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.

Message

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

Outline

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

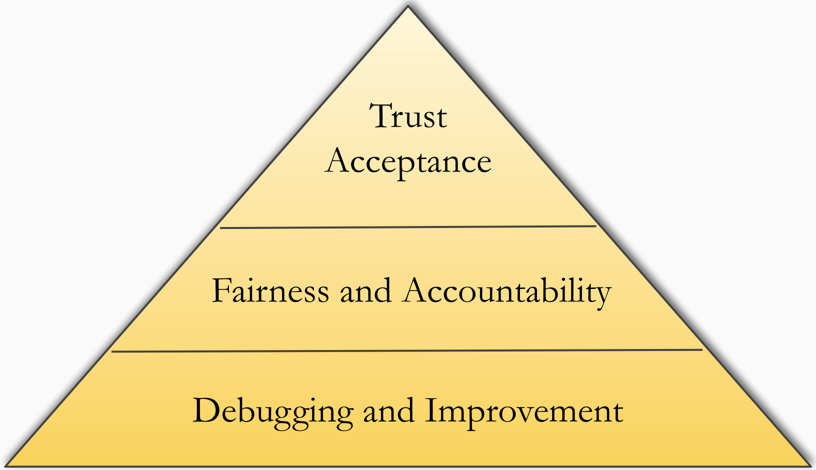
1. Explanations Advance Visual Learning
2. Computer Vision Applications
    a. Fine-Grained Recognition
    b. Zero-Shot Learning
    c. Self-Driving Cars
    d. Explanations as a means for effective communication
3. Summary and Open Questions
4. Live Q&A for Part II

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

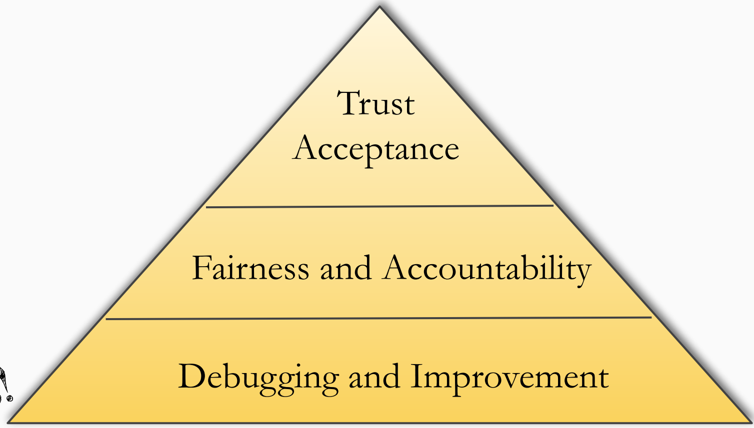
For XAI to achieve these goals, explanations should be *at least*

- 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!



Introduction

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."*

H. Kaur et al. ‚Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning, CHI 2020.
T. Miller, Explanation in Artificial Intelligence:Insights from the Social Sciences, Elsevier, 2019.
M. de Graaf, B. Malle, How People Explain Action (and Autonomous Intelligent Systems Should Too), in: AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction, 2017.

Introduction

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 ≠ 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)

O. Camburu et al., e-SNLI: Natural Language Inference with Natural Language Explanations, NeurIPS, 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.

- Druzdzel (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.

D. A. Wilkenfeld, T. Lombrozo, Inference to the best explanation (IBE) versus explaining for the best inference (EBI), Science & Education, 2015.
M. Druzdzel, Qualitative Verbal Explanations in Bayesian Belief Networks, Artificial Intelligence and Simulation of Behaviour Quarterly, Special issue on Bayesian belief networks, 1996.

Introduction

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?

F. Heider, The psychology of interpersonal relations, New York: Wiley, 1958
Y. Belinkov et al., Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference, ACL, 2019.

Introduction

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"*.

Y. Alufaisan et al., Does Explainable Artificial Intelligence Improve Human Decision-Making?, AAAI 2021

Introduction

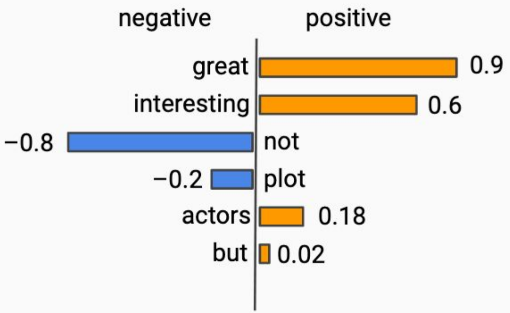
Types of explanations

Introduction

Types of explanations

1. Feature-based

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

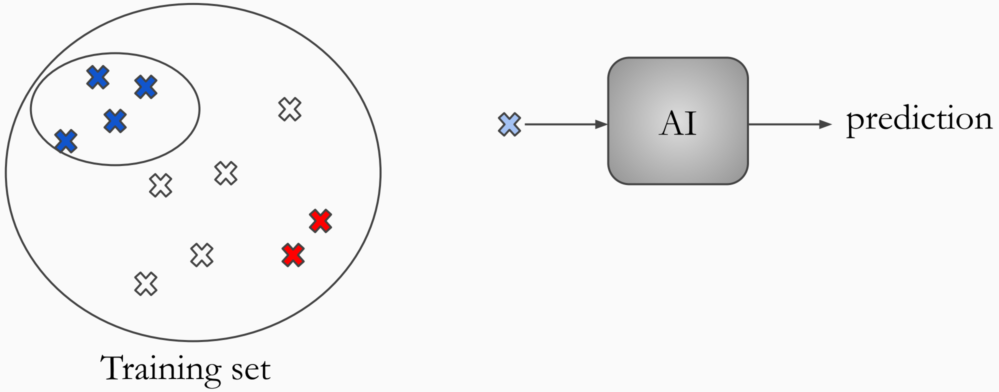


M. Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD, 2016.
S. Lundberg and S. Lee, A Unified Approach to Interpreting Model Predictions, NeurIPS, 2017.
M. Sundararajan, Axiomatic Attribution for Deep Networks, ICML, 2017.

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

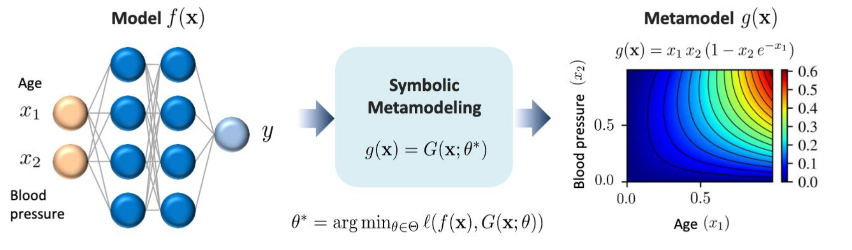


Figure 1: **Pictorial depiction of the symbolic metamodeling framework.** Here, the model $f(\mathbf{x})$ is a deep neural network (left), and the metamodel $g(\mathbf{x})$ is a closed-form expression $x_1\,x_2\,(1 - x_2\,\exp(-x_1))$ (right).

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

Introduction

Types of explanations

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

.

.

.

Introduction

Types of explanations

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

.

.

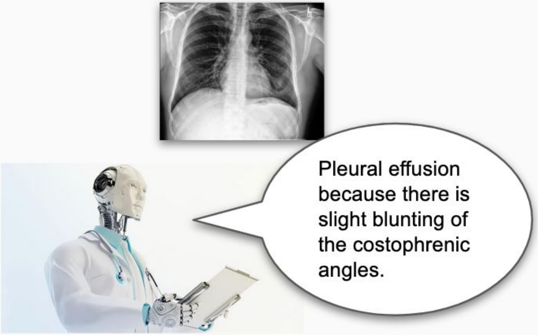
.

Complementary!

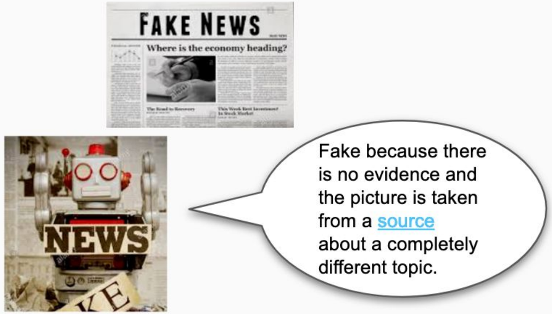
Natural-XAI

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



Natural-XAI

The Potential

1. Audience-friendly explanations

2. Better AI

3. Interactive XAI

Natural-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)

Replicate with NLEs

- 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.

H. Kaur et al., Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.
T. Miller, Explanation in Artificial Intelligence:Insights from the Social Sciences, Elsevier, 2019.
M. Druzdzel, Qualitative Verbal Explanations in Bayesian Belief Networks, Artificial Intelligence and Simulation of Behaviour Quarterly, Special issue on Bayesian belief networks, 1996.

Natural-XAI

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)

N. Rajani et al., Explain Yourself! Leveraging Language Models for Commonsense Reasoning, ACL, 2019.
P. Atanasova et al., Generating Fact Checking Explanations, ACL, 2020.

Natural-XAI

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?

Natural-XAI

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)

S. Kumar and P. Talukdar, NILE: Natural Language Inference with Faithful Natural Language Explanations, ACL, 2020.
P. Hase et al., Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?, ACL, 2020.
O. Camburu et al., Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations, ACL, 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)

S. Narang et al., WT5?! Training Text-to-Text Models to Explain their Predictions, 2020

Natural-XAI

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

O. Camburu et al., e-SNLI: Natural Language Inference with Natural Language Explanations, NeurIPS, 2018.
M. Kayser et al., e-ViL: A Dataset and Benchmark for Natural Language Explanations inVision-Language Tasks, 2021.

Natural-XAI

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

Natural-XAI

The Puzzle of Natural-XAI

NLP Applications

- 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

S. Bowman et al., A large annotated corpus for learning natural language inference, EMNLP, 2015.

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.

S. Bowman et al., A large annotated corpus for learning natural language inference, EMNLP, 2015.

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

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

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 → Label : 67% accuracy due to artifacts in SNLI

- correlations between tokens in hypotheses and labels:
  - "tall", "sad" → neutral, "animal", "outside" → entailment, "sleeping", negations → contradiction
- sentence length

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

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

Premise agnostic

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

- correlations between tokens in hypotheses and labels:
  - "tall", "sad" → neutral, "animal", "outside" → entailment, "sleeping", negations → contradiction
- sentence length

Our experiment

Hypothesis → Label : 66% correct*

Hypothesis → Explanation : 6% correct**

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

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

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

Premise agnostic

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

- correlations between tokens in hypotheses and labels:
  - "tall", "sad" → neutral, "animal", "outside" → entailment, "sleeping", negations → contradiction
- sentence length

Our experiment

Hypothesis → Label : 66% correct*

Hypothesis → Explanation : 6% correct**

10x more difficult to rely on spurious correlation to generate correct explanations than to produce correct labels

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

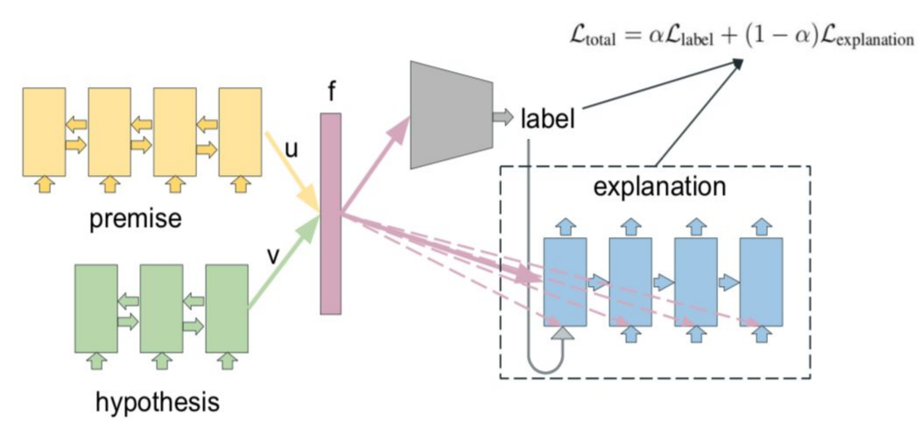
S. Gururangan et al., Annotation Artifacts in Natural Language Inference Data, NAACL, 2019.

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

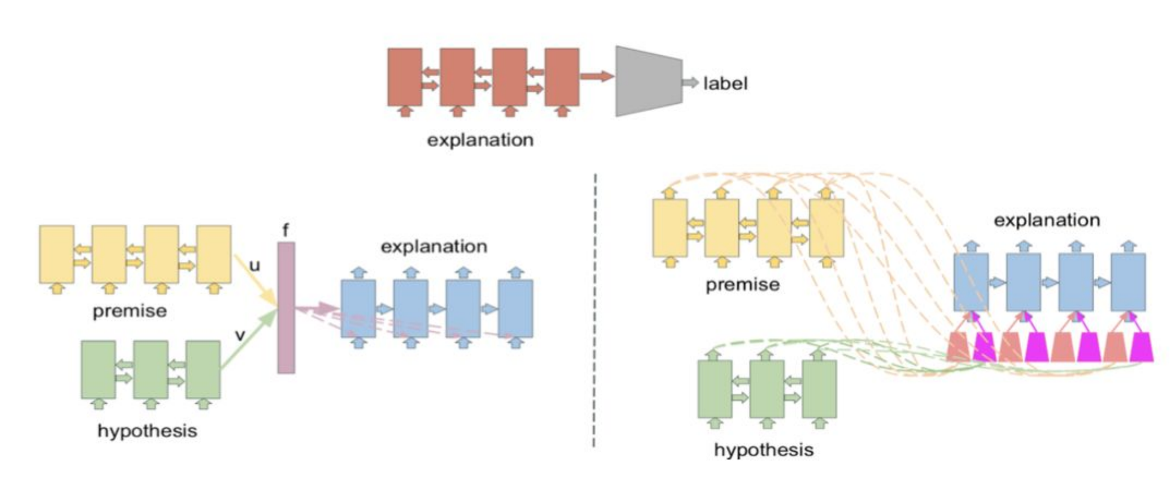
$$f = [u, v, |u - v|, u \circledast v]$$



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

Explain then Predict (BiLSTM-Max-ExplPred)

- (premise, hypothesis) → explanation
  - Seq2Seq (BiLSTM-Max-ExpPred-Seq2Seq)
  - Seq2Seq-Attention (BiLSTM-Max-ExplPred-Att)
- explanation → 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)	(b)	(c)
(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)	(b)	(c)
(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]

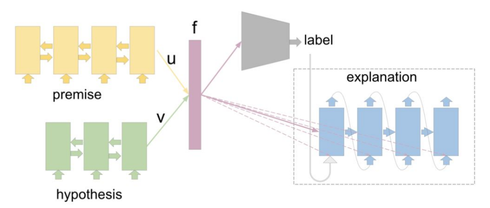
(3) Premise: A blond-haired doctor and her African American assistant looking threw new medical manuals.
Hypothesis: A man is eating pb and j.
Gold label: contradiction

(a)	(b)	(c)
(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

A. Williams et al., A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, NAACL, 2018.
M. Marelli et al., A SICK cure for the evaluation of compositional distributional semantic models, LREC, 2014.

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

Q: Why are you stopping?

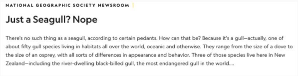
Q': Why are you stopping? ***There is no one crossing.***

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

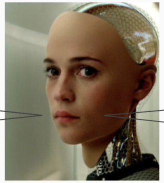
A': I stopped because **there is no one crossing.**

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?

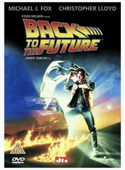
Q2: Is there a ***Husky*** in the image?

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.**

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

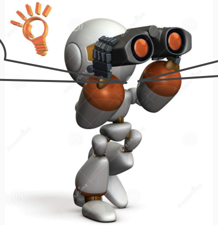
A model providing **inconsistent explanations** can have either of the **two undesired behaviours**:

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?

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 (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.

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)

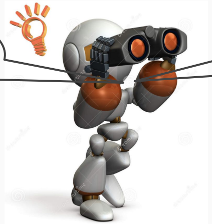
Context-free vs. Context-dependent Inconsistencies

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**.

