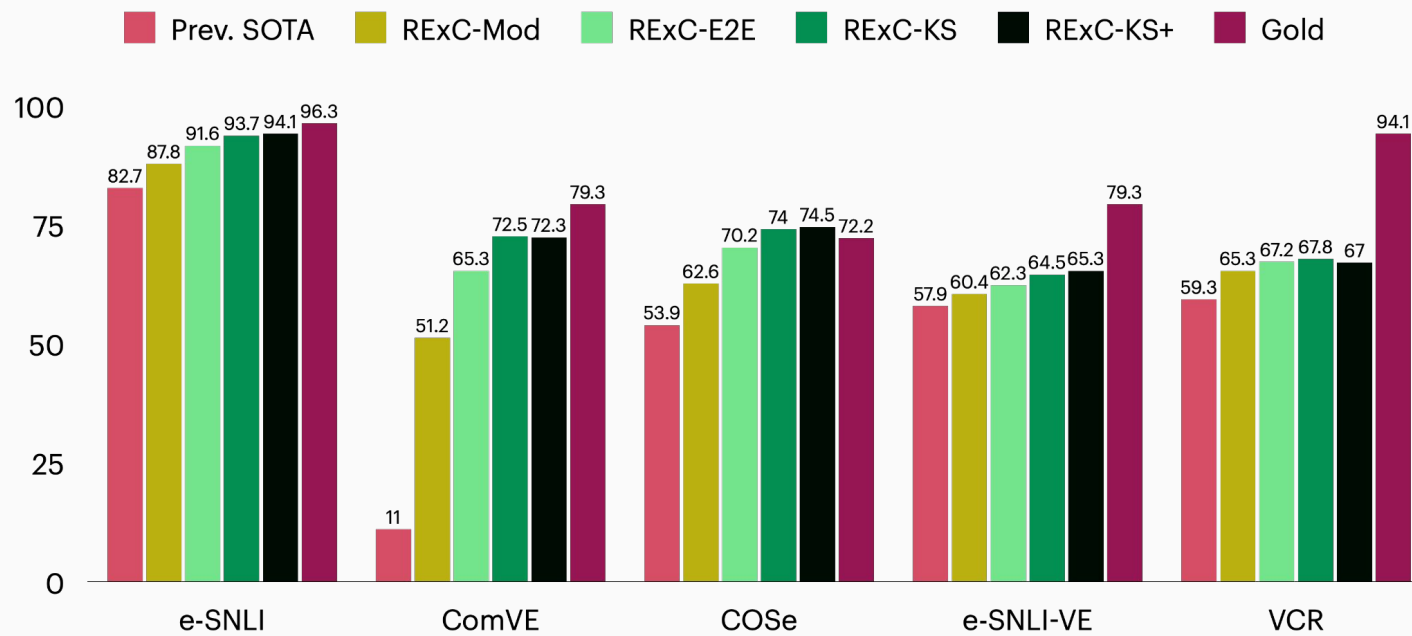
BLEURT (Sellam et al., 2020)



Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)

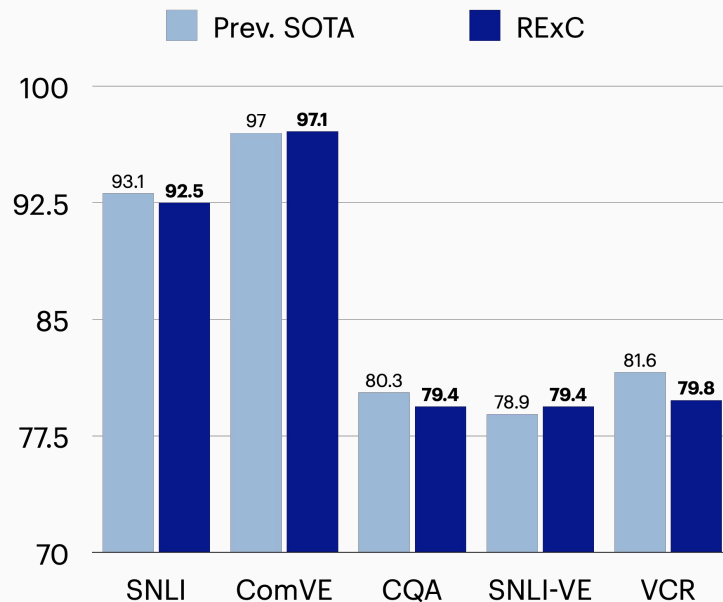
Human evaluation



Rationale-Inspired Natural Language Explanations with Commonsense

(Majumder et al., 2021)


Task performance



Summary Part 1



Thank you!

 @oanacamb

Questions?