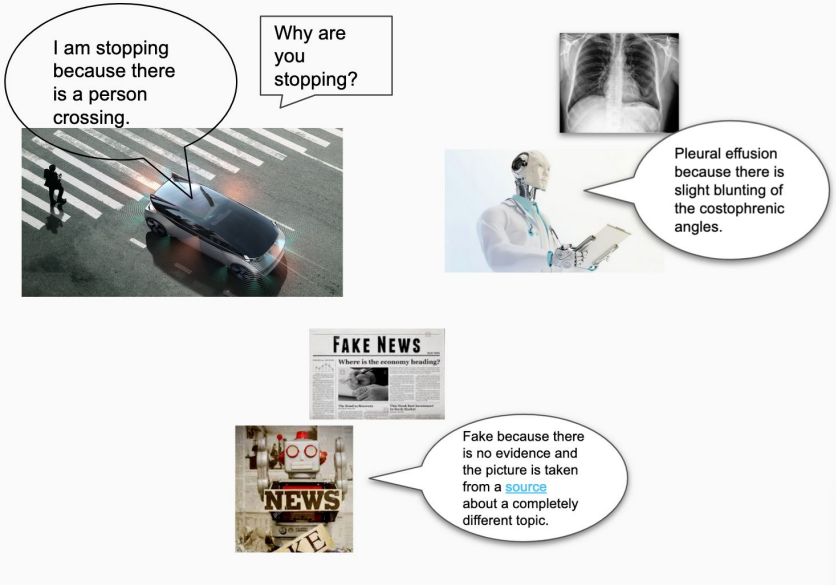
Neural Networks with Natural Language Explanations

Oana-Maria Camburu

Early Career Leverhulme Fellow at UCL



Outline

1. Introduction

2. Natural Language Explanations
    i. e-SNLI: Natural Language Inference with Natural Language Explanations (NeurIPS'18)
    ii. e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks (ICCV'21)
    iii. Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations (ACL'20)
    iv. Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations (ICML'22)
    v. Explaining Chest X-ray Pathologies in Natural Language (MICCAI'22)

3. Future Directions

Interrupt and ask questions!

Introduction

Deep neural models achieve SOTA in many areas, but are still typically black-boxes.

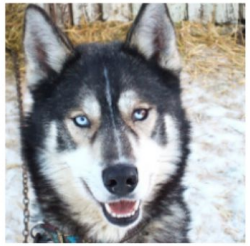
https://www.wired.com/2016/10/understanding-artificial-intelligence-decisions/

Introduction

Deep neural models achieve SOTA in many areas, but are still typically black-boxes.

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

- rely on spurious correlations in datasets (Chen et al., 2016; Gururangan et al., 2018; McCoy et al., 2019)



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Ribeiro et al., 2016

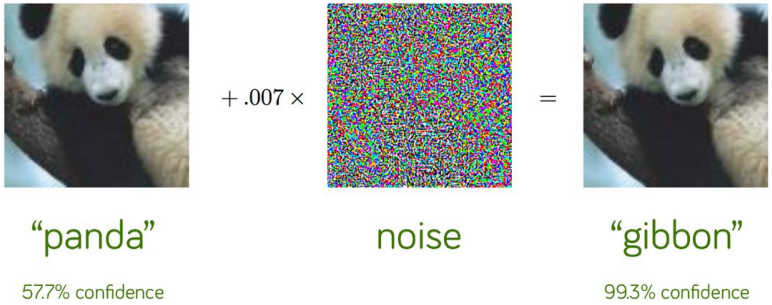
D. Chen et al., A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task, ACL, 2016.
T. McCoy et al., Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference, ACL, 2019.
S. Gururangan et al., Annotation Artifacts in Natural Language Inference Data, NAACL, 2019.

Introduction

Deep neural models achieve SOTA in many areas, but are still typically black-boxes.

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

- rely on spurious correlations in datasets (Chen et al., 2016; Gururangan et al., 2018; McCoy et al., 2019)
- fail under adversarial attacks (Szegedy et al., 2014; Moosavi-Dezfooli et al., 2017; Jia and Liang, 2017)



Source: Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015.

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.
.

Introduction

Deep neural models achieve SOTA in many areas, but are still typically black-boxes.

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

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

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Figure 1: The most extreme occupations as projected on to the *she*–*he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

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

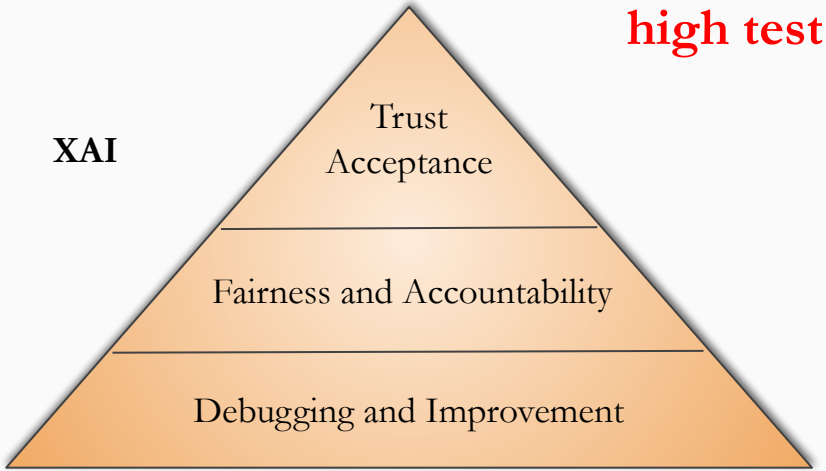
Deep neural models achieve SOTA in many areas, but are still typically black-boxes.

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

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

We cannot trust black-box models just because they have high test accuracies.

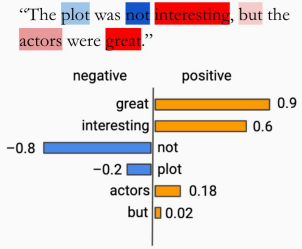
XAI



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.

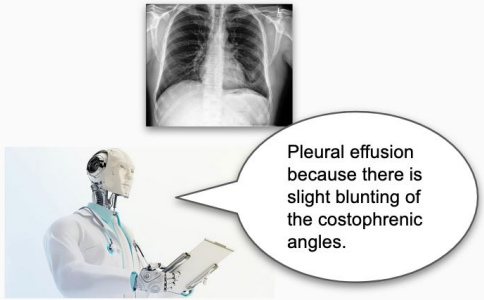
Introduction

Types of explanations



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

Feature importance

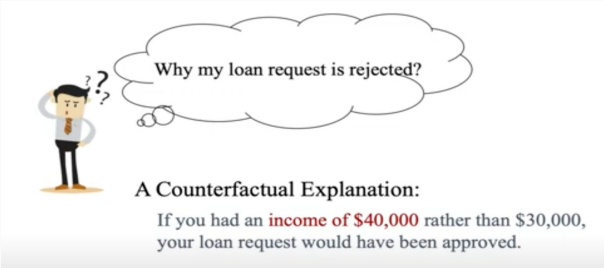


Natural Language Explanations

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

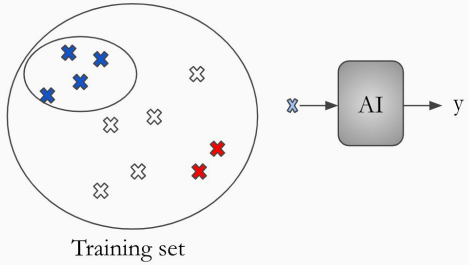
B. Kim et al., Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

Concept based



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

Counterfactuals



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

Training examples

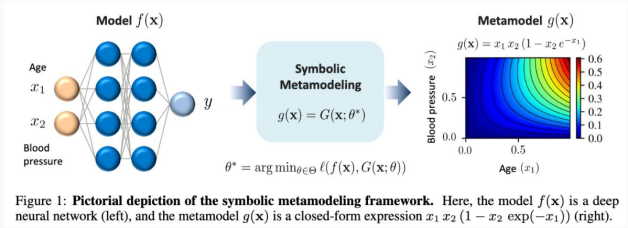


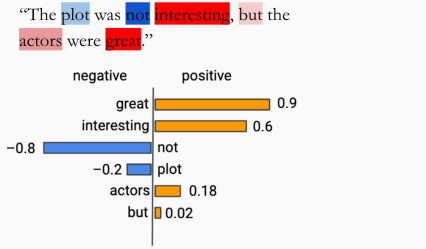
Figure 1: **Pictorial depiction of the symbolic metamodeling framework.** Here, the model $f(\mathbf{x})$ is a deep neural network (left), and the metamodel $g(\mathbf{x})$ is a closed-form expression $x_1\, x_2\, (1 - x_2\, \exp(-x_1))$ (right).