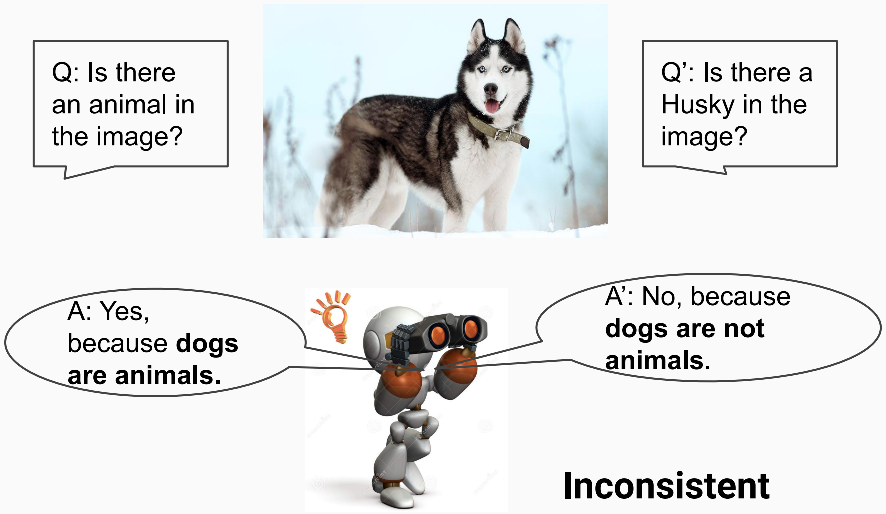
Inconsistent

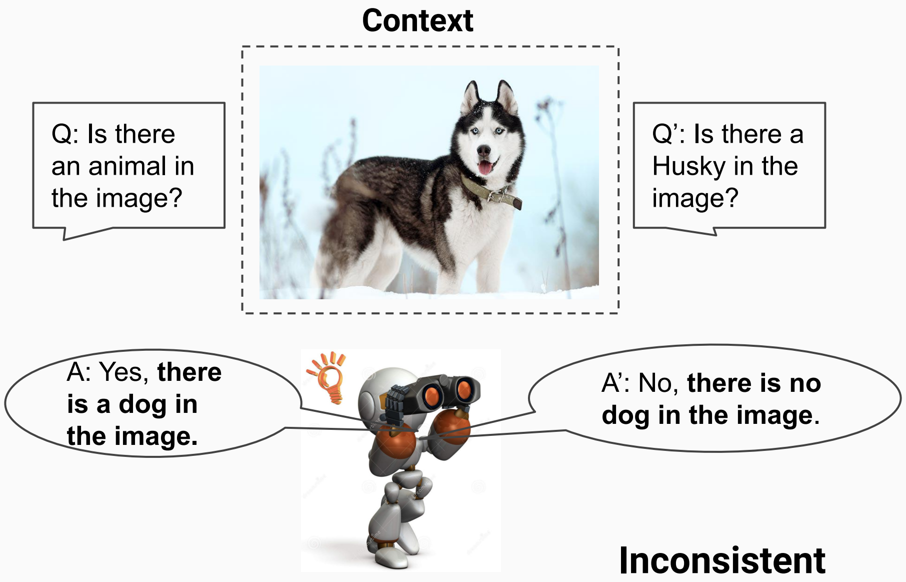
Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)

Context-free vs. Context-dependent Inconsistencies

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

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





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

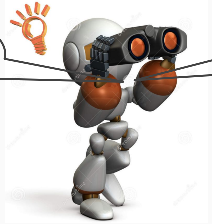
Context-free vs. Context-dependent Inconsistencies

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.**

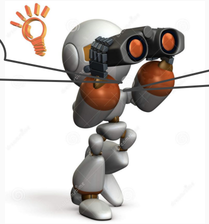
Inconsistent

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.**

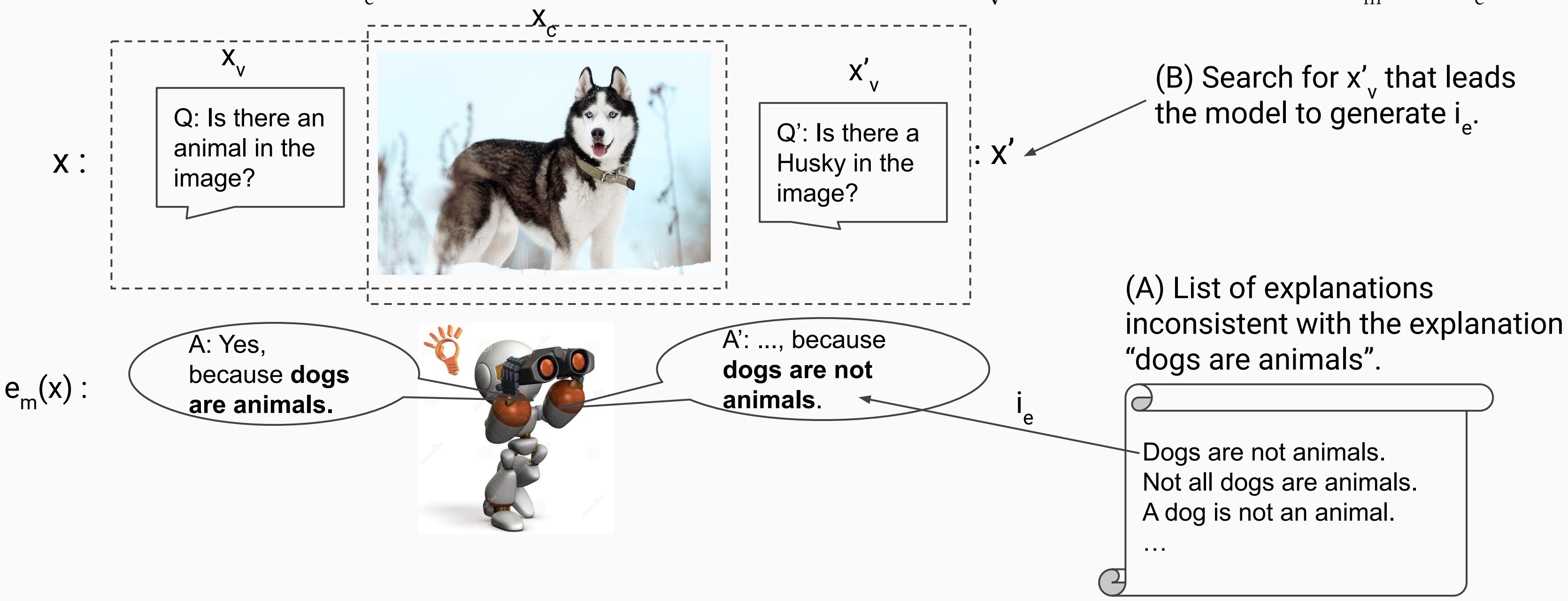
NOT Inconsistent

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 $e_m(x') = i_e$.

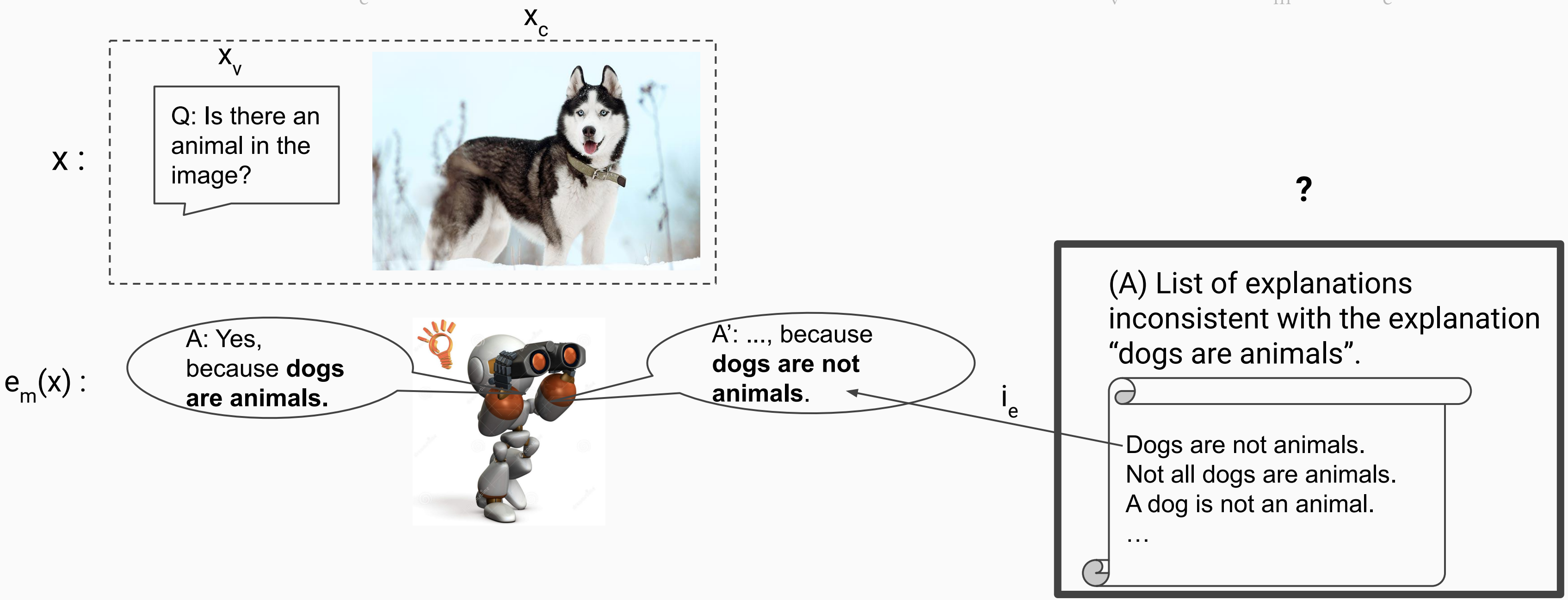


Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (Camburu et al., ACL'20)

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.*" ⟺ "*A dog is not an animal.*"
2. **Task-specific antonyms**: "*The car continues because it is green light.*" ⟺ "*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*." ⟺ Recommender(movie Z, user U) = **No** because "*Z is a comedy*."

   Recommender(movie Y, user U) = **Yes** because "*Z is a comedy*." ⟺ 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 $e_m(x)$.

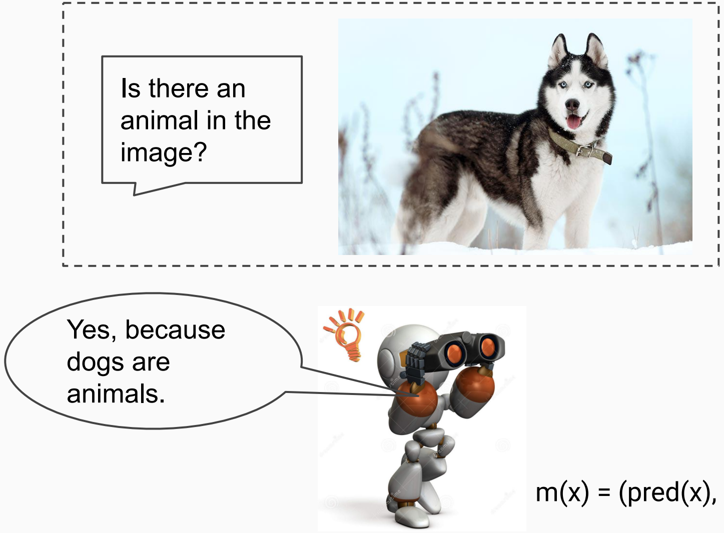
(B) For an inconsistent statement i_e created at step (A), find the variable part of an input x'_v such that $e_m(x') = i_e$.

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

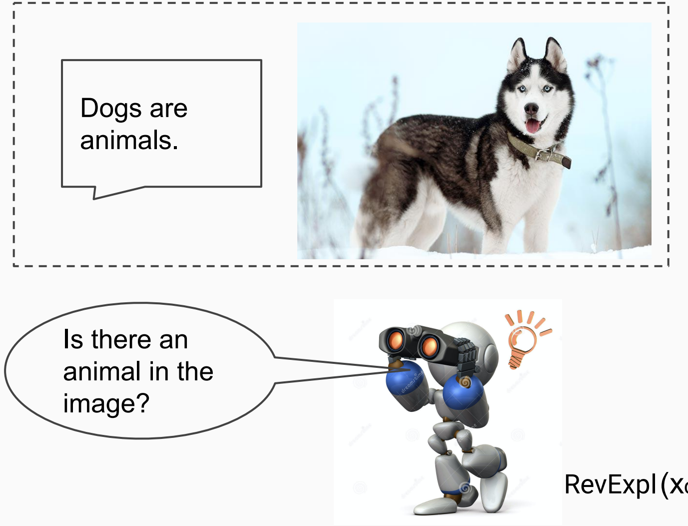
High-level Approach

(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.



$m(x) = (pred(x), e_m(x))$



$RevExpl(x_c, e_m(x)) = x_v$

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

Approach

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

- 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 = 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 = (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 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: $e_m(x)$ = *"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

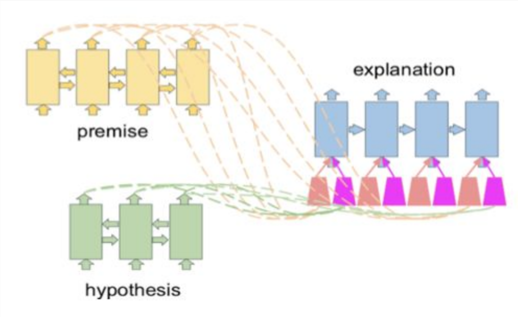
- *cannot be dog and animal at the same time*
- *dog is not animal*
- *dog is the opposite of animal*
- *it is either dog or animal*

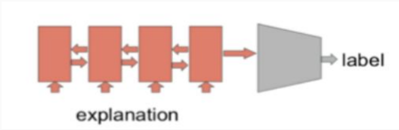
. . .

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

BiLSTM-Max-ExplPred-Att model

- 64.27% correct explanations





- RevExpl(premise, explanation) = 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 ~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. 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: A man talks to two guards as he holds a drink.	
Original Hypothesis: The prisoner is talking to two guards in the prison cafeteria. Predicted Label: neutral Original explanation: **The man is not necessarily a prisoner.**	Reverse Hypothesis: A prisoner talks to two guards. Predicted Label: entailment 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. Predicted label: contradiction Original explanation: **There are either two women and a man or three women.**	Reverse Hypothesis: Three women are sitting down eating. Predicted label: neutral Reverse explanation: **Two women and a man are three women.**
Premise: Biker riding through the forest.	
Original Hypothesis: Man riding motorcycle on highway. Predicted Label: contradiction Original explanation: **Biker and man are different.**	Reverse Hypothesis: A man rides his bike through the forest. Predicted label: entailment Reverse explanation: **A biker is a man.**
Premise: A hockey player in helmet.	
Original Hypothesis: They are playing hockey Predicted Label: entailment Original explanation: **A hockey player in helmet is playing hockey.**	Reverse Hypothesis: A man is playing hockey. Predicted Label: neutral 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. Predicted label: entailment Original explanation: **A woman is a female.**	Reverse Hypothesis:The students are all female. Predicted label: neutral 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. 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.**
Premise: A family walking with a soldier.	
Original Hypothesis: A group of people strolling. Predicted label: entailment Original explanation: **A family is a group of people.**	Reverse Hypothesis: A group of people walking down a street. Predicted label: contradiction 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.

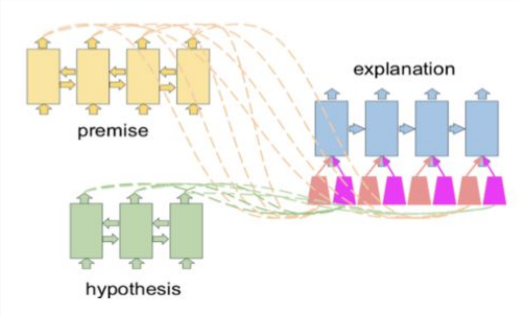
P: A small **child** watches the outside world through a window.
H: A small **toddler** watches the outside world through a window.
E: Not every child is a toddler.

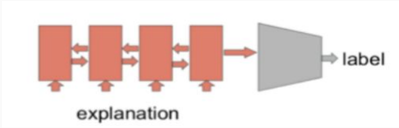
P: A small **toddler** watches the outside world through a window.
H: A small **child** watches the outside world through a window.
E: A toddler is a small child.

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

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?





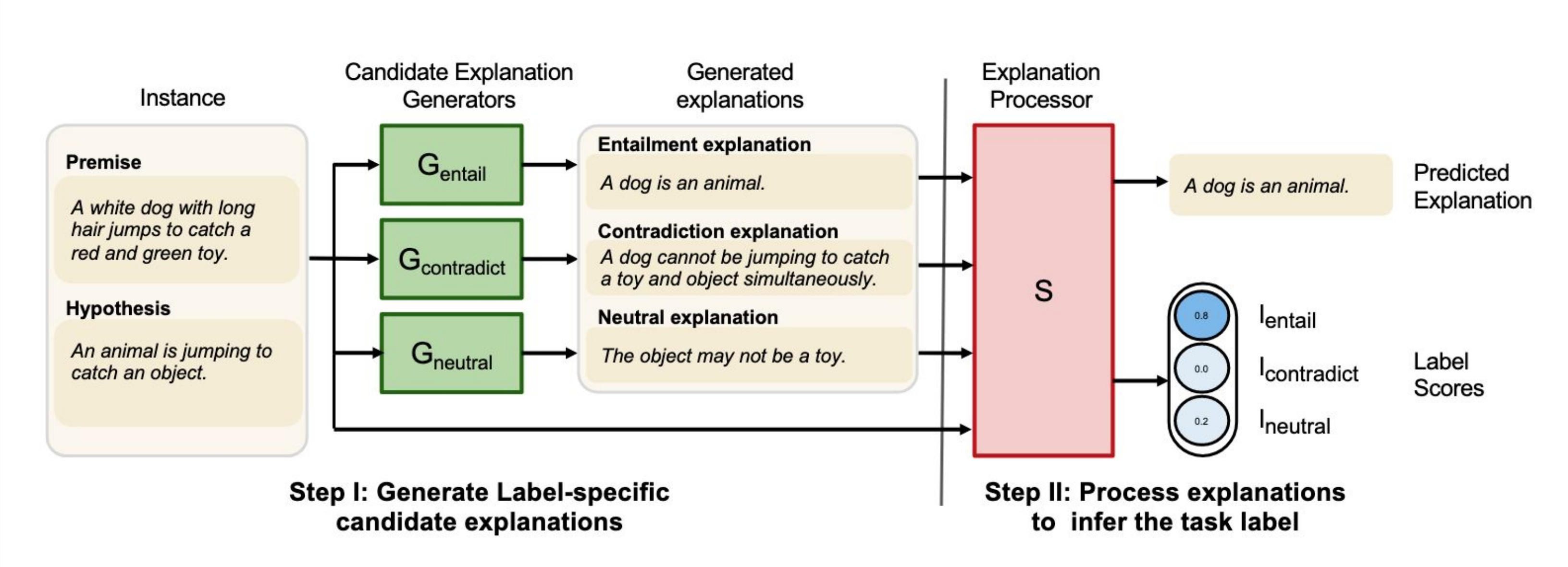
- The form of the explanation is enough to get predict the label, likely undermining faithfulness.
- How can we probe faithfulness?

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.

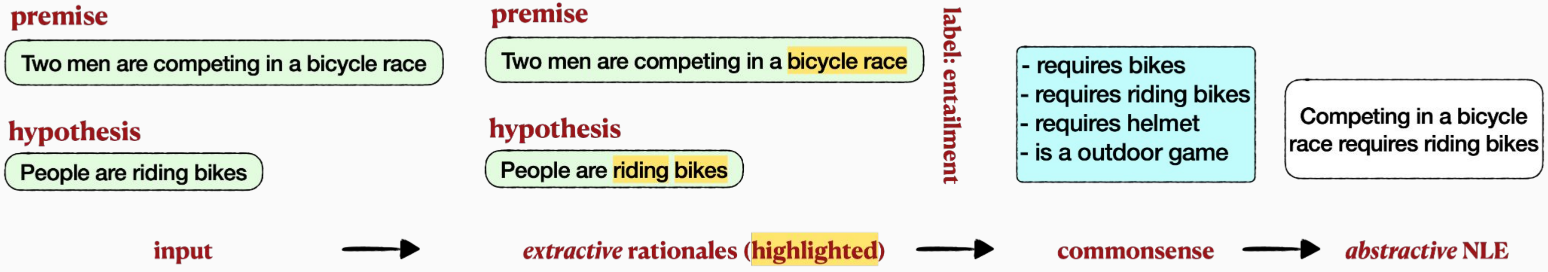
ORIGINAL HYPOTHESIS: A child rides a swing in the daytime.
PREDICTED LABEL: entailment
ORIGINAL EXPLANATION: **The sun is in the daytime.**

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?



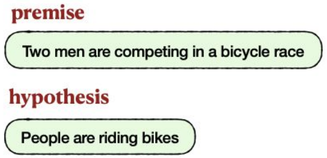
Rationale-Inspired Natural Language Explanations with Commonsense

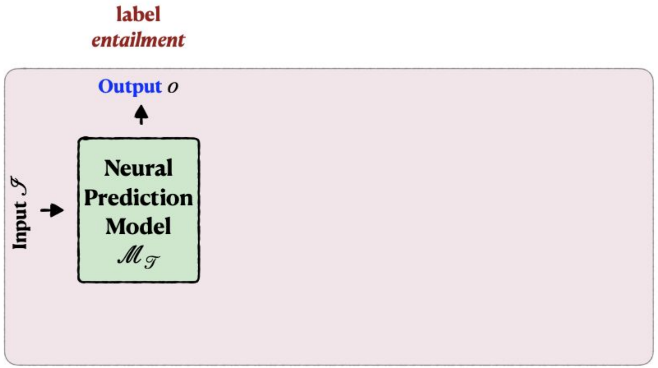
REXC

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

RExC

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





Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

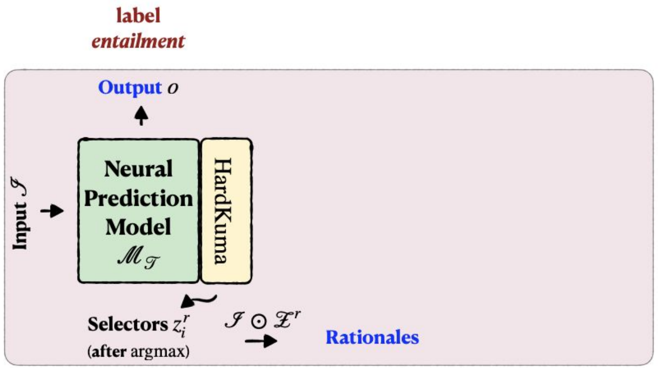
RExC

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes



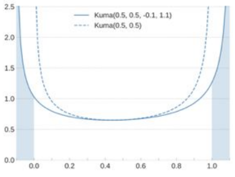
premise

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*

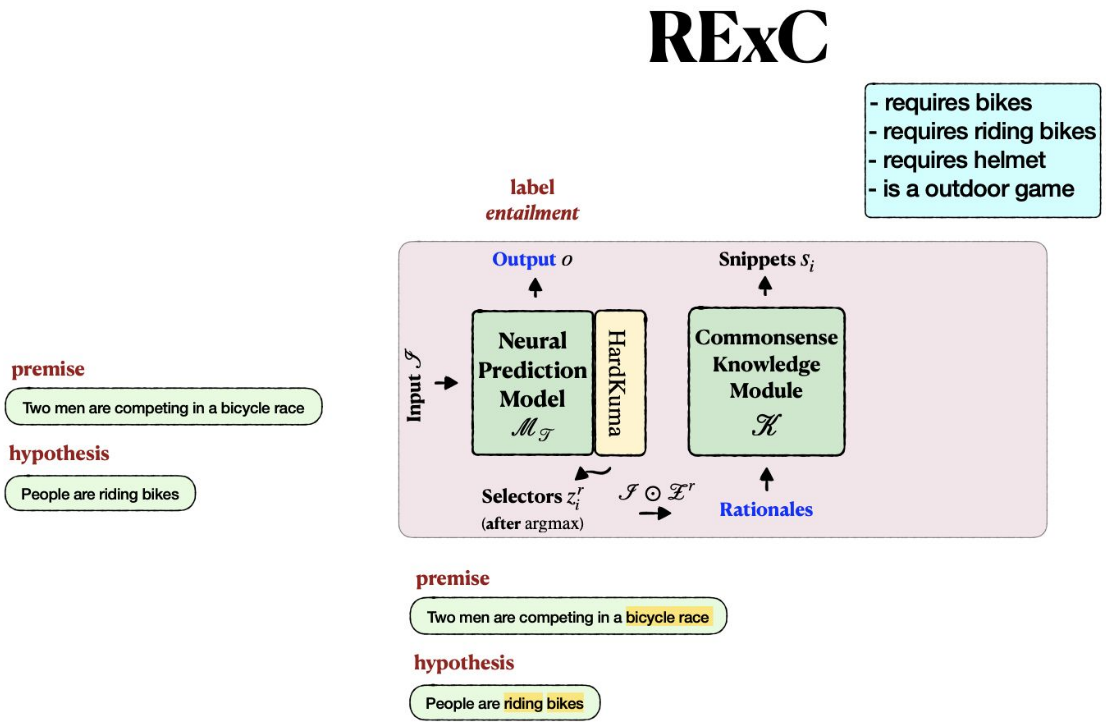


Bastings et al., 2020

L_1 regularization for sparsity

X

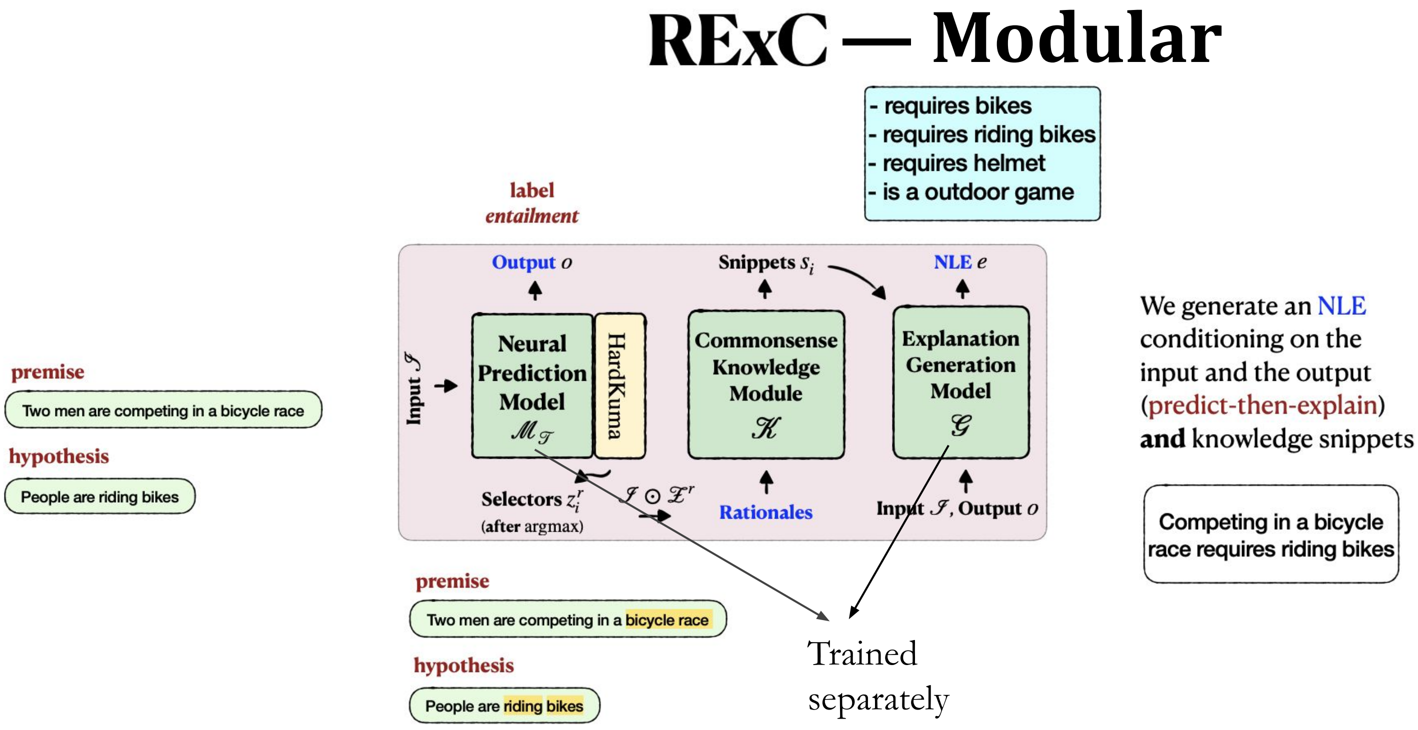
Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)



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

x

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

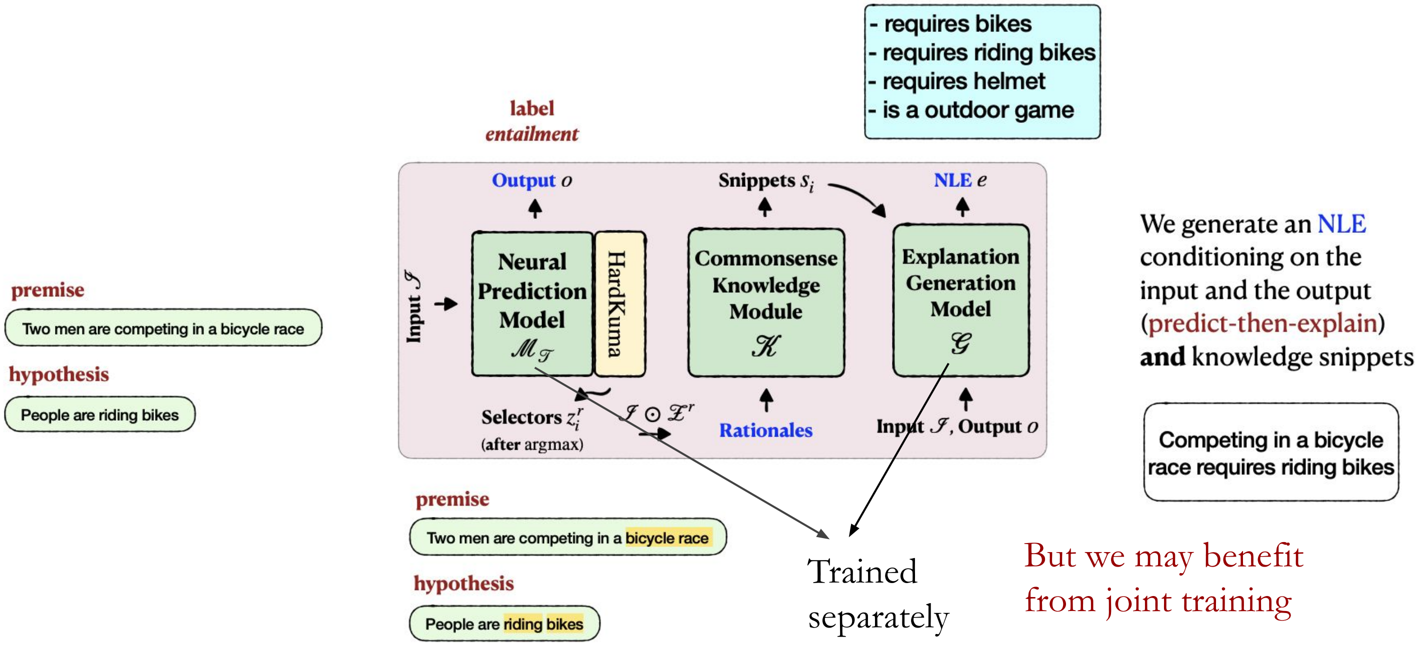


Trained separately

We generate an NLE conditioning on the input and the output (predict-then-explain) **and** knowledge snippets

x

RExC — Modular



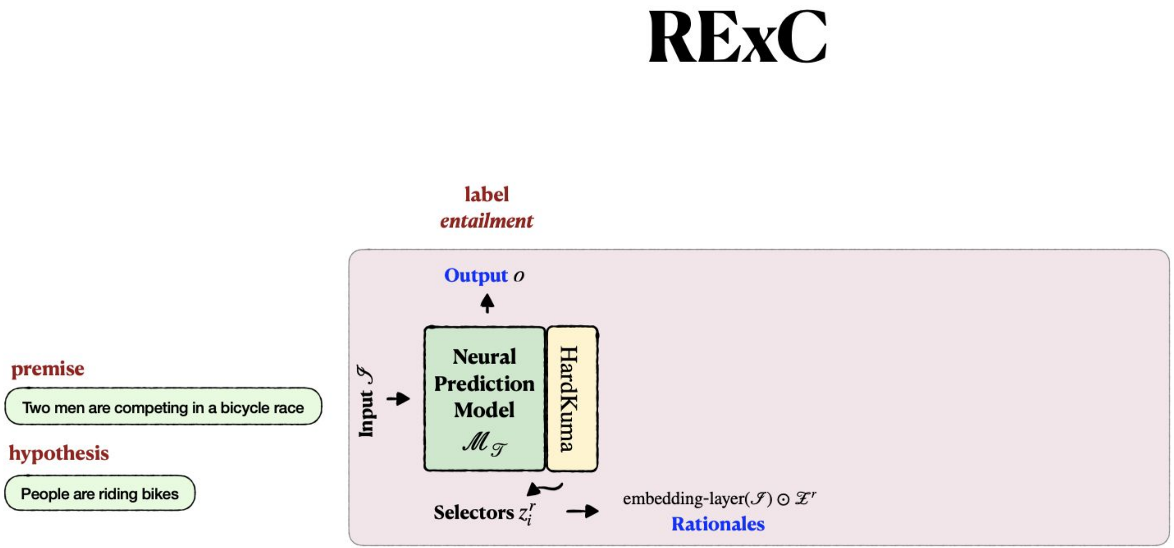
Trained separately

We generate an NLE conditioning on the input and the output (predict-then-explain) **and** knowledge snippets

But we may benefit from joint training

x

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

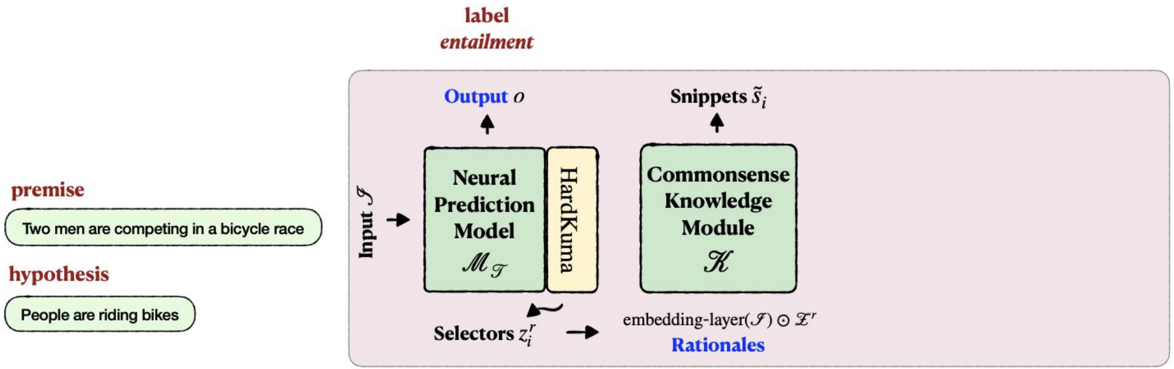


The series of binary latent variables z_i^r are used as masks on the embedded input

x

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

RExC

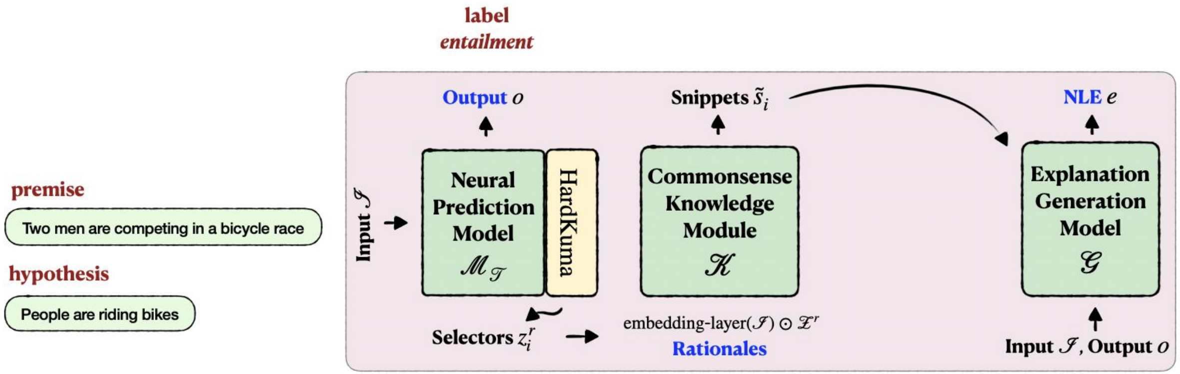


... and directly sent to a generative commonsense module $\mathcal{K}$, mirroring the modular approach