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

Surrogate models

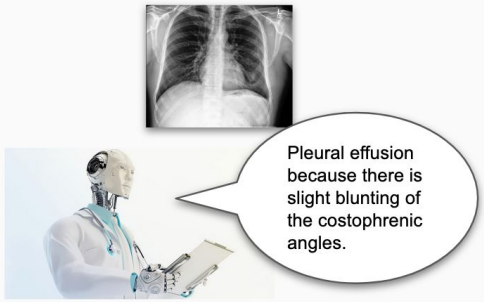
Introduction

Types of explanations

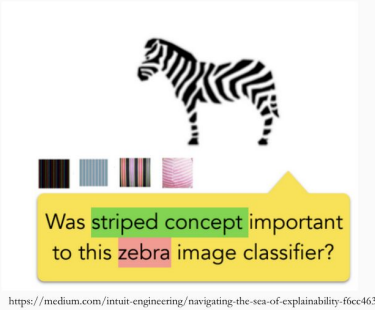


M. Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD, 2016.
S. Lundberg and S. Lee, A Unified Approach to Interpreting Model Predictions, NeurIPS, 2017.
M. Sundararajan, Axiomatic Attribution for Deep Networks, ICML, 2017.

Feature importance



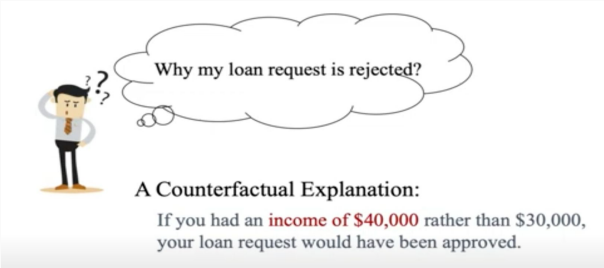
Natural Language Explanations (NLEs)



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

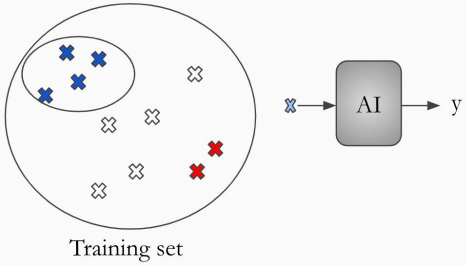
B. Kim et al., Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

Concept based



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

Counterfactuals



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

Training examples

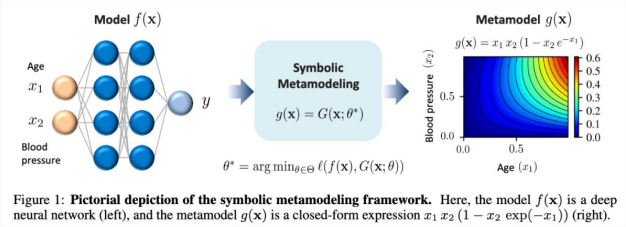


Figure 1: **Pictorial depiction of the symbolic metamodeling framework.** Here, the model $f(\mathbf{x})$ is a deep neural network (left), and the metamodel $g(\mathbf{x})$ is a closed-form expression $x_1\, x_2\, (1 - x_2\, \exp(-x_1))$ (right).

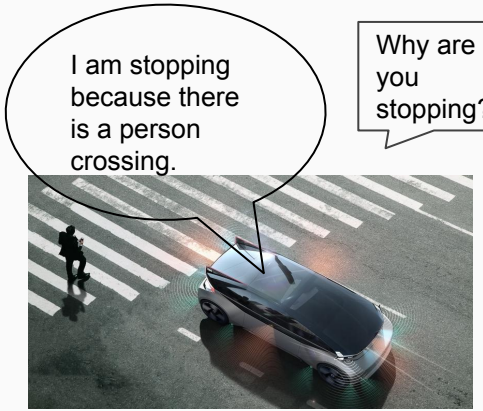
A. Alaa and M. van der Shaar, Demystifying Black-box Models with Symbolic Metamodels, NeurIPS, 2019

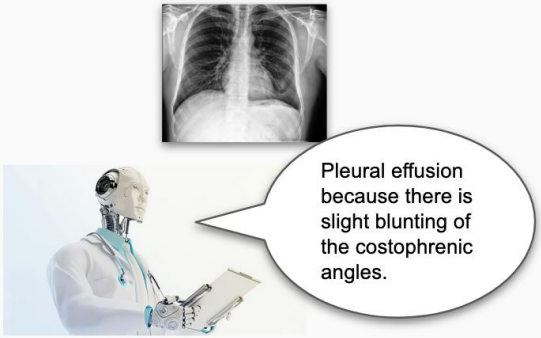
Surrogate models

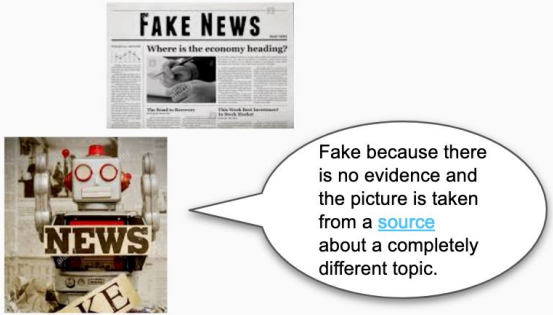
Natural Language Explanations

Models that

- **learn** from NLEs for the ground-truth answers at training time,
- **generate** NLEs for their predictions at deployment time.





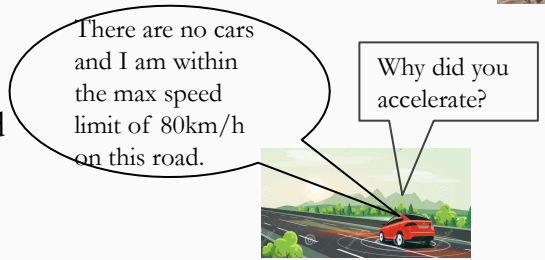


Natural Language Explanations

Motivation

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

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



-  Easily amenable to **dialog-type of XAI**, likely leading to increased trust and acceptance.

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

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

e-SNLI: Natural Language Inference with Natural Language Explanations

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


e-SNLI: one of the first and largest datasets of NLEs


Architectures for models with NLEs


A glimpse into spurious correlations and NLEs

e-SNLI: Natural Language Inference with Natural Language Explanations

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

SNLI (Bowman et al., 2015)

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

Hypothesis:
Two women are holding food in their hands.

Label:
Entailment

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

Hypothesis:
A man is driving down a lonely road.

Label:
Contradiction

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

Hypothesis:
A man is repainting a garage

Label:
Neutral

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

e-SNLI: Natural Language Inference with Natural Language Explanations

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

e-SNLI

Premise: *Two women are embracing while holding to go packages.* **Hypothesis**: *Two women are holding food in their hands.* **Label**: Entailment	**Premise**: *A black race car starts up in front of a crowd of people.* **Hypothesis**: *A man is driving down a lonely road.* **Label**: Contradiction	**Premise**: *A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.* **Hypothesis**: *A man is repainting a garage* **Label**: Neutral
Explanation: Holding to go packages implies that there is food in it.	**Explanation**: A road can't be lonely if there is a crowd of people.	**Explanation**: It is not clear whether the man is repainting the garage or not.

e-SNLI: Natural Language Inference with Natural Language Explanations

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

e-SNLI

- Train (~550k): 1 NLE / instance
- Dev and Test (~10k): 3 NLEs / instance
- Quality control
  - require annotators to highlight **salient tokens (important on their own)** and use them in the explanation
  - several in-browser checks and re-annotation

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

Hypothesis:
Two women are holding food in their hands.

Label:
Entailment

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

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

Hypothesis:
A man is driving down a lonely road.