x

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

RExC — E2E

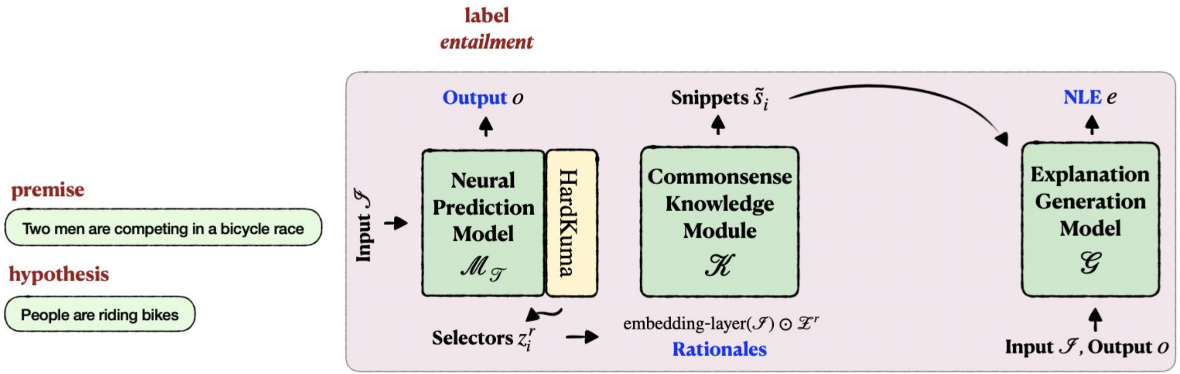


... and directly sent to a generative commonsense module $\mathcal{K}$, mirroring the modular approach

x

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

RExC — E2E

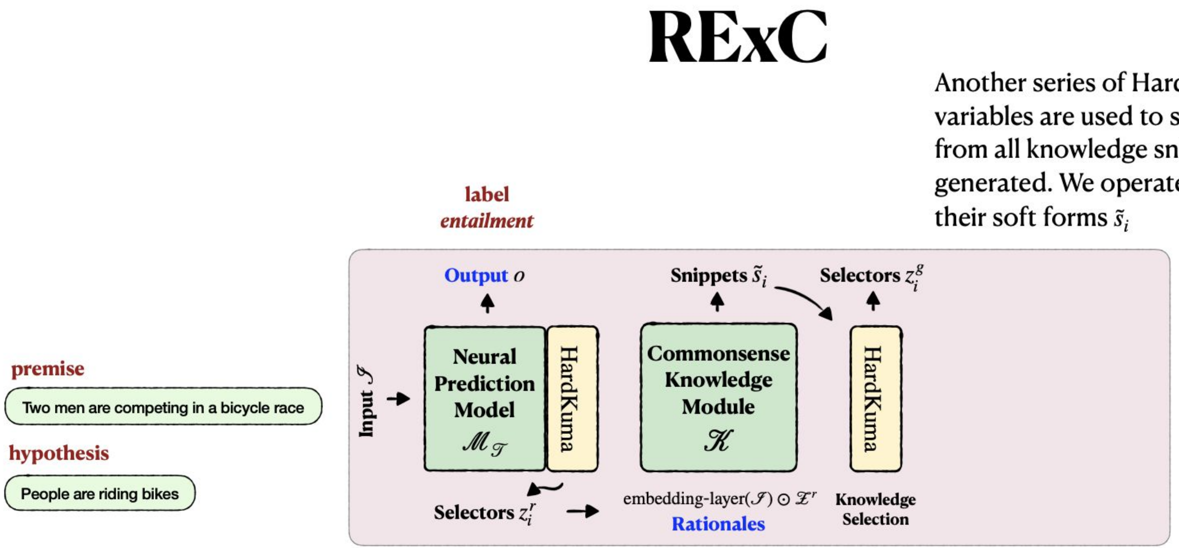


... and directly sent to a generative commonsense module $\mathcal{K}$, mirroring the modular approach

But we may benefit from doing a selection of the snippets

x

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

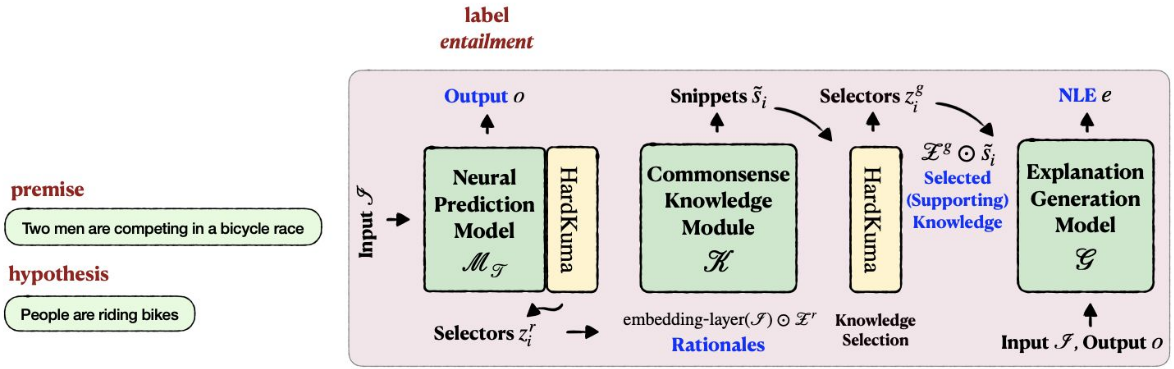


Another series of HardKuma variables are used to sample from all knowledge snippets generated. We operate on their soft forms $\tilde{s}_i$

x

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

RExC



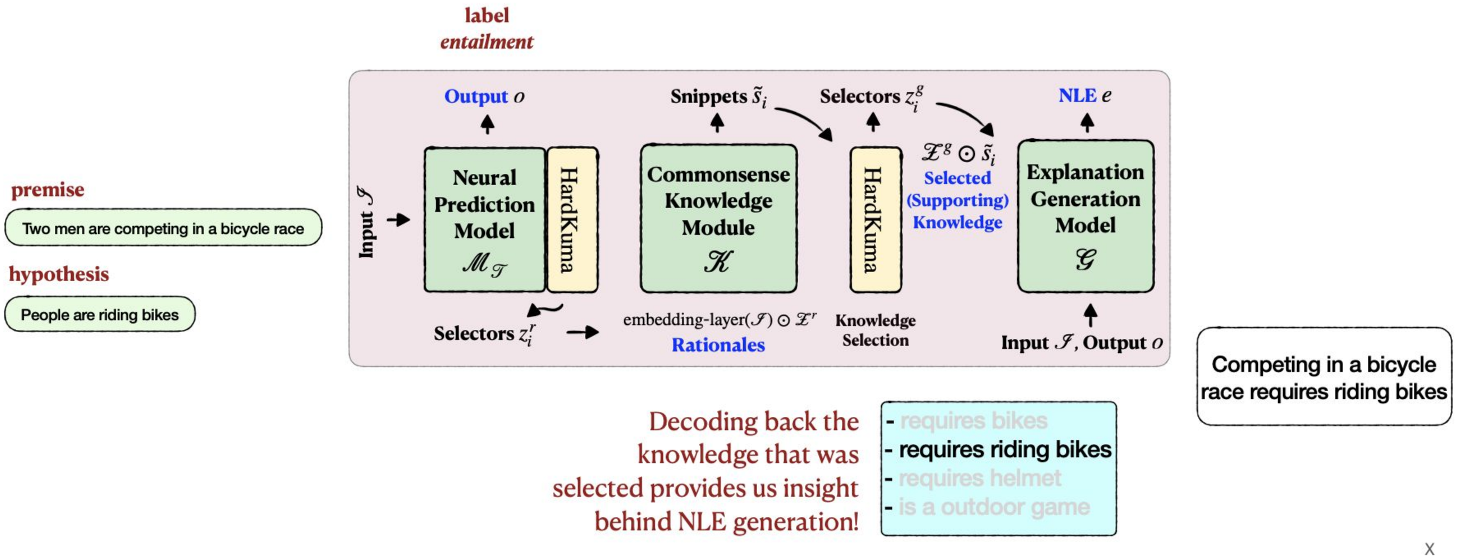
With the selected knowledge representations, generator $\mathcal{G}$ generates the NLE

Competing in a bicycle race requires riding bikes

X

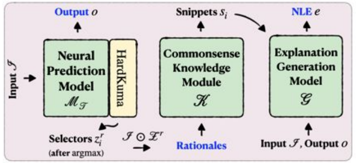
Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

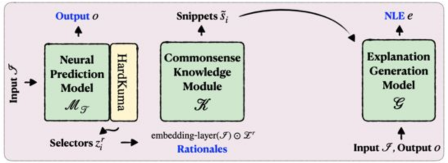
RExC — KS

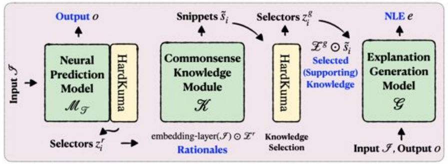


Variants of RExC

RExC-Mod | **RExC-E2E** | **RExC-KS** | **RExC-KS+**







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

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

X

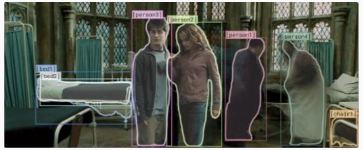
Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

Natural Language Tasks

Task	Example	Label	Dataset
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.

Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

NLP Tasks

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

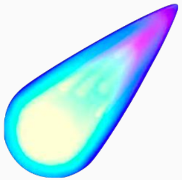
BART: a Seq2Seq pretrained transformer with a MLP prediction head

(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

M. Lewis et al., BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. ACL, 2020.
A. Bosselut et al., COMET: Commonsense transformers for automatic knowledge graph construction. ACL, 2019.

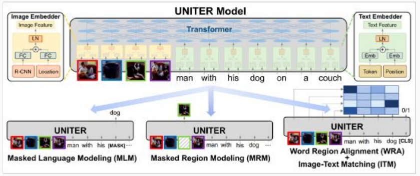
Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

Vision-Language Tasks

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

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

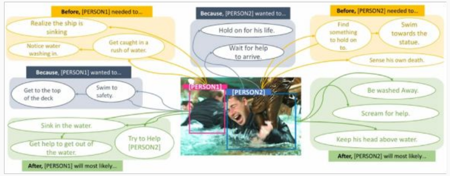
(Chen et al., 2020)



$\mathcal{K}$

Visual-COMET: Commonsense Transformer trained on Visual Commonsense Graph

(Park et al., 2020)



$\mathcal{G}$

GPT2: a pretrained transformer-based Language Model

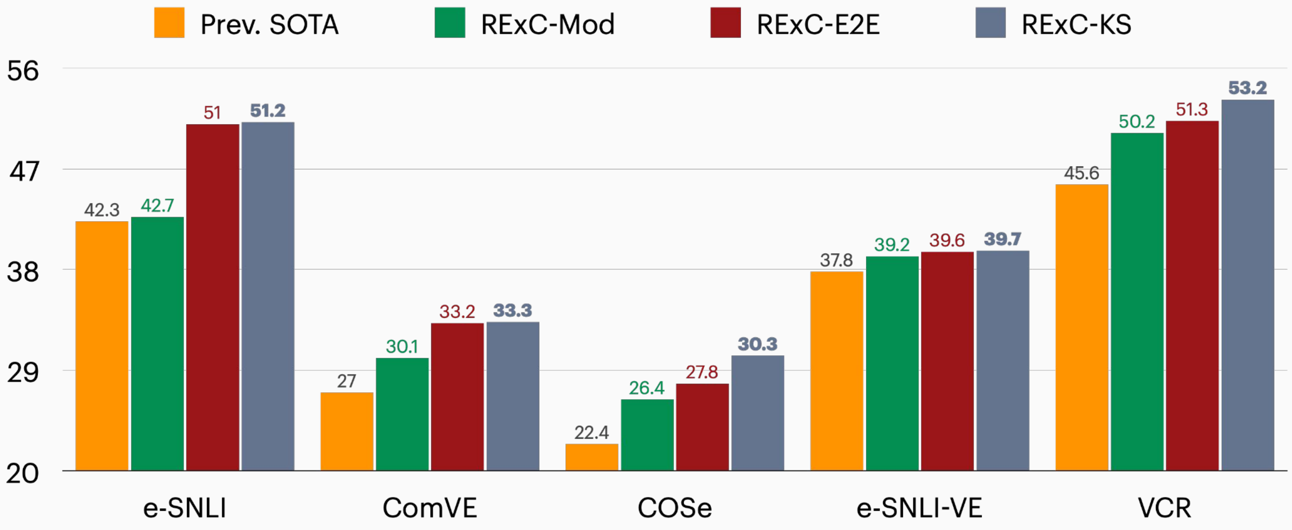
(Radford et al., 2020)

GPT-2

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)

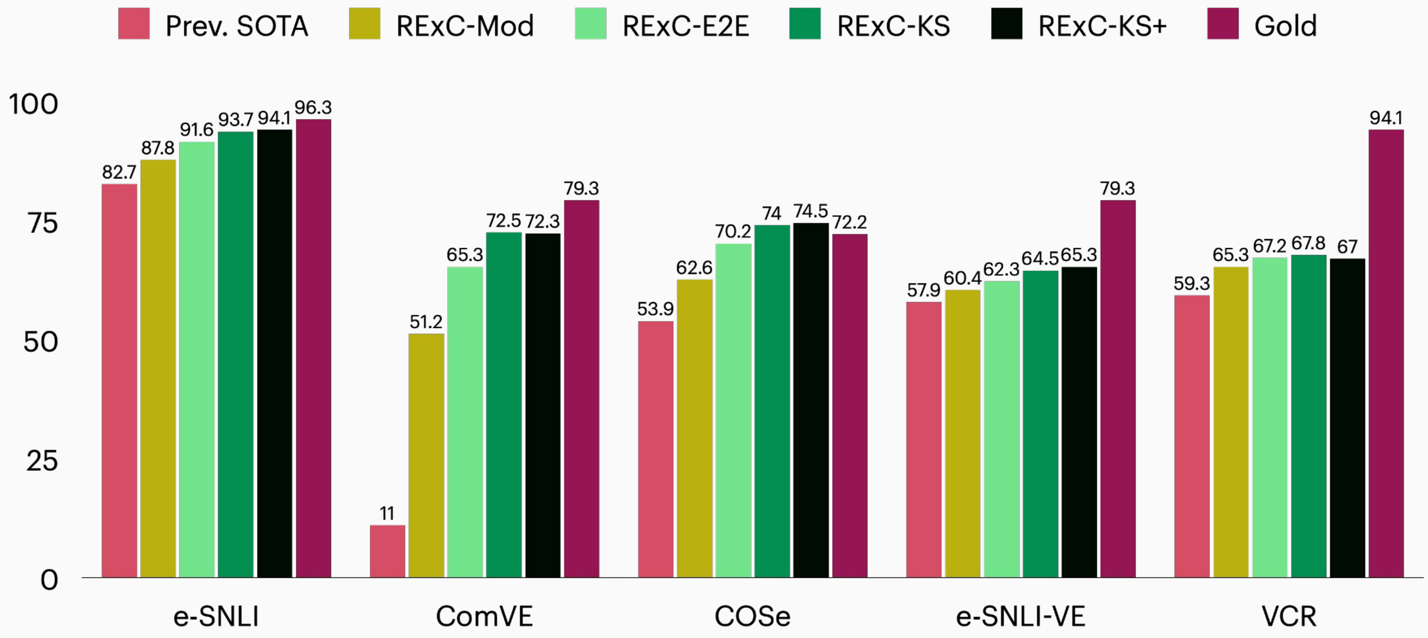
BLEURT (Sellam et al., 2020)



T. Sellam et al., BLEURT :Learning robust metrics for text generation, ACL, 2020.

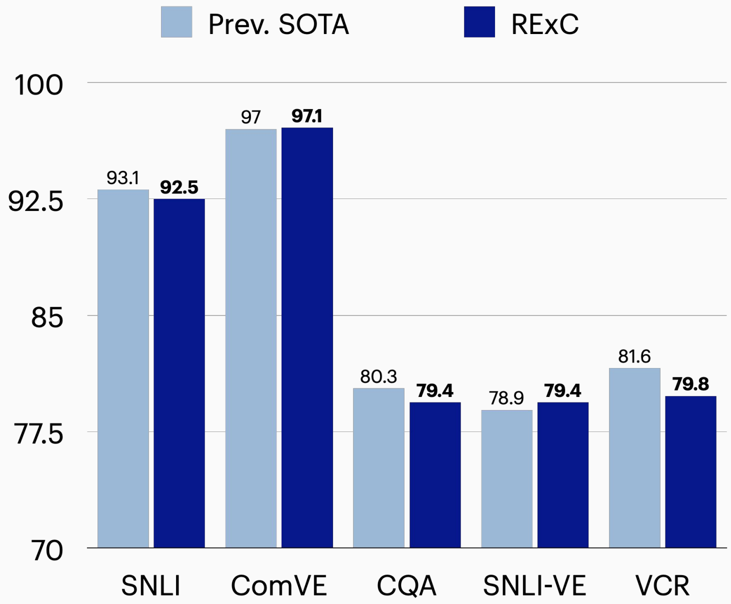
Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

Human evaluation



Rationale-Inspired Natural Language Explanations with Commonsense (Majumder et al., 2021)

Task performance



Summary Part 1

Thank you!

@oanacamb

Questions?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

EML
Explainable Machine Learning

Natural-XAI: ICML Tutorial – Part 2

Prof. Dr. Zeynep Akata
University of Tübingen, Cluster of Excellence Machine Learning
Explainable Machine Learning (EML) Group

19 July 2021

Outline

Outline

Explanation and Learning are Related

Generating Natural Language Explanations for Visual Decisions

Modeling Conceptual Understanding of the User

Summary and Future Work

The Ultimate Goal of Learning

the supreme goal of all theory is to make
the irreducible basic elements as simple and as few as possible

The Ultimate Goal of Learning

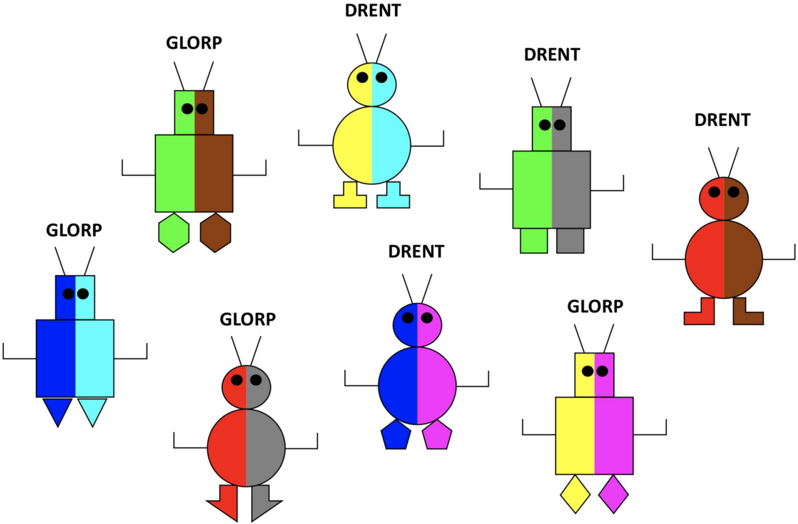
the supreme goal of all theory is to make
the irreducible basic elements as simple and as few as possible

without having to surrender
the adequate representation of a single datum of experience

Albert Einstein, 1934

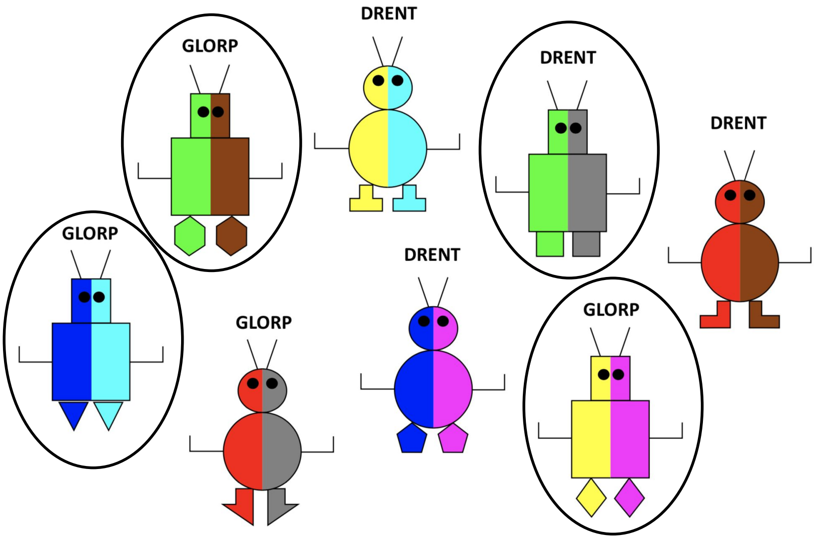
Learning via Explanation

Tania Lombrozo TICS'16



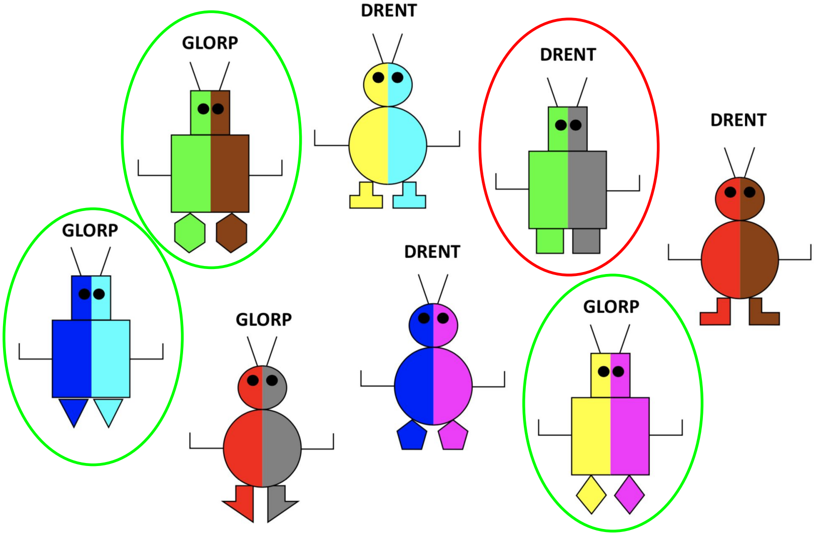
Learning via Explanation

Tania Lombrozo TICS'16



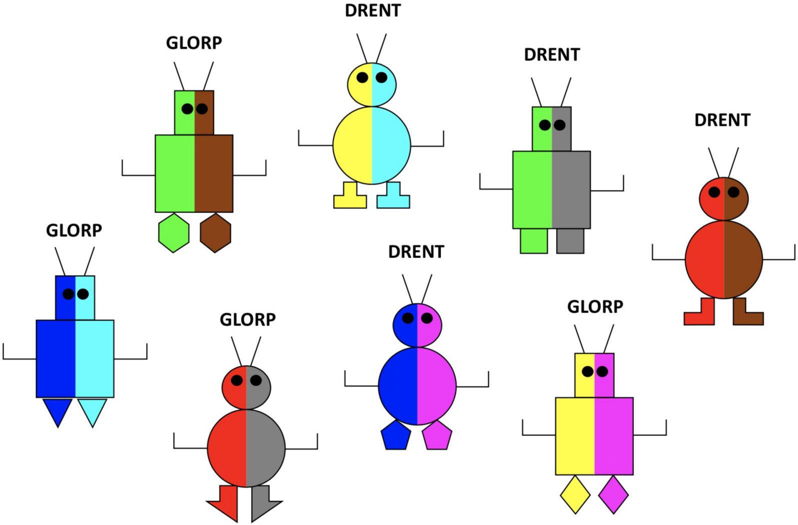
Learning via Explanation

Tania Lombrozo TICS'16



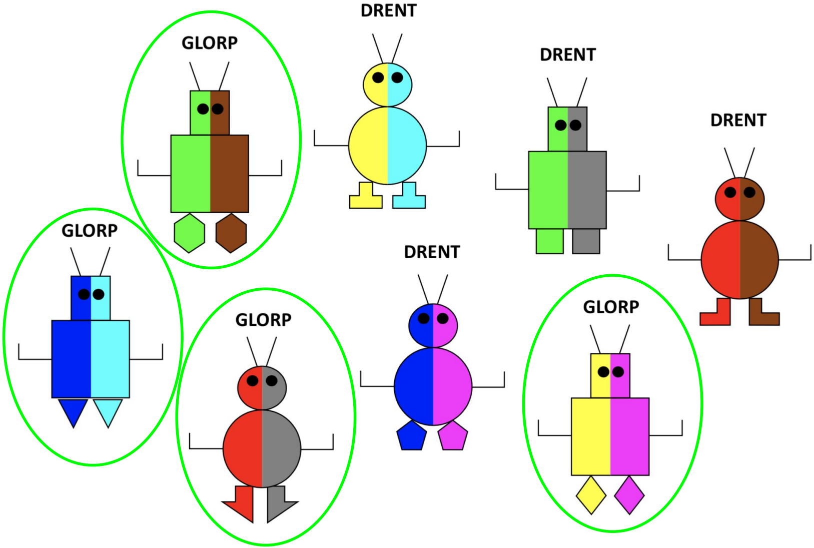
Learning via Explanation

Tania Lombrozo TICS'16



Learning via Explanation

Tania Lombrozo TICS'16



Explanations are ...

Lombrozo TICS'16

- Broad: they justify a broader range of observations or phenomena
- Simple: they provide a concise description for the communication partner
- Contrastive: they differentiate two alternative decisions
- Helpful for another task: they entail transferable information

Explanations are ...

Lombrozo TICS'16

- Broad: they justify a broader range of observations or phenomena
- Simple: they provide a concise description for the communication partner
- Contrastive: they differentiate two alternative decisions
- Helpful for another task: they entail transferable information

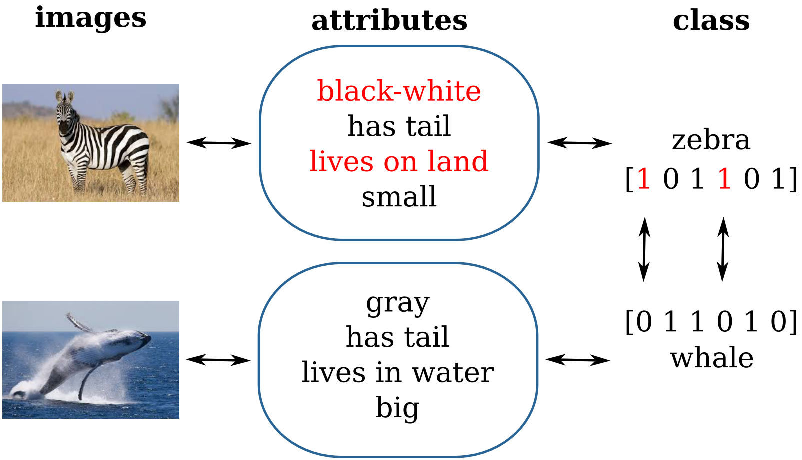
Attributes as Explanations

Lampert et al. CVPR'09

images	attributes	class
	black-white has tail lives on land small	zebra [1 0 1 1 0 1]
	gray has tail lives in water big	[0 1 1 0 1 0] whale

Attributes as Explanations

Lampert et al. CVPR'09



Attributes as Explanations

Lampert et al. CVPR'09

images	**attributes**	**class**
	black-white has tail lives on land small	zebra [1 0 1 1 0 1]
	gray has tail lives in water big	[0 1 1 0 1 0] whale

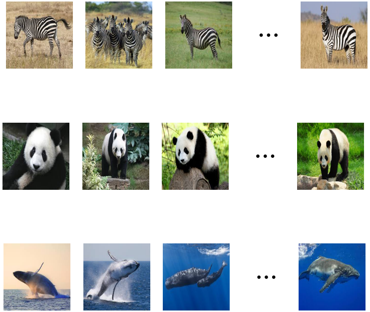
Explanations are ...

Lombrozo TICS'16

- Broad: they justify a broader range of observations or phenomena
- Simple: they provide a concise description for the communication partner
- Contrastive: they differentiate two alternative decisions
- Helpful for another task: they entail transferable information

Generalized Zero-Shot Learning

images



attributes

black-white
has tail
lives on land
small

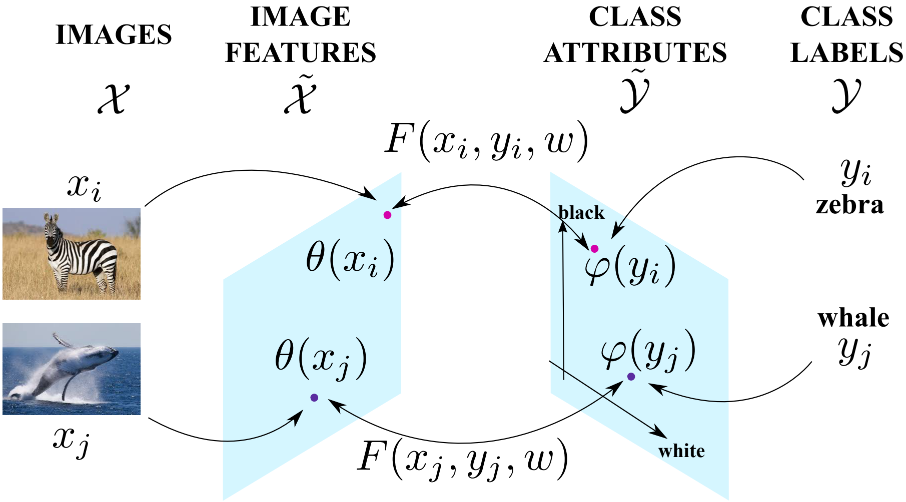
black-white
no tail
lives on land
medium

gray
has tail
lives in water
big

white
has tail
lives on land
tiny

Muldimodal Embeddings

Akata et al. CVPR'13 & TPAMI'16



Benchmark Example Datasets for Zero-Shot Learning

Animals with Attributes (AWA) [Lampert et.al. CVPR'09]	50 cls	85 att	
Caltech UCSD-Birds (CUB) [Wah et.al.'11]	200 cls	312 att	

Zero-Shot Learning: A Comprehensive Evaluation of the Good, the Bad, the Ugly;
Xian, Lampert, Schiele, Akata at IEEE TPAMI 2019

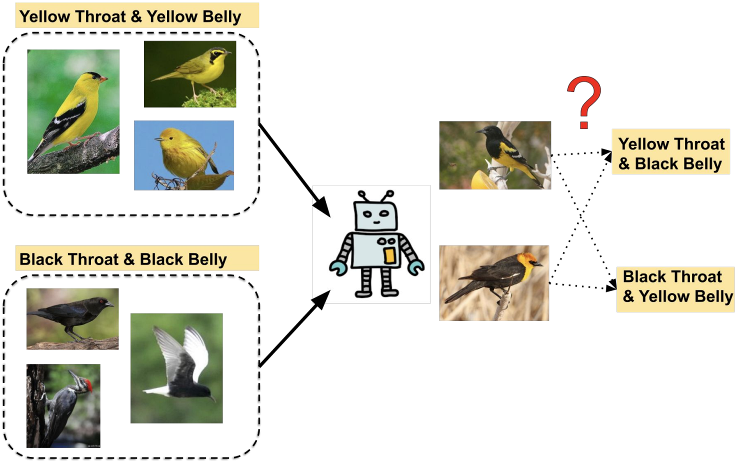
Attribute Explanations in Zero Shot Learning

	AWA	CUB
`class labels`	0	0
`attributes`	66.7	50.1

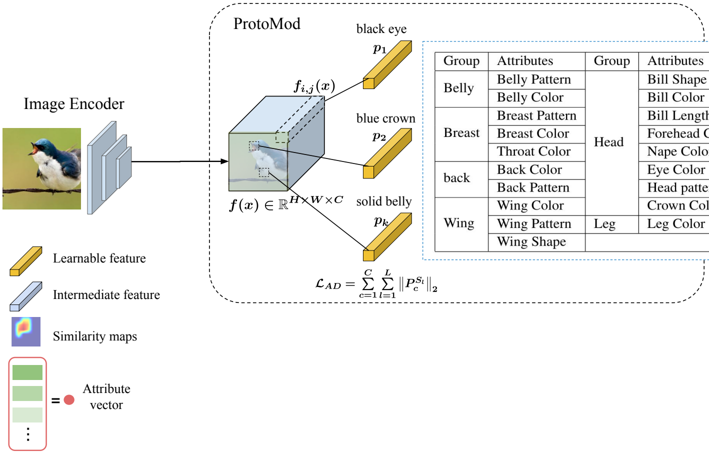
$$\text{Top-1 accurracy} = \frac{1}{\|\mathcal{Y}^u\|} \sum_{c=1}^{\|\mathcal{Y}^u\|} \frac{\#\ \text{correct in c}}{\#\ \text{samples in c}}$$

Attributes of Fine-Grained Objects Can Be Confusing

Incidental correlations between attributes as they often co-occur



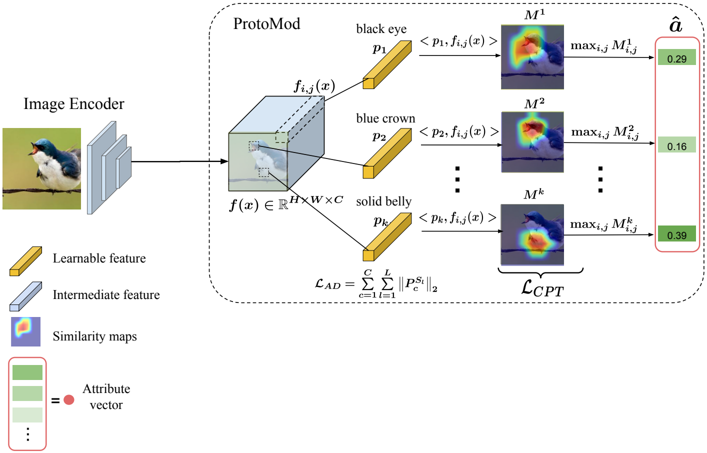
Attribute Prototype Network



Group	Attributes	Group	Attributes	Group	Attributes
Belly	Belly Pattern	Head	Bill Shape	Tail	Upper Tail Color
	Belly Color		Bill Color		Udder Tail Color
Breast	Breast Pattern		Bill Length		Tail Pattern
	Breast Color		Forehead Color		Tail Shape
	Throat Color		Nape Color	Others	Bird Size
back	Back Color		Eye Color		Bird Shape
	Back Pattern		Head pattern		Primary Color
Wing	Wing Color		Crown Color		Under Parts Color
	Wing Pattern	Leg	Leg Color		Upper Parts Color
	Wing Shape				

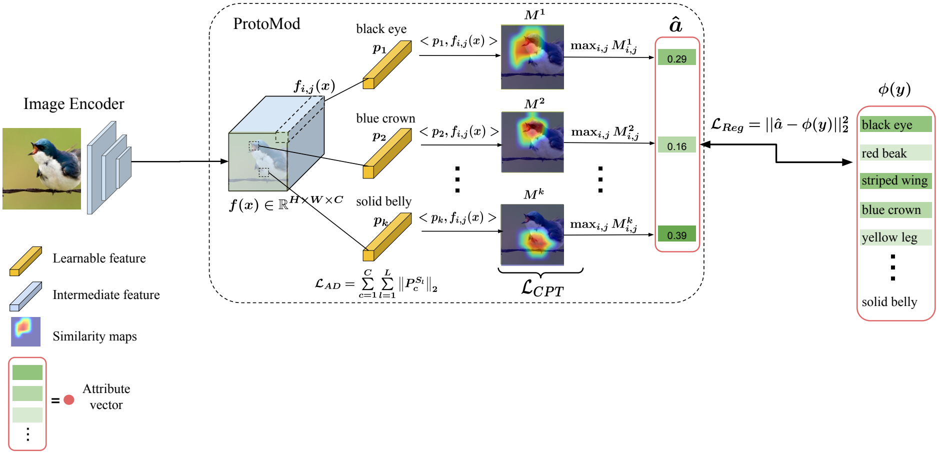
Attribute Prototype Network for Zero-Shot Learning; Xu, Xian, Wang, Schiele, Akata at NeurIPS 2020