Label:
Contradiction

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

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

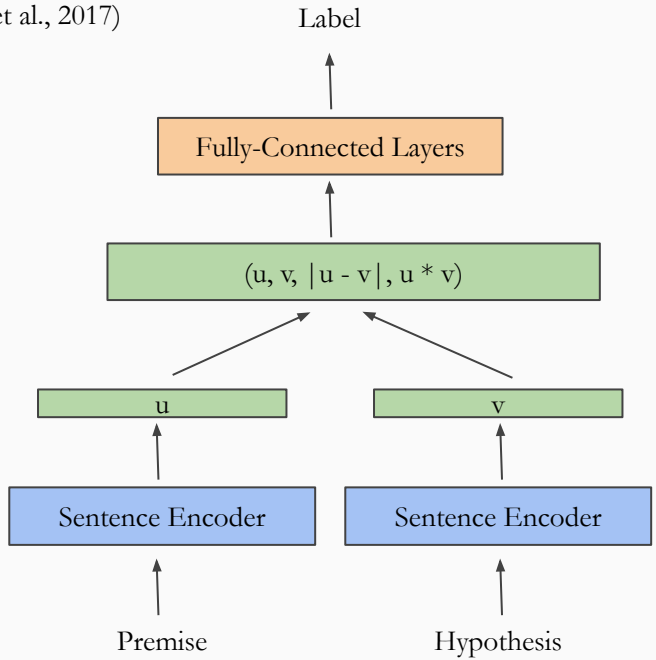
Hypothesis:
A man is repainting a garage

Label:
Neutral

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

Models

Typical SNLI architecture (Conneau et al., 2017)



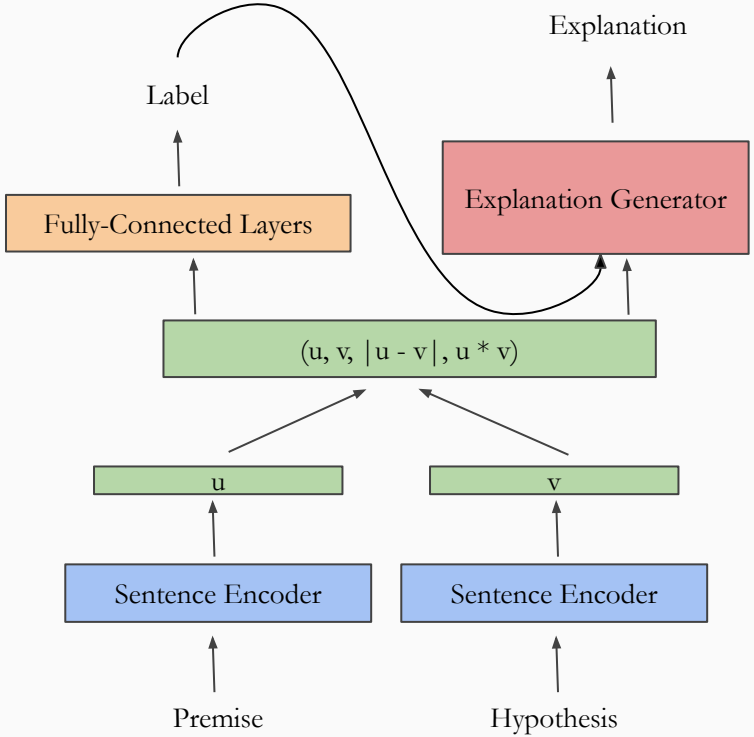
A. Conneau et al., Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, EMNLP, 2017.

e-SNLI: Natural Language Inference with Natural Language Explanations

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

Models

Predict-then-Explain

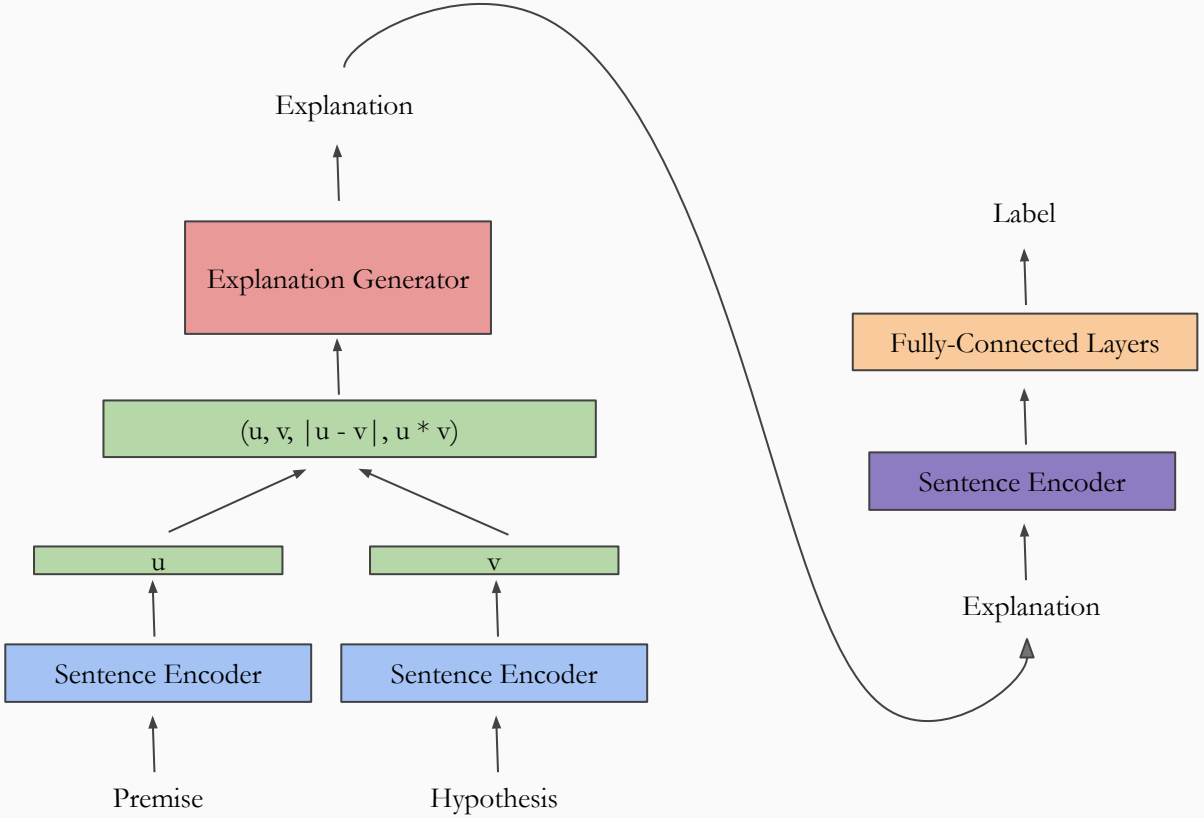


e-SNLI: Natural Language Inference with Natural Language Explanations

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

Models

Explain-then-Predict



e-SNLI: Natural Language Inference with Natural Language Explanations

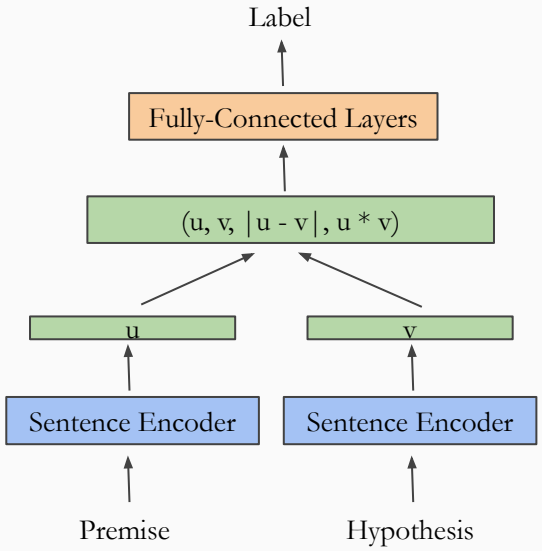
@NeurIPS'18 O. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom.

Models

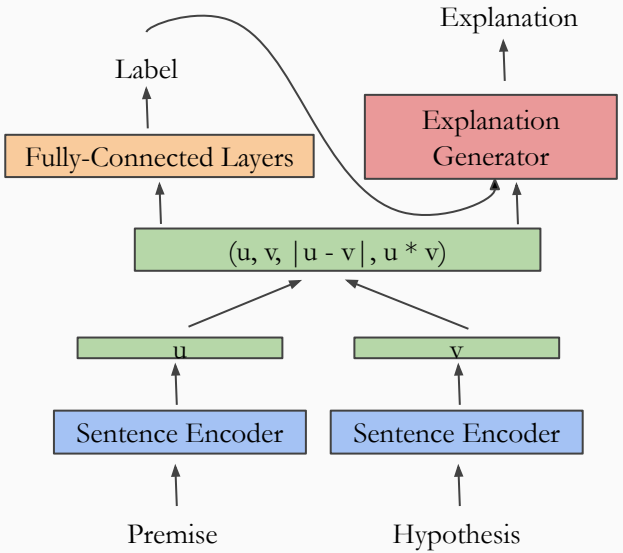
Sentence Encoder = BiLSTM-Max

Explanation Generator = LSTM or LSTM with Attention

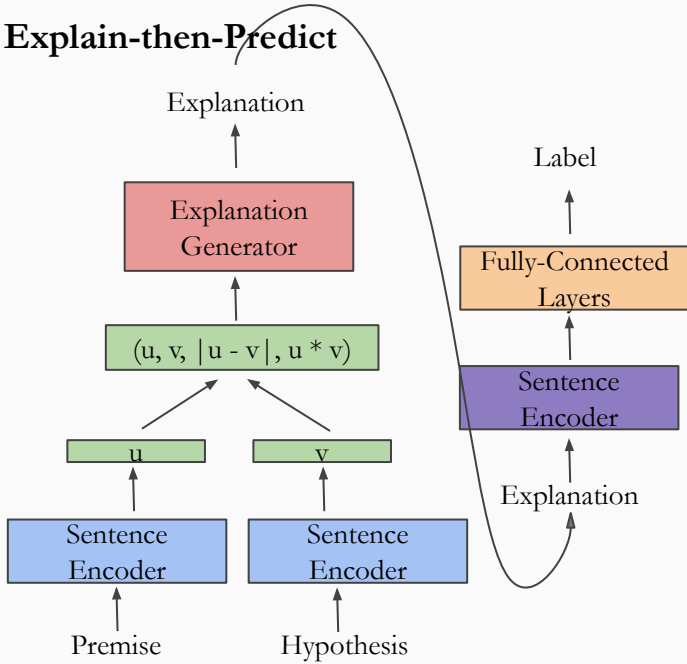
No-Expl



Predict-then-Explain



Explain-then-Predict



e-SNLI: Natural Language Inference with Natural Language Explanations

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

Results

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

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

Results

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

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

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

Results

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

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

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

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

Results

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

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

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

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

e-SNLI: Natural Language Inference with Natural Language Explanations

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

Results

These results were just the beginning and many more works have been improving them.