Attribute Prototype Network



Attribute Prototype Network for Zero-Shot Learning; Xu, Xian, Wang, Schiele, Akata at NeurIPS 2020

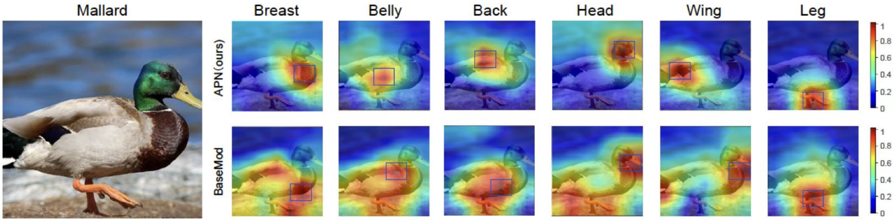
Attribute Prototype Network



Attribute Prototype Network for Zero-Shot Learning; Xu, Xian, Wang, Schiele, Akata at NeurIPS 2020

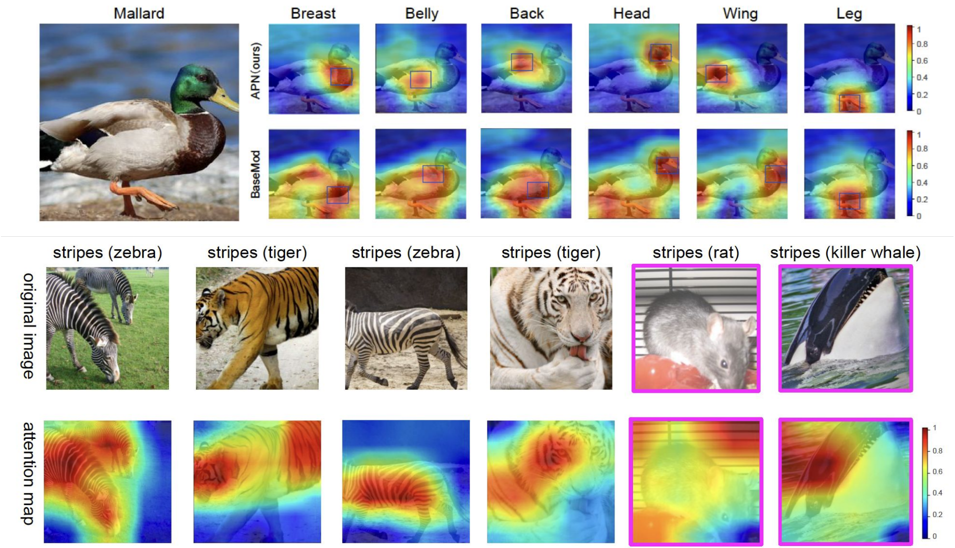
Visualizing Attribute Prototypes

Xu et al. NeurIPS 2020



Visualizing Attribute Prototypes

Xu et al. NeurIPS 2020



Wikipedia and WordNet as Explanations

Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Zebra

From Wikipedia, the free encyclopedia

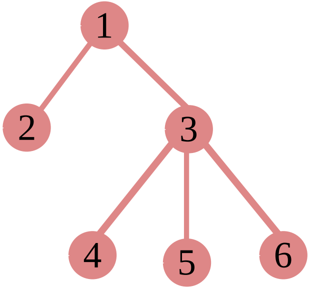
For other uses, see Zebra (disambiguation).

Zebras (/ˈzɛbrə/ ZEB-rə or /ˈziːbrə/ ZEE-brə)[1] are several species of African equids (horse family) united by their distinctive black and white striped coats. Their stripes come in different patterns, unique to each individual. They are generally social animals that live in small harems to large herds. Unlike their closest relatives, horses and donkeys, zebras have never been truly domesticated.

There are three species of zebras: the plains zebra, the Grévy's zebra and the mountain zebra. The plains zebra and the mountain zebra belong to the subgenus *Hippotigris*, but Grévy's zebra is the sole species of subgenus *Dolichohippus*. The latter resembles an ass, to which it is closely related, while the former two are more horse-like. All three belong to the genus *Equus*, along with other living equids.

Zebra

A herd of plains zebra (*Equus quagga*)



$$2 = [1\ 0\ 2\ 3\ 3\ 3]$$

Word2Vec [Mikolov et.al. NIPS'13]
GloVe [Pennington et.al EMNLP'14]

Hierarchical similarity measures

Wikipedia, WordNet Explanations in Zero Shot Learning

	AWA	**CUB**
`w2v`	51.2	28.4
`glo`	58.8	24.2
`hie`	51.2	20.6
`w2v + glo + hie`	**60.1**	**29.9**

Wikipedia, WordNet Explanations in Zero Shot Learning

	AWA	CUB
`w2v`	51.2	28.4
`glo`	58.8	24.2
`hie`	51.2	20.6
`w2v + glo + hie`	**60.1**	**29.9**
`att`	66.7	50.1
`w2v + glo + hie + att`	**73.9**	**51.7**

Natural Language as a Proxy for Explanations

The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.

This bird has distinctive-looking brown and white stripes all over its body, and its brown tail sticks up.

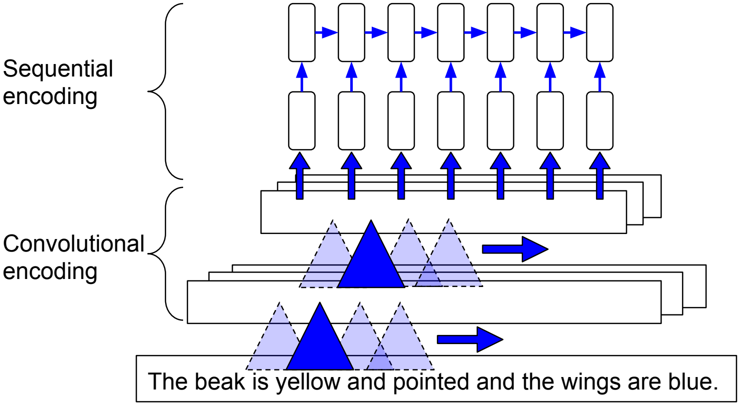
This flower has a central white blossom surrounded by large pointed red petals which are veined and leaflike.

Light purple petals with orange and black middle green leaves

Learning Deep Representations of Fine-Grained Visual Descriptions;
Reed, Akata, Schiele, Lee at IEEE CVPR 2016

Deep Representations of Text

Reed et.al. CVPR'16



Text-Based Explanations in Zero-Shot Learning

	AWA	CUB
`w2v + glo + hie`	60.1	29.9
`att`	66.7	50.1
`w2v + glo + hie + att`	73.9	51.7

Text-Based Explanations in Zero-Shot Learning

	AWA	**CUB**
`w2v + glo + hie`	60.1	29.9
`att`	66.7	50.1
`w2v + glo + hie + att`	73.9	51.7
`text`	N/A	**56.8**

Conclusions for: Explanations and Learning are Related

Attribute-based and Natural Language Explanations

1. Provide an intuitive interface for the model
2. Provide side information to learn strong and generalizable representations
3. Complement visual information in limited data regimes

Outline

Explanation and Learning are Related

Generating Natural Language Explanations for Visual Decisions

Modeling Conceptual Understanding of the User

Summary and Future Work

Natural Language for Fine-Grained Explanations

The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.

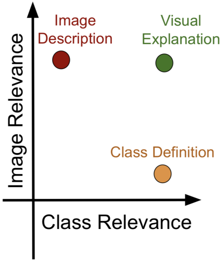
This bird has distinctive-looking brown and white stripes all over its body, and its brown tail sticks up.

This flower has a central white blossom surrounded by large pointed red petals which are veined and leaflike.

Light purple petals with orange and black middle green leaves

Learning Deep Representations of Fine-Grained Visual Descriptions;
Reed, Akata, Schiele, Lee at IEEE CVPR 2016

Difference between: Definition, Description and Explanation



Western Grebe

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Western Grebe* is a waterbird with a yellow pointy beak, white neck and belly, and black back.
Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

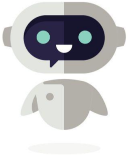
Laysan Albatross

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

Laysan Albatross

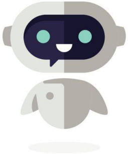
Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck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

Natural Language Explanations for Human Machine Communication

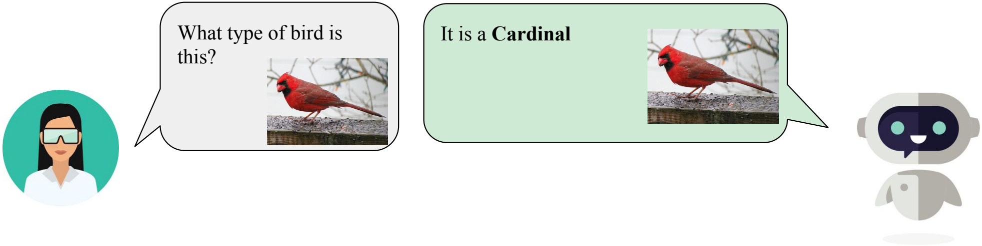


Natural Language Explanations for Human Machine Communication

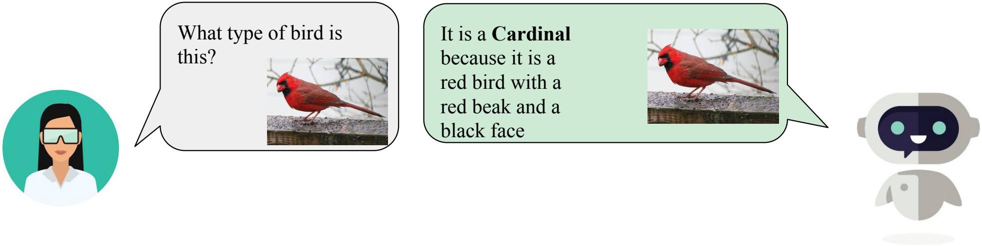
What type of bird is this?



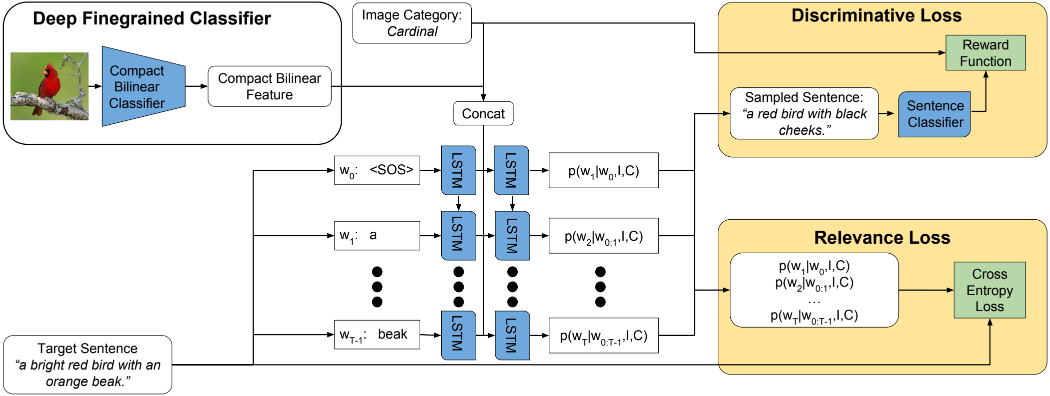
Natural Language Explanations for Human Machine Communication



Natural Language Explanations for Human Machine Communication



Generating Visual Explanations



Generating Visual Explanations;

Hendricks, Akata, Rohrbach, Donahue, Schiele, Darrell at ECCV 2016

Generating Visual Explanations Results

Hendricks et al. ECCV'16

This is a ***Downy Woodpecker*** *because...*

D: this bird has a white breast black wings and a **red spot** on its head.
E: this is a black and white bird with a **red spot** on its crown.

This is a ***Downy Woodpecker*** *because...*

D: this bird has a white breast black wings and a **red spot** on its head.
E: this is a white bird with a black wing and a black and white striped head.

Generating Visual Explanations Results

Hendricks et al. ECCV'16

*This is a **Downy Woodpecker** because...*

D: this bird has a white breast black wings and a **red spot** on its head.
E: this is a black and white bird with a **red spot** on its crown.

*This is a **Downy Woodpecker** because...*

D: this bird has a white breast black wings and a **red spot** on its head.
E: this is a white bird with a black wing and a black and white striped head.

Correct: Laysan Albatross, **Predicted:** Cactus Wren

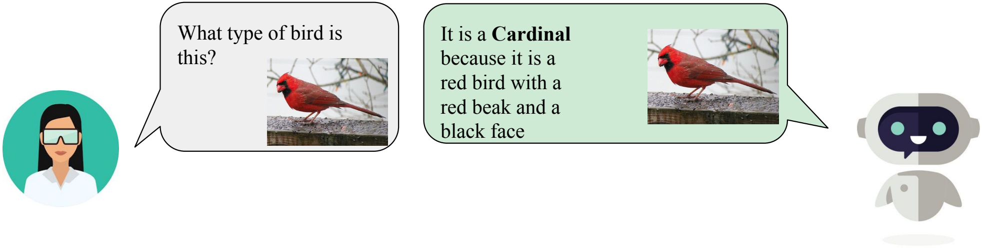
Explanation: ...this is a brown and white spotted bird with a long pointed beak.

Correct & Predicted: Laysan Albatross

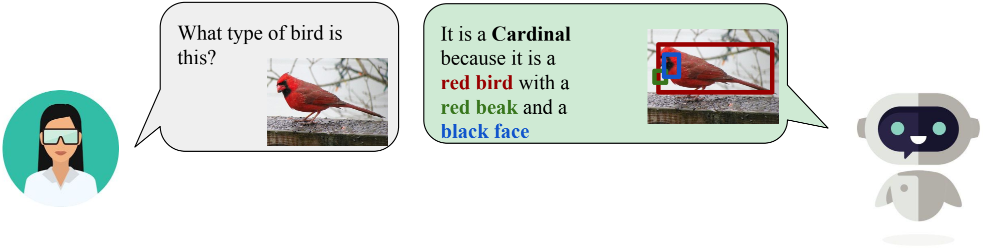
Explanation: ...this bird has a white head and breast with a long hooked bill.

***Cactus Wren* Definition:** ...this bird has a long thin beak with a brown body and black spotted feathers.
***Laysan Albatross* Definition:** ...this bird has a white head and breast a grey back and wing feathers and an orange beak.

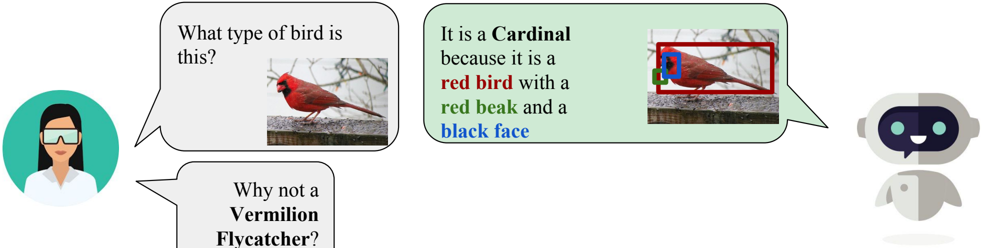
Natural Language Explanations for Human Machine Communication



Natural Language Explanations for Human Machine Communication

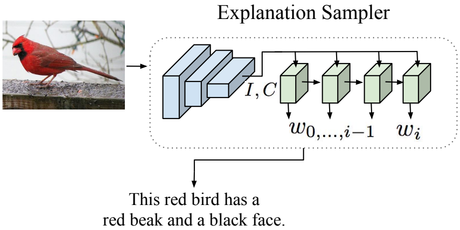


Natural Language Explanations for Human Machine Communication



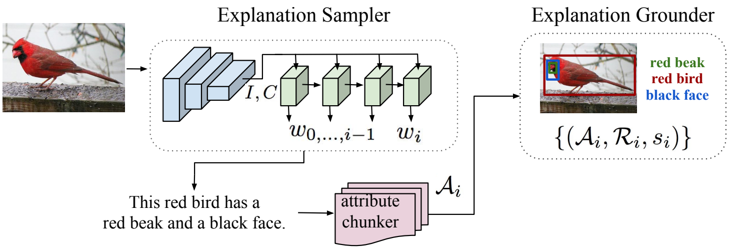
Natural Language Explanations for Human Machine Communication

Generating and Grounding Visual Explanations



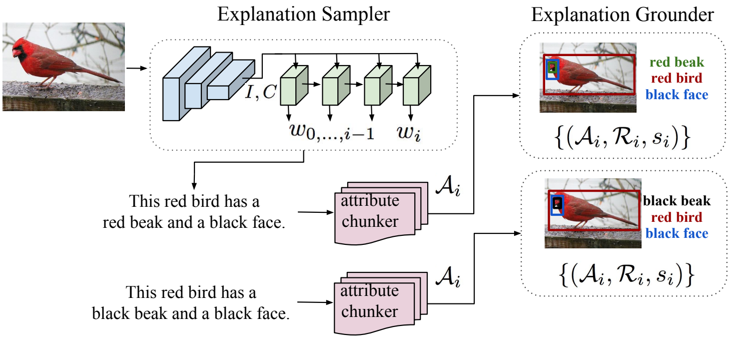
Grounding Visual Explanations; Hendricks, Hu, Darrell, Akata at ECCV 2018

Generating and Grounding Visual Explanations



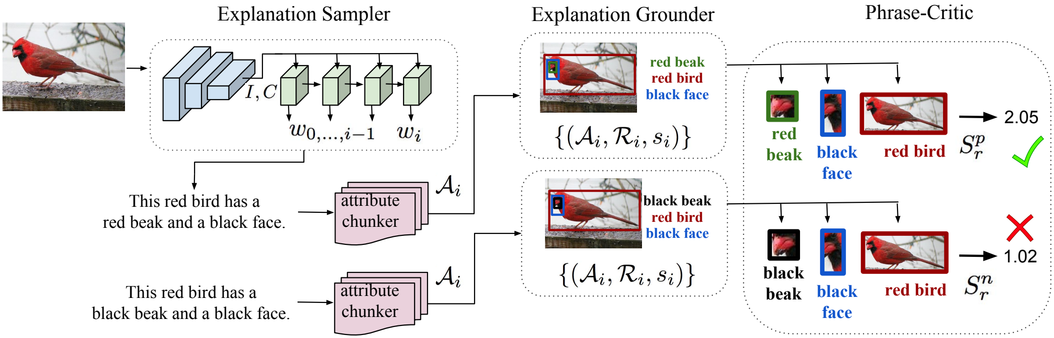
Grounding Visual Explanations; Hendricks, Hu, Darrell, Akata at ECCV 2018

Generating and Grounding Visual Explanations



Grounding Visual Explanations; Hendricks, Hu, Darrell, Akata at ECCV 2018

Generating and Grounding Visual Explanations



Grounding Visual Explanations; Hendricks, Hu, Darrell, Akata at ECCV 2018

Grounding Visual Explanations and Counterfactuals

This is a **Red Winged Blackbird** because ….

this is a **black bird** with a **red spot on its wingbars**.

Score: -11.29

this is a black bird with a red wing and a pointy black beak.

Grounding Visual Explanations and Counterfactuals

This is a **Red Winged Blackbird** because ....

this is a **black bird** with a **red spot on its wingbars**.

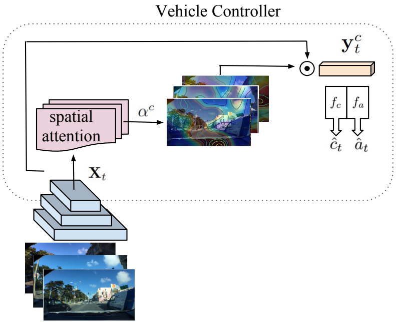
Score: -11.29

this is a black bird with a red wing and a pointy black beak.

Counterfactuals: Contrasting explanations are intuitive and informative

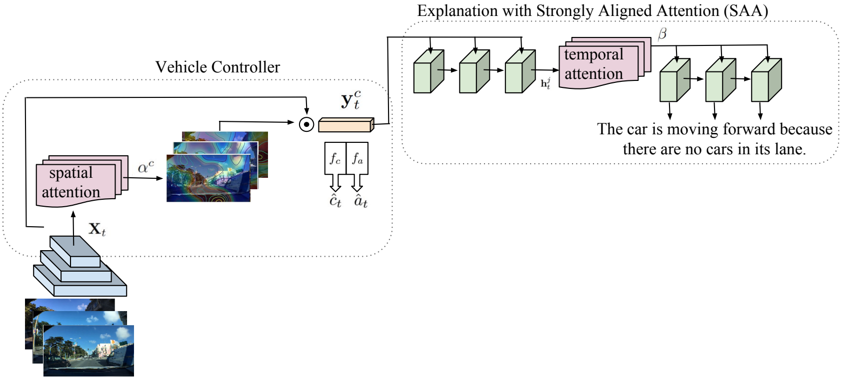
This bird is a **Crested Auklet** because this is a black bird with a small orange beak and it is not a **Red Faced Cormorant** because it does not have a long flat bill.

Textual Explanations for Self Driving Vehicles



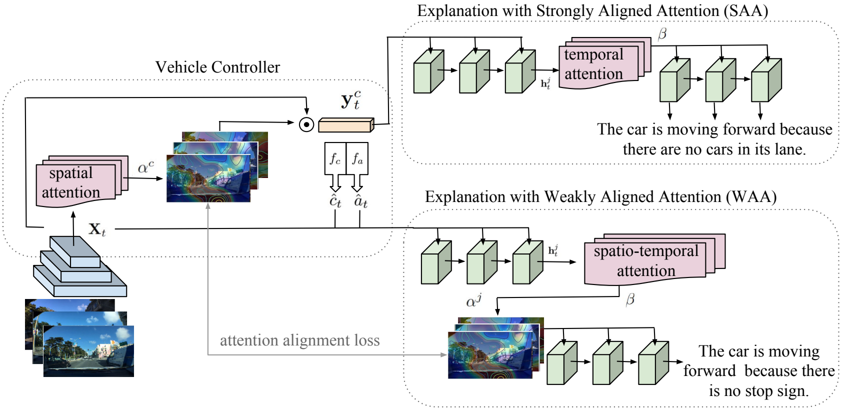
Textual Explanations for Self-Driving Vehicles; Kim, Rohrbach, Darrell, Canny, Akata at ECCV 2018

Textual Explanations for Self Driving Vehicles



Textual Explanations for Self-Driving Vehicles; Kim, Rohrbach, Darrell, Canny, Akata at ECCV 2018

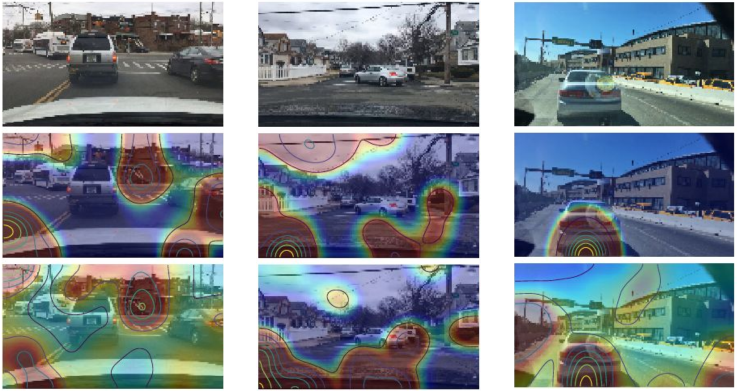
Textual Explanations for Self Driving Vehicles



Textual Explanations for Self-Driving Vehicles; Kim, Rohrbach, Darrell, Canny, Akata at ECCV 2018

Textual Explanations for Self Driving Vehicles

Kim et al. ECCV'18

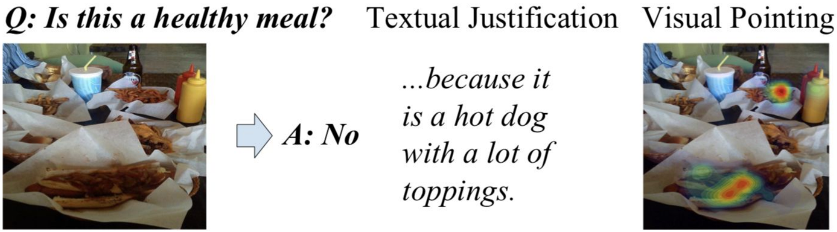


The car heads down the road because traffic is moving at a steady pace.

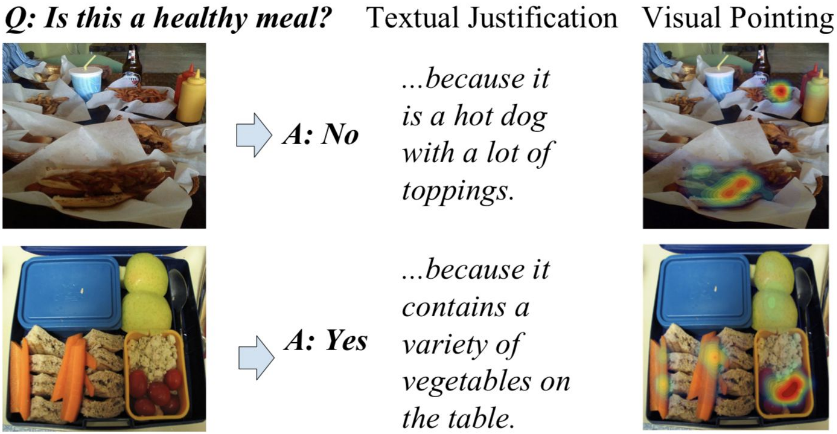
The car is slowing because it is approaching a stop sign.

The car is stopped because the car in front of it is stopped.

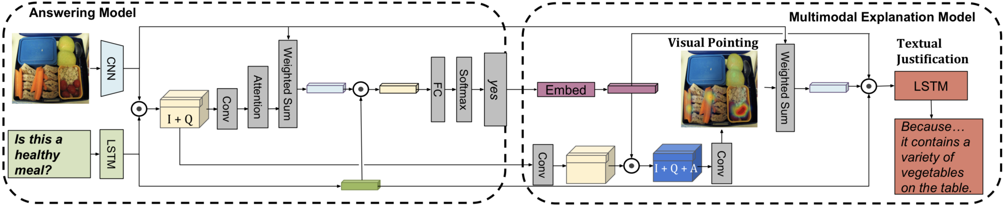
Explaining the Answers of Questions about the Image



Explaining the Answers of Questions about the Image



Justifying Decisions and Pointing to the Evidence

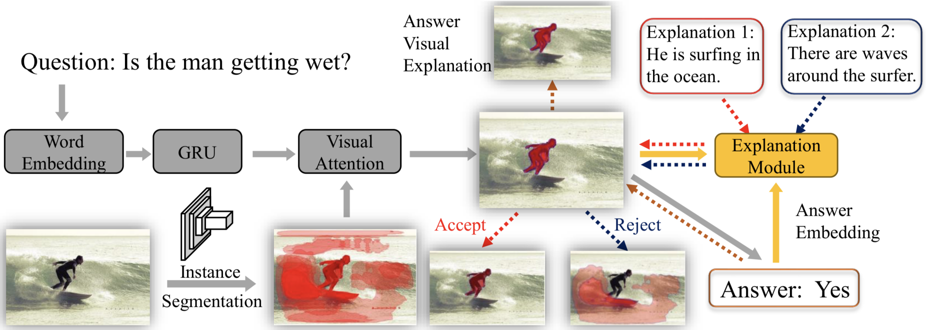


Ablation study shows that

- image attention and answer conditioning improves explanation generation quality

Multimodal Explanations: Justifying Decisions and Pointing to the Evidence;
Park, Hendricks, Akata, Schiele, Darrell, Rohrbach at IEEE CVPR 2018

Faithful Multimodal Explanation

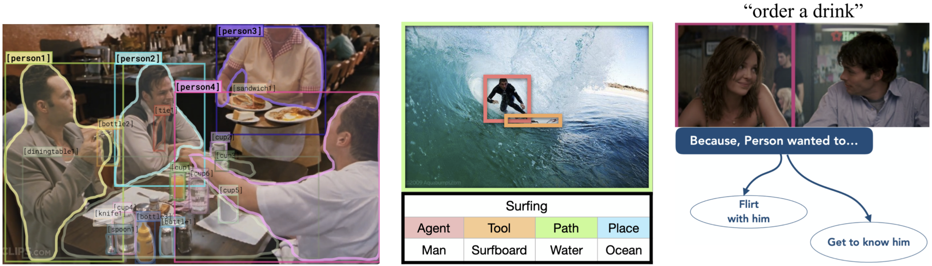


A feedback loop from the generated explanation aims to ensure that

- explanation utilizes the same visual features used to produce the answer

Faithful Multimodal Explanation for Visual Question Answering; Wu, Mooney at ACL 2019

Rationale VT Transformer

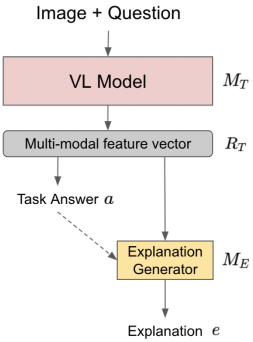


(a) Object Detection (b) Grounded Situation Recognition (c) Visual Commonsense Graph

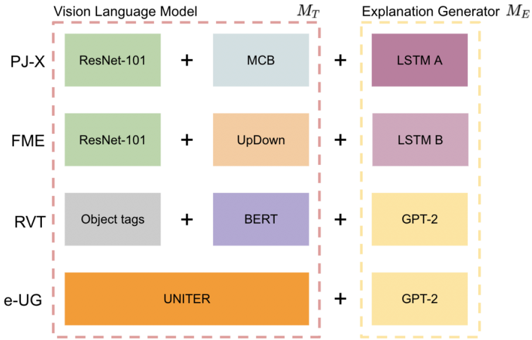
Image feature + gt answer + question + pre-trained GPT-2 model → explanation

Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs; Marasovic, Bhagavatula, Park, Bras, Smith, Choi at EMNLP 2020

e-ViL: Generating Explanations for Visual Entailment



(a) High-level structure of VL models.



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

e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks; Kayser, Camburu, Salewski, Emde, Do, Akata, Lukasiewicz; Ongoing Work

e-ViL Dataset for Explaining Visual Entailment

Kayser et al. Ongoing

Hypothesis: The people are flying kites at the beach.
Answer: Contradiction
RVT: People can't be riding kites while they are flying kites.
PJ-X: People cannot be flying and flying at the same time.
FME: People cannot be walking and flying kites at the same time
e-UG: People cannot be flying kites while they are standing on a street.
GT Explanation: construction site is different from the beach

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
FME: a woman can be a man or a woman
RVT: Just because a lady is holding a book does not mean she is the owner of the store.
e-UG: Just because a lady is working at a store does not mean she is the owner.

Conclusions for: Generating NL Explanations for Visual Decisions

Natural Language Explanations are

1. Class-specific, image-relevant, groundable and contrastive
2. Generalizable to image and video data as well as visual question answering
3. An effective means for evaluating the conceptual understanding of the model

Outline

Visual Dialog

Proposes

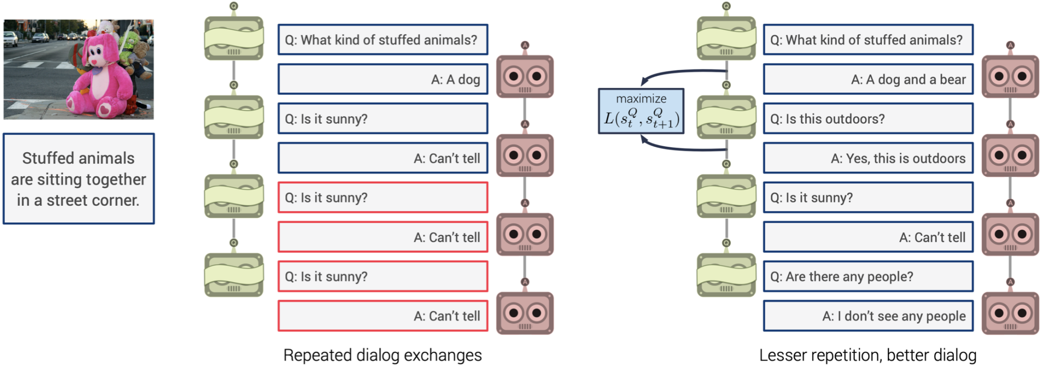
- a large-scale dataset
- data collection platform
- benchmark study on the Visual Dialog dataset

Finds that

- Naively incorporating history doesn't help
- Looking at the image is necessary

Visual Dialog; Das, Kottur, Gupta, Singh, Yadav, Foura, Parikh, Batra at IEEE CVPR 2017

Diversity Improved Visual Dialog

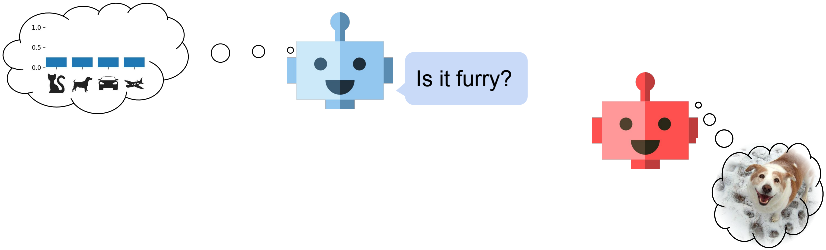


Improving Generative Visual Dialog by Answering Diverse Questions

Murahari, Chattopadhyay, Batra, Parikh, Das at EMNLP 2019

Visual Explanation Through Communication

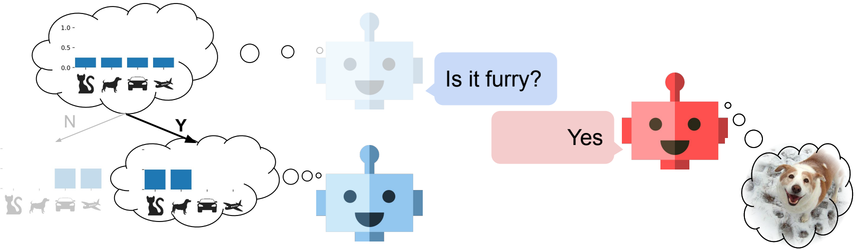
Alaniz et al. CVPR'21



Learning Decision Trees Recurrently Through Communication;
Alaniz, Marcos, Schiele, Akata at CVPR 2021

Visual Explanation Through Communication

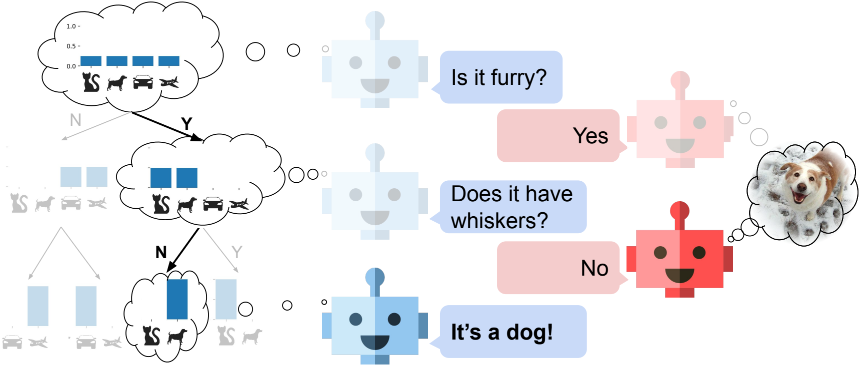
Alaniz et al. CVPR'21



Learning Decision Trees Recurrently Through Communication;
Alaniz, Marcos, Schiele, Akata at CVPR 2021