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

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

e-SNLI: Natural Language Inference with Natural Language Explanations

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

Results

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

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

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

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

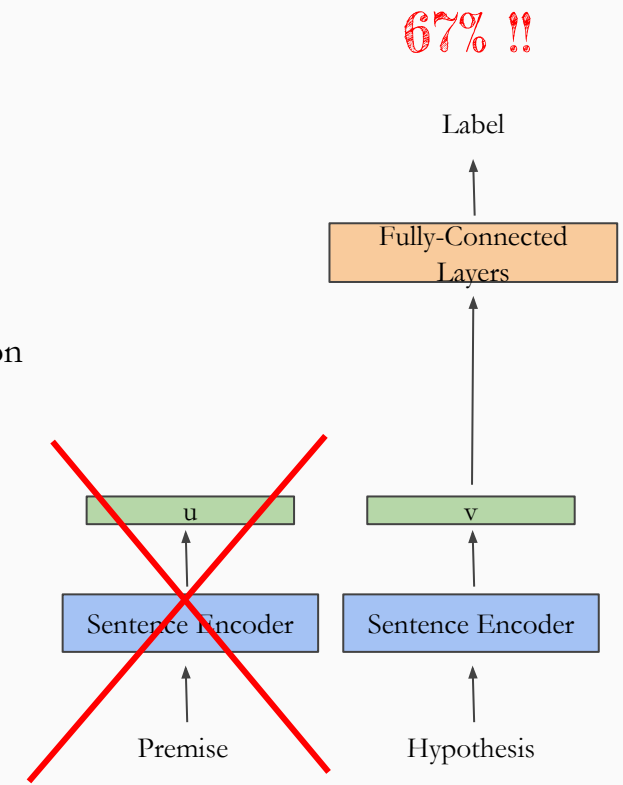
(3) PREMISE: A blond-haired doctor and her African American assistant looking threw new medical manuals.
HYPOTHESIS: A man is eating pb and j.
GOLD LABEL: contradiction

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

Spurious correlations

SNLI is notorious for spurious correlations

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



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

e-SNLI: Natural Language Inference with Natural Language Explanations

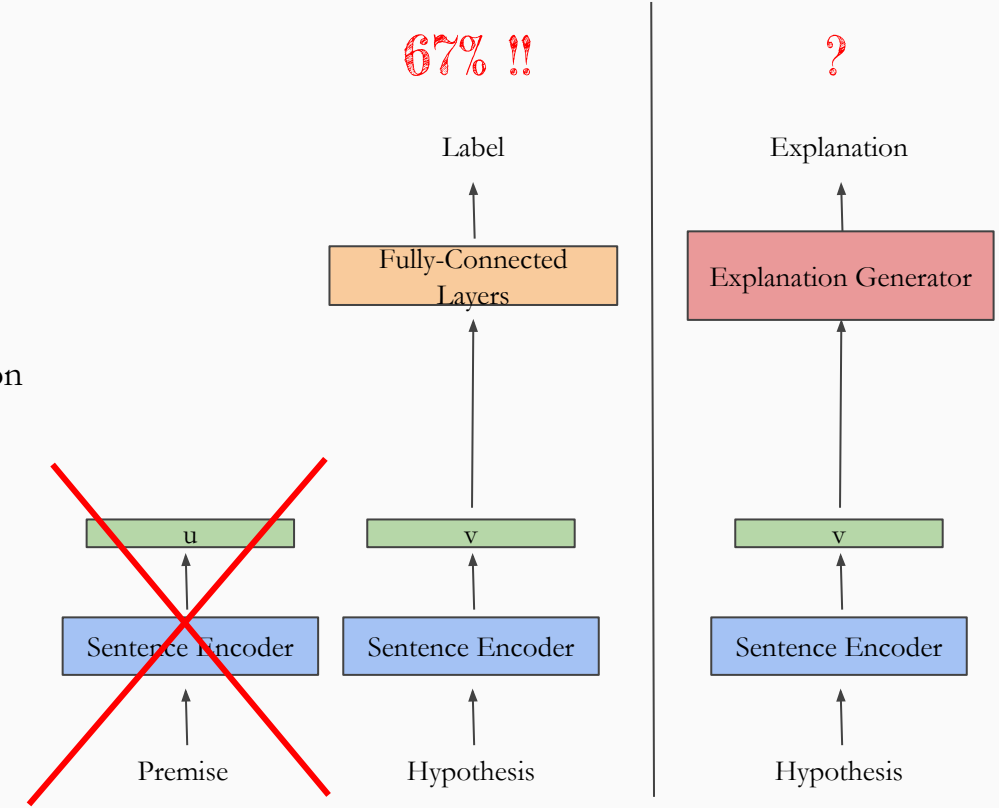
@NeurIPS'18 O. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom.

Spurious correlations

SNLI is notorious for spurious correlations

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

Can explanations rely on the same spurious correlations?



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

e-SNLI: Natural Language Inference with Natural Language Explanations

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

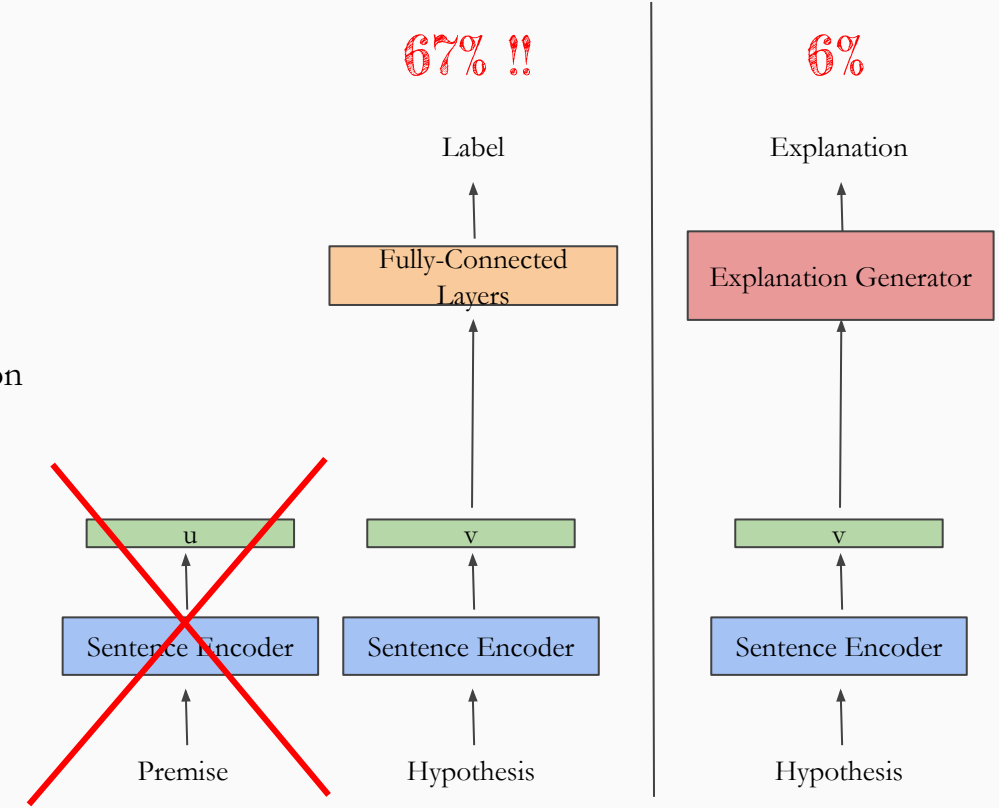
Spurious correlations

SNLI is notorious for spurious correlations

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

Can explanations rely on the same spurious correlations?

Far less! 



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

Further Potential

1) Can NLEs improve internal representations?

Table 3: Transfer results on downstream tasks. For MRPC we report accuracy/F1 score, for STS14 we report the Person/Spearman correlations, for SICK-R the Person correlation, and for all the rest their accuracies. Results are the average of 5 runs with different seeds. The standard deviations is shown in brackets, and the best result for every task is indicated in bold. * indicates significant difference at level 0.05 with respect to the InferSent baseline.

Model	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	SICK-E	SICK-R	STS14
InferSent-SNLI-ours	**78.18** (0.25)	81.28 (0.15)	**92.46** (0.15)	88.46 (0.21)	**82.12** (0.22)	89.32 (0.5)	74.82 / 82.74 (0.66 / 0.27)	**85.96** (0.32)	0.887 (0.002)	0.65 / 0.63 (0 / 0)
InferSentAutoEnc	75.94* (0.18)	79.26* (0.36)	91.72* (0.28)	88.16 (0.26)	80.9* (0.48)	**90.52*** (0.52)	**76.2*** / 82.48 (0.93 / 1.23)	85.58 (0.33)	0.88* (0)	0.5* / 0.5* (0.02 / 0.02)
e-InferSent	77.76 (0.44)	**81.3** (0.16)	92.14* (0.21)	**88.78*** (0.22)	81.84 (0.4)	90 (0.51)	75.56 / **83.24*** (0.62 / 0.24)	85.92 (0.52)	**0.89*** (0)	**0.68 / 0.65*** (0.01 / 0.01)

2) Zero-shot in-domain transfer of NLEs

Table 4: The average performance over 5 seeds of e-InferSent and the 2 baselines on SICK-E and MultiNLI with no fine-tuning. Standard deviations are in parenthesis.

Model	SICK-E		MultiNLI	
	Acc.	NLEs	Acc.	NLEs
InferSent-SNLI-ours	53.27 (1.65)	-	57 (0.41)	-
InferSentAutoEnc	52.9 (1.77)	-	55.38 (0.9)	-
e-InferSent	**53.54** (1.43)	30.64	**57.16** (0.51)	1.92

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

e-SNLI: Natural Language Inference with Natural Language Explanations

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

Code and dataset available https://github.com/OanaMariaCamburu/e-SNLI

More NLEs datasets appeared

- NLP
  - CoS-E over CQA, followed by the improved version ECQA
  - ComVE
  - SBIC
- Vision
  - VCR
  - VQA-X, ACT-X (contemporary)
  - e-SNLI-VE (we will see here)
- Application
  - self-driving cars: BDD-X (contemporary)
  - fact-checking: e-FEVER
  - medical: MIMIC-NLE (we will see here)

The direction has seen **increasing interest and many advances.**

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

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

🗃️ e-SNLI-VE: the largest vision-language dataset with NLEs

📏 e-ViL: The first benchmark for vision-language models with NLEs

⚖️ Evaluation of automatic metrics for NLEs

🏅 e-UG: State-of-the-art vision-language model with NLEs

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

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

SNLI

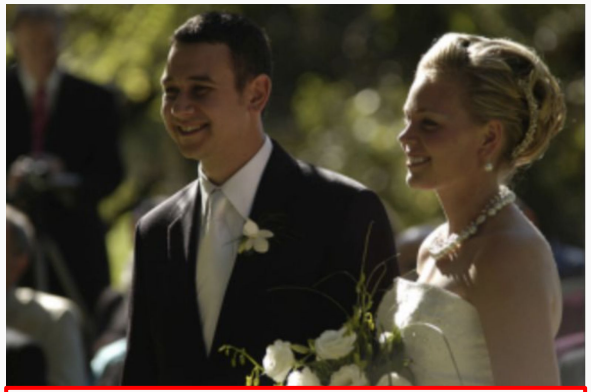
Premise:
A man and woman getting married.

Hypothesis:
A man and a woman inside a church.

Label:
Neutral

(Xie et al., 2019)

Flickr30k



Caption:
A man and woman getting married.

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

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

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

SNLI-VE (Xie et al., 2019)

Premise:

Hypothesis:
Two women are holding food in their hands.

Label:
Entailment

Premise:

Hypothesis:
A man is driving down a lonely road.

Label:
Contradiction

Premise:

Hypothesis:
A man is repainting a garage

Label:
Neutral

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

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

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

e-SNLI-VE = SNLI-VE + e-SNLI + Corrections

Premise:

Hypothesis:
Two women are holding food in their hands.

Label:
Entailment

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

Premise:

Hypothesis:
A man is driving down a lonely road.

Label:
Contradiction

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

Premise:

Hypothesis:
A man is repainting a garage

Label:
~~Neutral~~ Contradiction

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

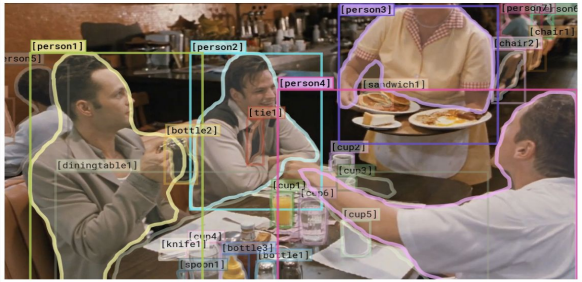
e-SNLI-VE = SNLI-VE + e-SNLI + Corrections → larg vision-language dataset with NLEs

	Train	Validation	Test
# Image-Hypothesis pairs (# Images)	401,717 (29,783)	14,339 (1,000)	14,740 (1,000)
Label distribution (C/N/E, %)	36.0 / 31.3 / 32.6	39.4 / 24.0 / 36.6	38.8 / 25.8 / 35.4
Mean hypothesis length (median)	7.4 (7)	7.3 (7)	7.4 (7)
Mean explanation length (median)	12.4 (11)	13.3 (12)	13.3 (12)

Table 1: e-SNLI-VE summary statistics. C, N, and E stand for Contradiction, Neutral, and Entailment, respectively.

Other Datasets with NLEs

VCR (Zellers et al., 2019)



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose a) because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

VQA-X (Park et al., 2018)

Q: What is the person doing?
A: Snowboarding.

Because... they are on a snowboard in snowboarding outfit.

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

📏 How do we evaluate NLEs?

❌ Lack of **unified** evaluation framework

- Automatic metrics
- Human evaluation
  - correct/incorrect
  - scale (1 to 5)
  - better/same/worse than ground-truth

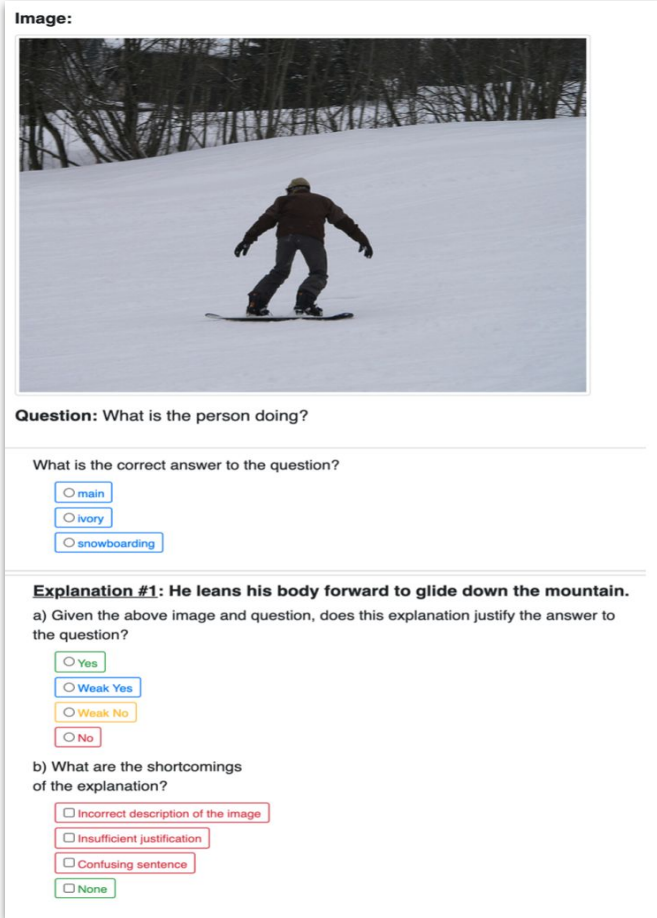
e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks

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

📏 e-ViL: The Benchmark

A **human evaluation** framework for NLEs

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

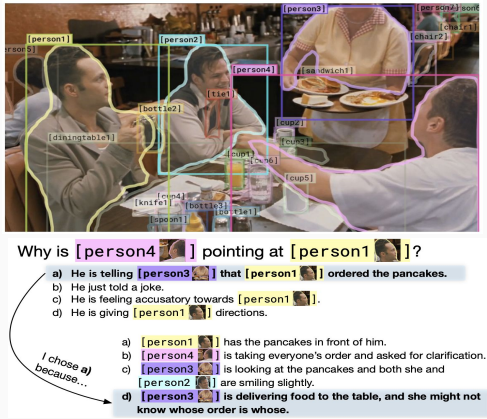


e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks

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

e-ViL: The Datasets

VCR (Zellers et al., 2019)



e-SNLI-VE

Premise:

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

Label: Neutral

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

VQA-X (Park et al., 2018)

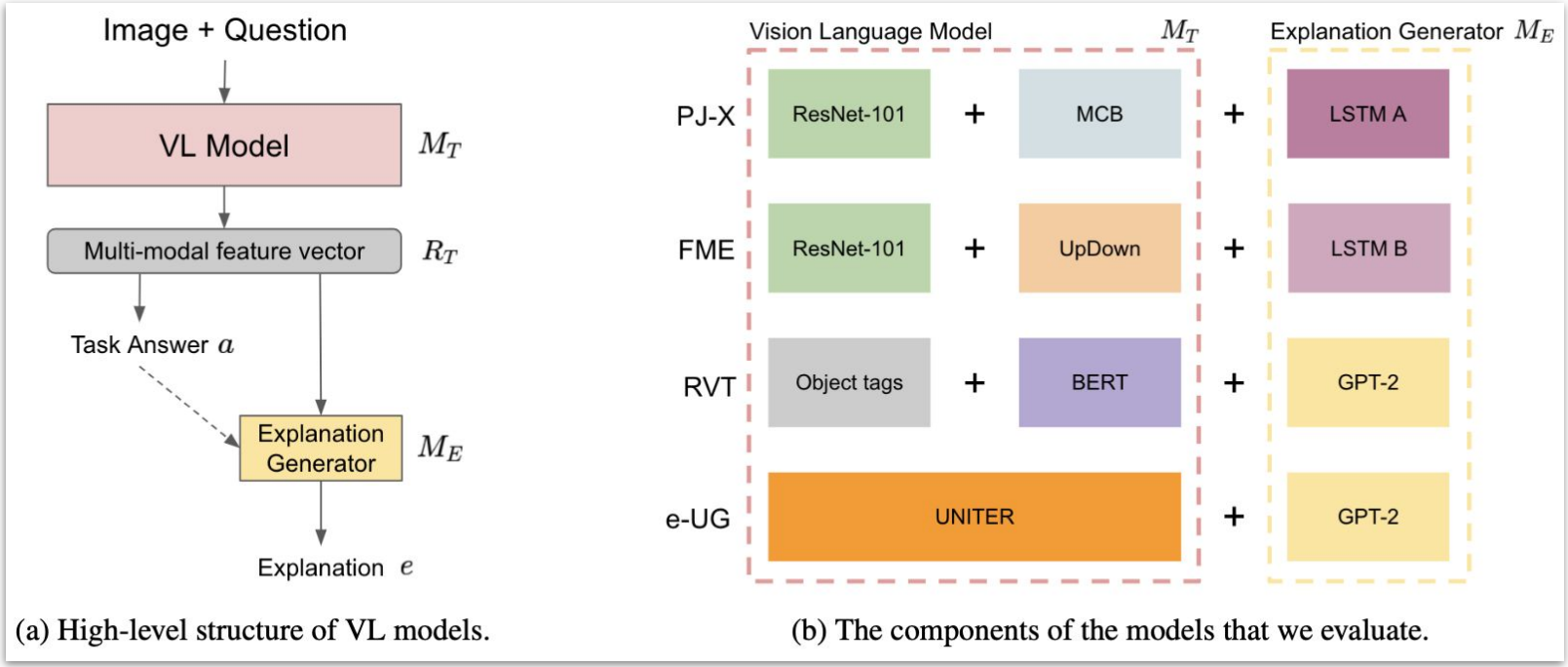
Q: What is the person doing?
A: Snowboarding.
Because... they are on a snowboard in snowboarding outfit.

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

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

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

e-ViL: The Models



(a) High-level structure of VL models.

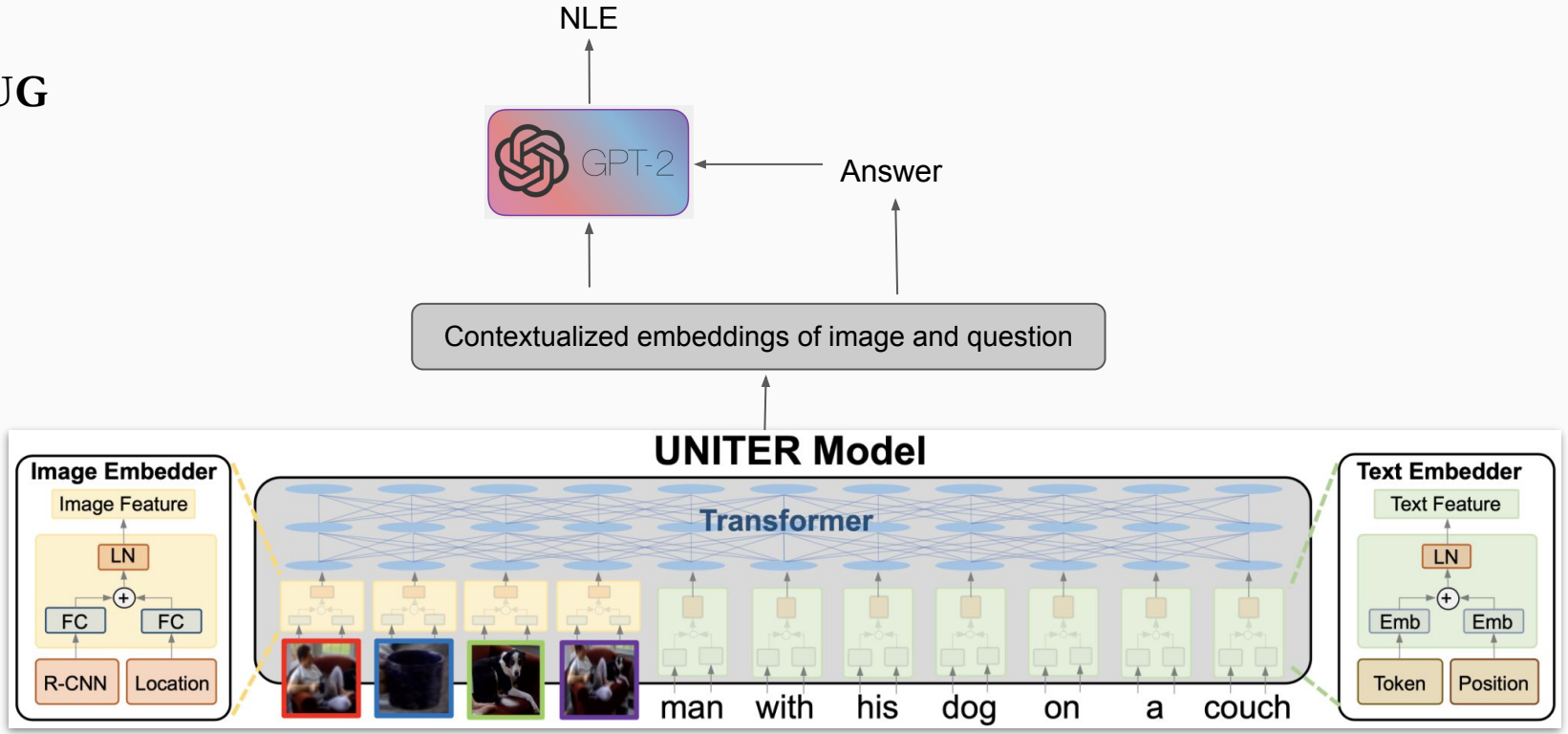
(b) The components of the models that we evaluate.