Visual Explanation Through Communication

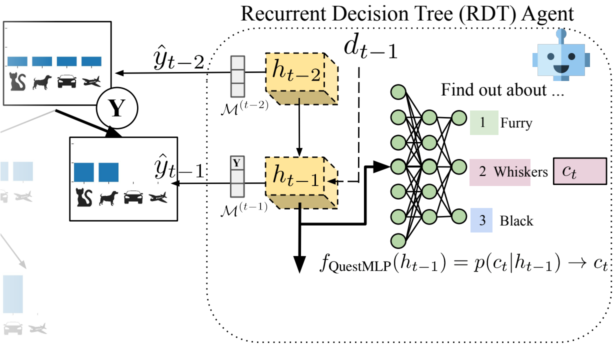
Alaniz et al. CVPR'21



Learning Decision Trees Recurrently Through Communication;
Alaniz, Marcos, Schiele, Akata at CVPR 2021

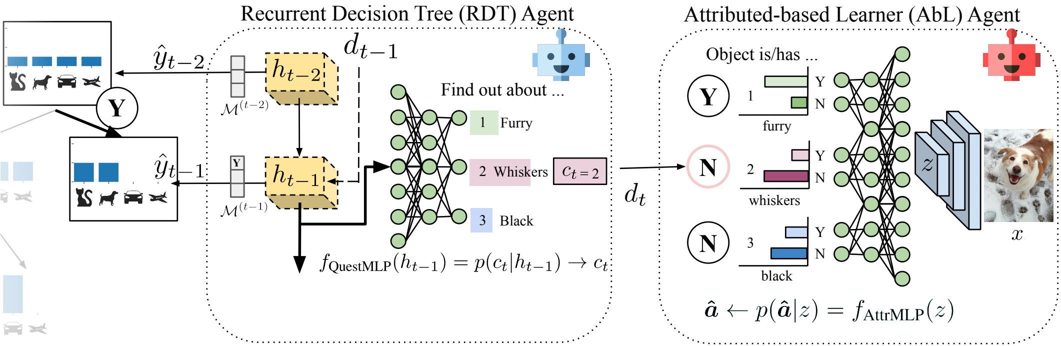
Recurrent Decision Tree with Attributes

Alaniz et al. CVPR'21



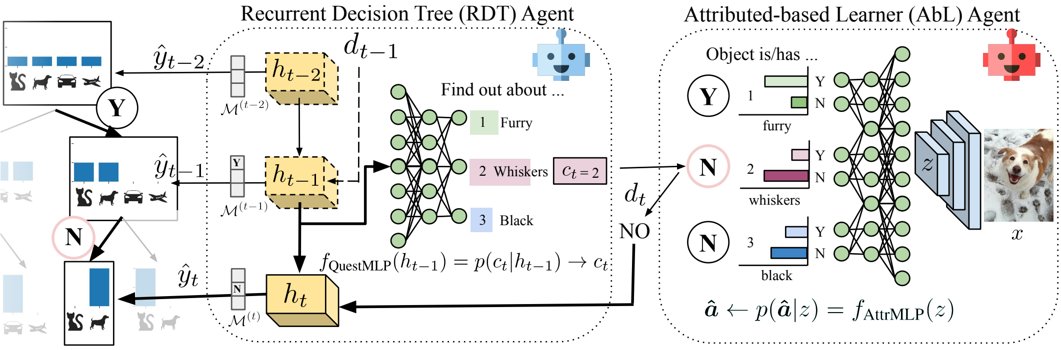
Recurrent Decision Tree with Attributes

Alaniz et al. CVPR'21



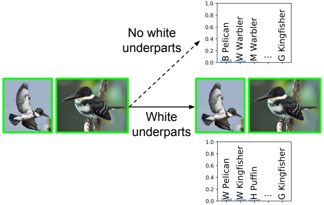
Recurrent Decision Tree with Attributes

Alaniz et al. CVPR'21



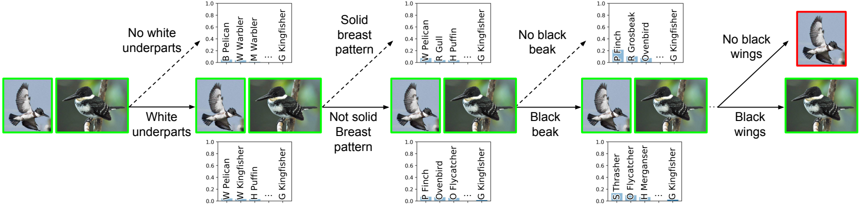
CUB Decision Sequence

CUB Decision Sequence



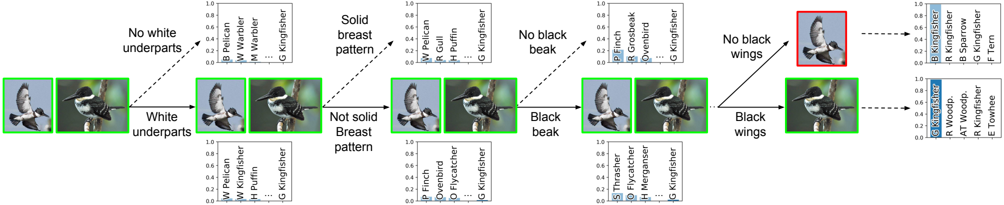
CUB Decision Sequence

Alaniz et al. CVPR'21

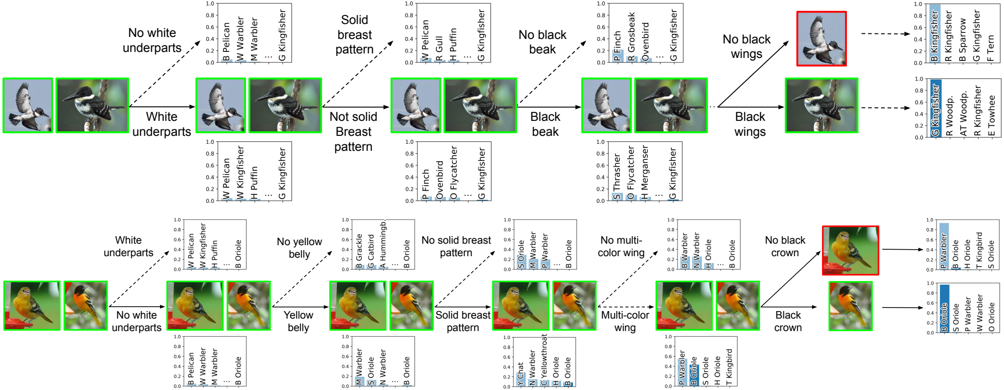


CUB Decision Sequence

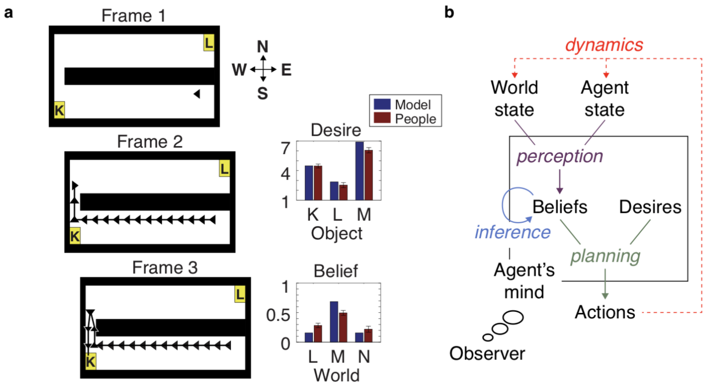
Alaniz et al. CVPR'21



CUB Decision Sequence



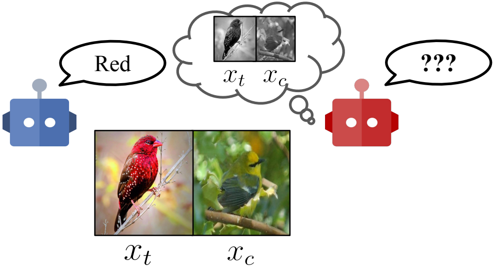
Machine Theory of Mind



Rational quantitative attribution of beliefs, desires and percepts in human mentalizing
Baker, Jara-Ettinger, Tenenbaum; Nature Human Behaviour, 2017

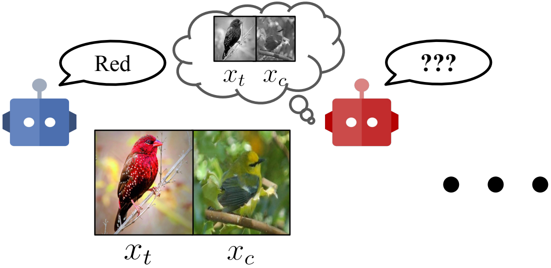
Machine Theory of Mind; Rabinowitz, Perbet, Song, Zhang, Eslami, Botvinick; ICML 2018

Modeling Conceptual Understanding



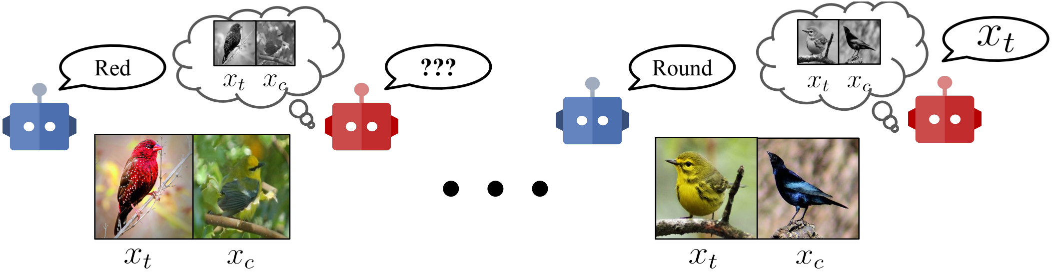
Modeling Conceptual Understanding in Image Reference Games;
Corona, Alaniz, Akata; NeurIPS 2019

Modeling Conceptual Understanding



Modeling Conceptual Understanding in Image Reference Games;
Corona, Alaniz, Akata; NeurIPS 2019

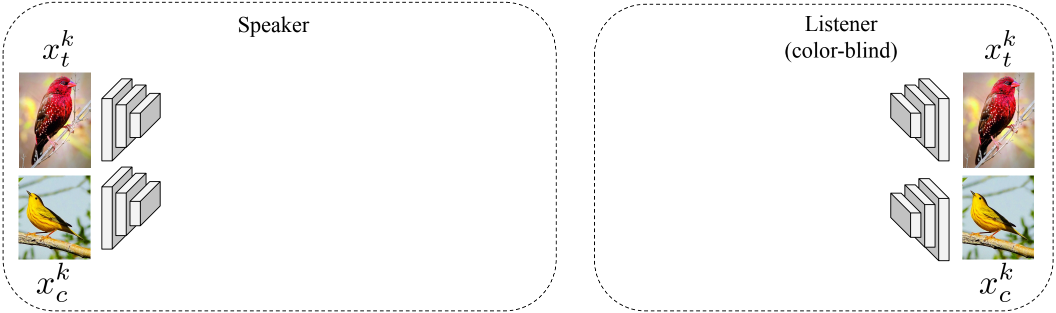
Modeling Conceptual Understanding



Modeling Conceptual Understanding in Image Reference Games;
Corona, Alaniz, Akata; NeurIPS 2019

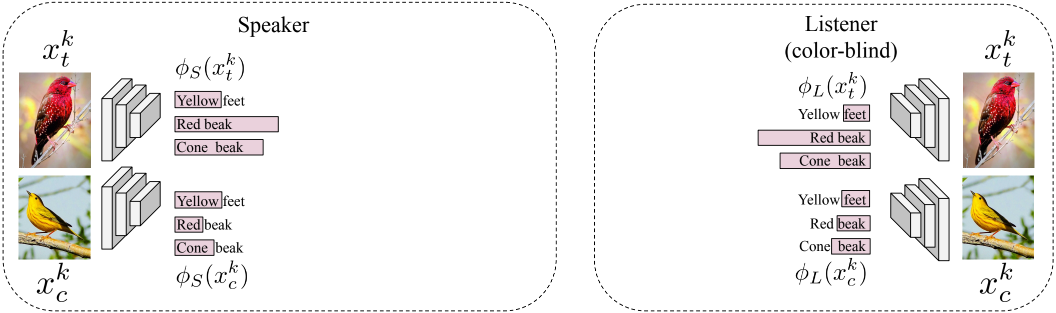
Modeling Conceptual Understanding

Corona et al. NeurIPS'19



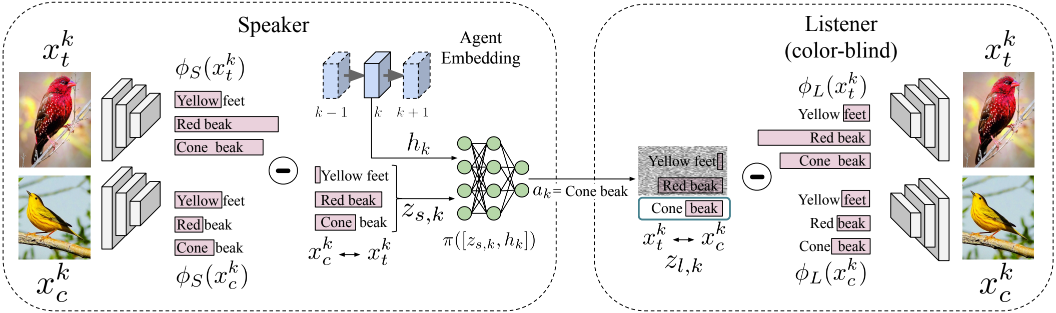
Modeling Conceptual Understanding

Corona et al. NeurIPS'19



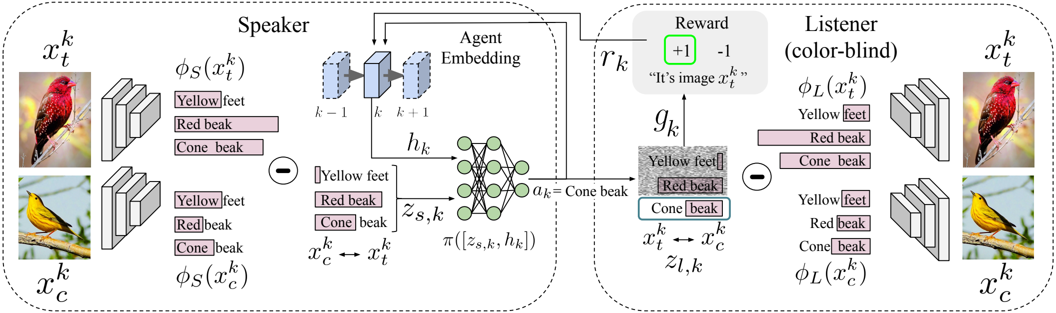
Modeling Conceptual Understanding

Corona et al. NeurIPS'19



Modeling Conceptual Understanding

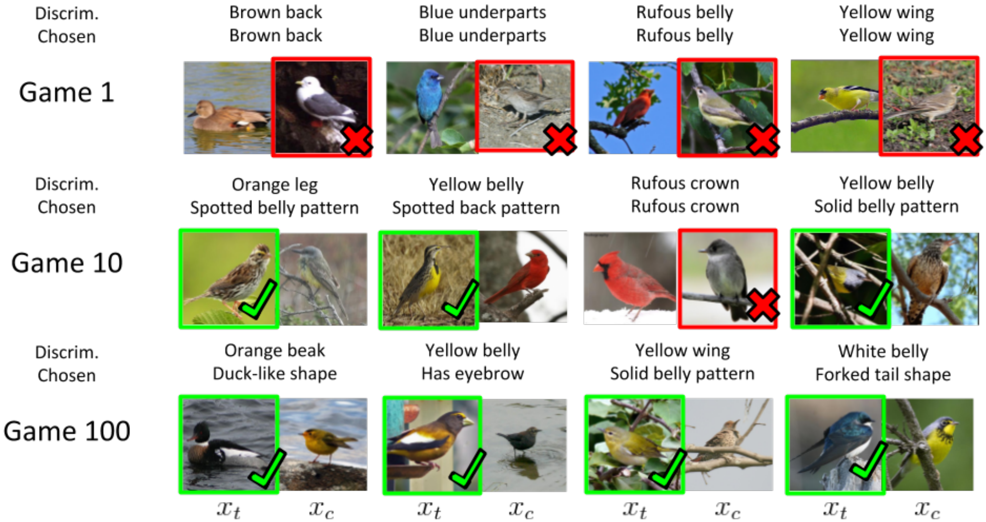
Corona et al. NeurIPS'19



- Speaker adapts to the listener by incorporating information after each game

Modeling Conceptual Understanding Results

Corona et al. NeurIPS'19



Conclusions for: Modeling Conceptual Understanding

Modeling the conceptual understanding of the user is

1. An important step towards a more natural communication
2. Necessary to build support and trust between the user and the machine
3. Difficult to evaluate: proxy tasks like classification maybe a solution

Outline

Summary

1. Learning with basic, simple, contrastive elements of understanding is important
   [Explanation via attributes, text embeddings, natural language]

Summary

1. Learning with basic, simple, contrastive elements of understanding is important
   [Explanation via attributes, text embeddings, natural language]
2. Vision is complementary for generating natural language explanations
   [Fine-grained image based, video-based and visual question answering explanations]

Summary

1. Learning with basic, simple, contrastive elements of understanding is important
   [Explanation via attributes, text embeddings, natural language]
2. Vision is complementary for generating natural language explanations
   [Fine-grained image based, video-based and visual question answering explanations]
3. Developing explainable deep models is important for user acceptance
   [Understanding the mental model of the user is necesary for effective communication]

Future of Deeply Explainable Artificial Intelligence

Future of Deeply Explainable Artificial Intelligence

User: What happened?

Future of Deeply Explainable Artificial Intelligence

User: What happened?

AI: I was driving down an empty road. I decided to slow down as a ball appeared on the right. I saw a child running towards the ball, so I decided to stop.

Future of Deeply Explainable Artificial Intelligence

User: What happened?

AI: I was driving down an empty road. I decided to slow down as a ball appeared on the right. I saw a child running towards the ball, so I decided to stop.

User: What would have happened if you did not stop ?

Future of Deeply Explainable Artificial Intelligence

User: What happened?

AI: I was driving down an empty road. I decided to slow down as a ball appeared on the right. I saw a child running towards the ball, so I decided to stop.

User: What would have happened if you did not stop ?

AI: If there was an impact, the child would have gotten hurt.

Thank you!