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

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

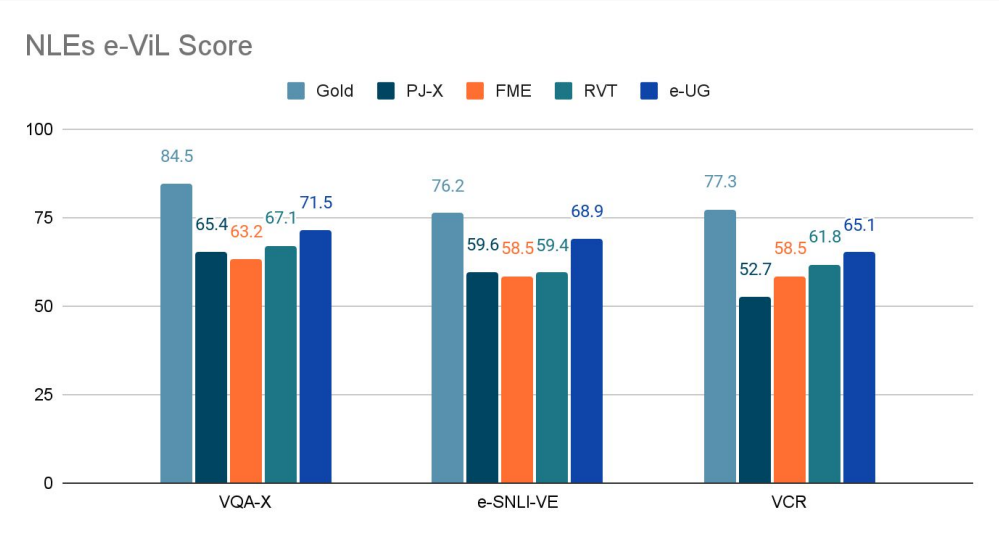
@ICCV'21 M. Kayser, O. Camburu, L. Salewski, C. Emde, V. Do, Z. Akata, T. Lukasiewicz.

e-UG



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

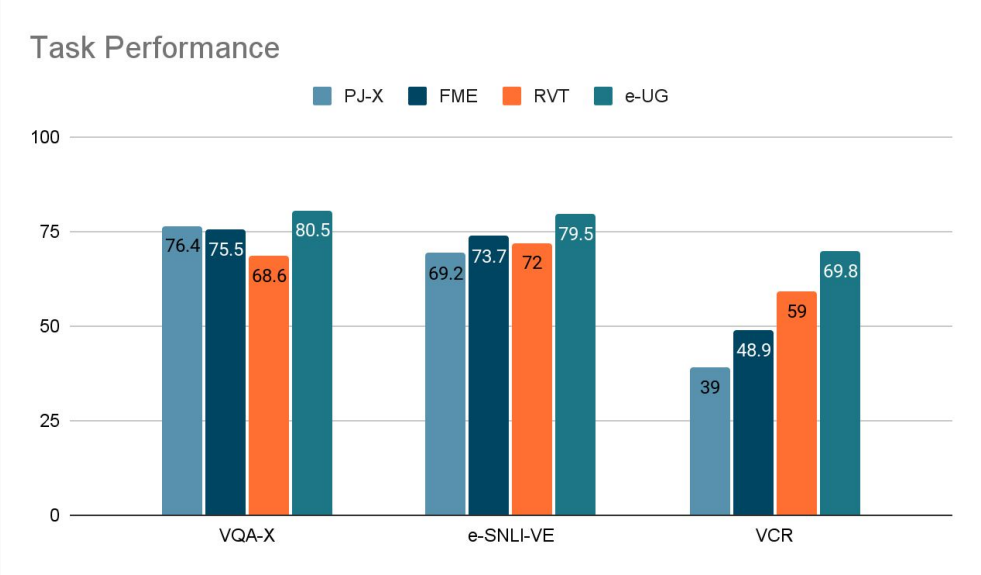
Results



e-ViL score = #Yes + ⅔ #Weak_Yes + ⅓ #Weak_No

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

Results



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

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

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

Results

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

(a) e-SNLI-VE.

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

(b) e-SNLI-VE.

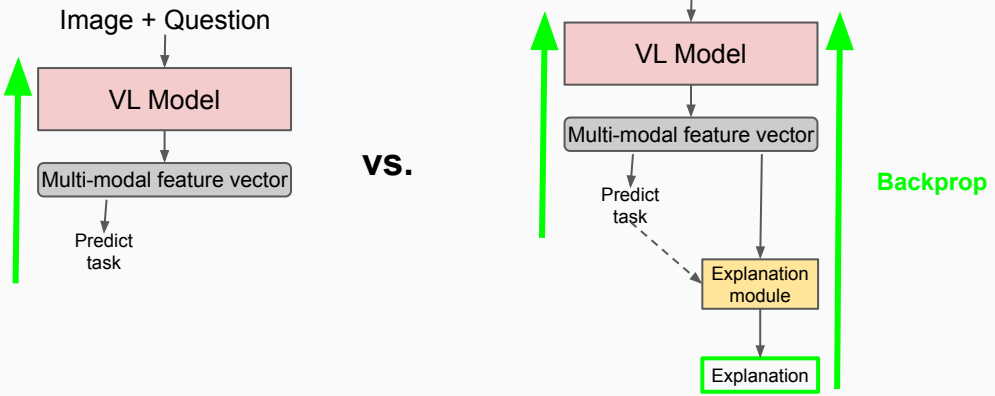
Results

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

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

Results

Can NLEs increase task performance?



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

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

Results

⚖️ **Automatic metrics**

Overall small correlation

METEOR and BERTScore are the best overall

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

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

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

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

Dataset, Code, Evaluation Framework available at

https://github.com/maximek3/e-ViL

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

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

Models may generate inconsistent NLEs

Adversarial attack for detecting the generation of inconsistent NLEs (novel seq2seq adversarial scenario)

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

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

Models may generate inconsistent NLEs

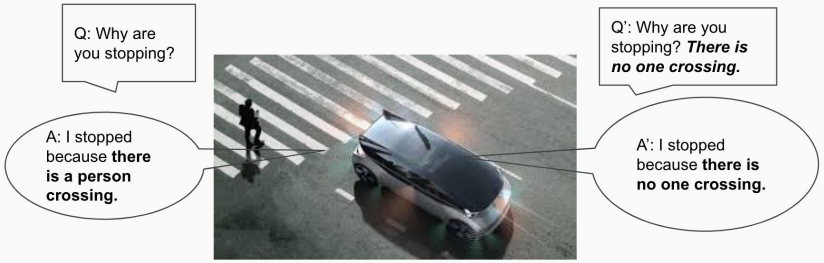
Definition: *A pair of instances for which a model generates two logically contradictory explanations forms an* ***inconsistency***.

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

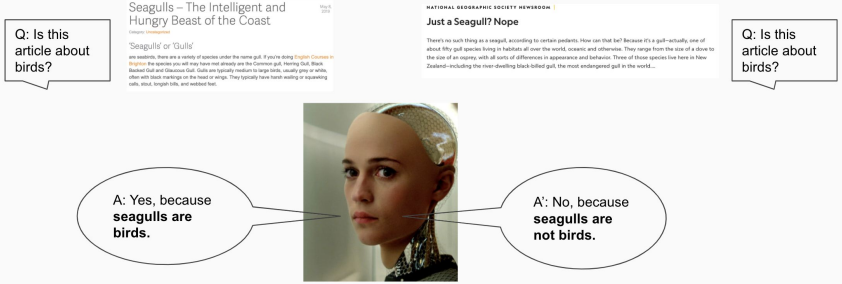
@ACL'20 O. Camburu, B. Shillingford, P. Minervini, T. Lukasiewicz, P. Blunsom.

Examples of inconsistencies

Self-Driving Cars



Question Answering



Visual Question Answering

Recommender Systems

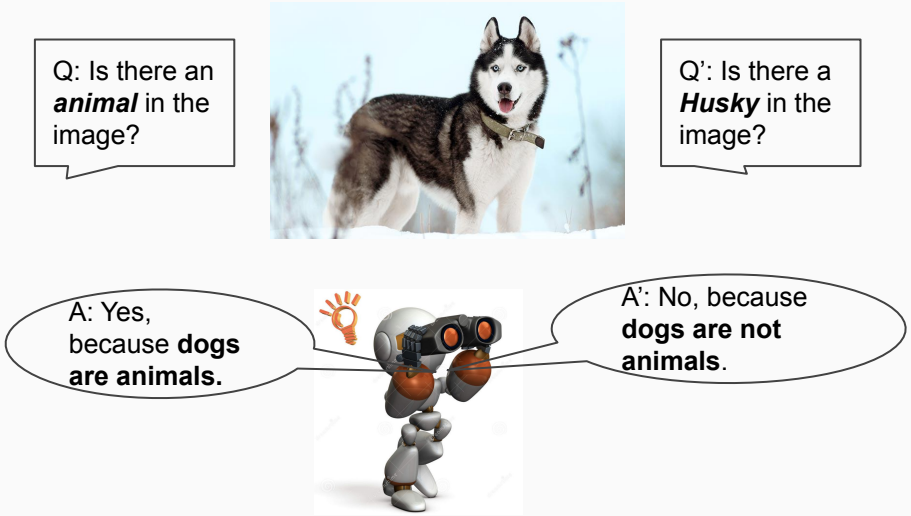
Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

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

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

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

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



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

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

Goal: Check models' robustness against generating inconsistent NLEs.

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

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

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

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

Goal: Check models' robustness against generating inconsistent NLEs.

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

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

High-level Approach

(A) For an instance x and the explanations $e_m(x)$, 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 $e_m(x') = i_e$.

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

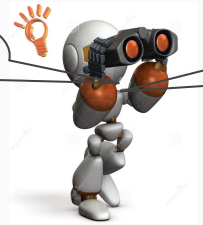
@ACL'20 O. Camburu, B. Shillingford, P. Minervini, T. Lukasiewicz, P. Blunsom.

Inconsistencies could be dependent on the context

Q: Is there an animal in the image?

Q': Is there a Husky in the image?

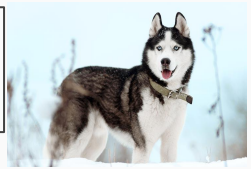
A: Yes, **there is a dog in the image.**



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

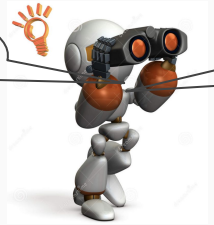
Inconsistent

Q: Is there an animal in the image?



Q': Is there a Husky in the image?

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



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

NOT Inconsistent

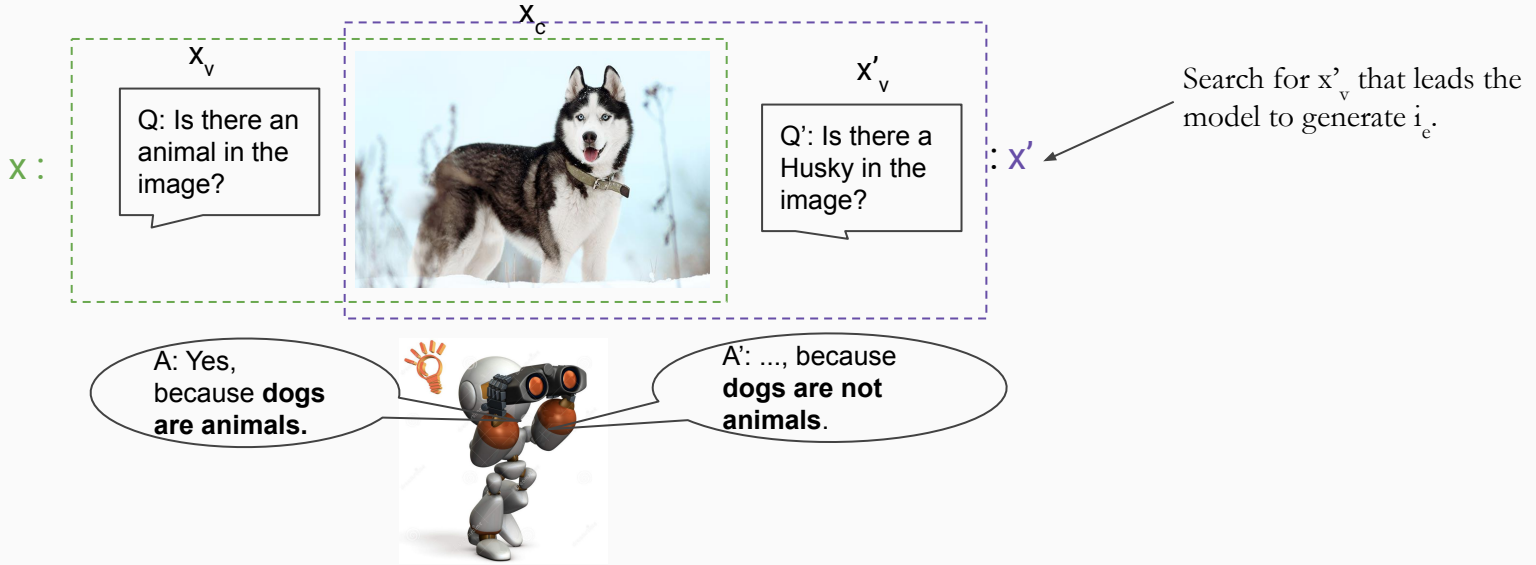
Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

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

Adversarial method

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

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



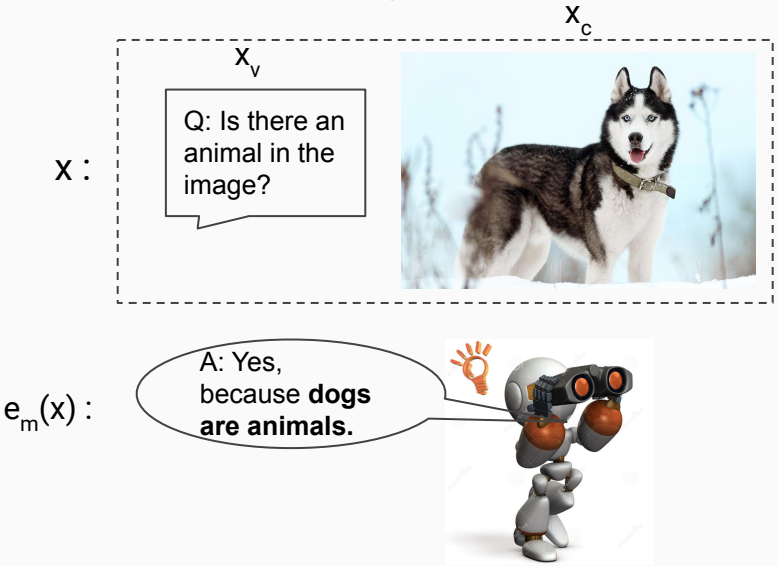
Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

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

Adversarial method

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

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



A set of logical rules:

- negation
- task-specific antonyms
- swap NLEs of mutually exclusive labels

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

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

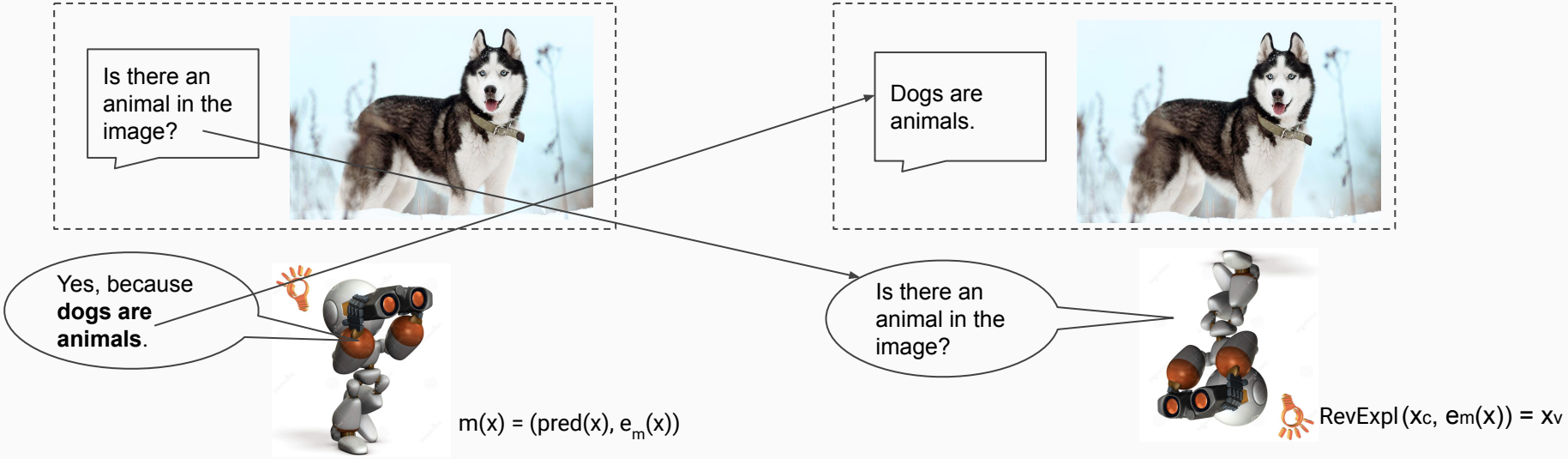
Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

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

Adversarial method

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

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



Adversarial method

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

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

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

- by using logic rules: negation, task-specific antonyms,

  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

Addressing a Novel Adversarial Setup

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

2) The model has to generate a **full target sequence**: the goal is to generate the **exact** statement that was identified as inconsistent with the original explanation. Previous 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 do not have to be a paraphrase or a small perturbation of the original input** (can happen as a byproduct). Previous works focus on adversaries being paraphrases or a minor deviation from the original input (Belinkov and Bisk, 2018).

Experiments: e-SNLI

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

- Best model from before: Expl-Pred-Att
  - 64.27% correct explanations

- RevExpl(premise, explanation) = hypothesis
  - same architecture as Expl-Pred-Att
  - 32.78% test accuracy (exact string match for the generated hypothesis)

- **Success rate** of our adversarial method for finding inconsistencies **4.51%** on the e-SNLI test set
  - **443** distinct pairs of inconsistent explanations

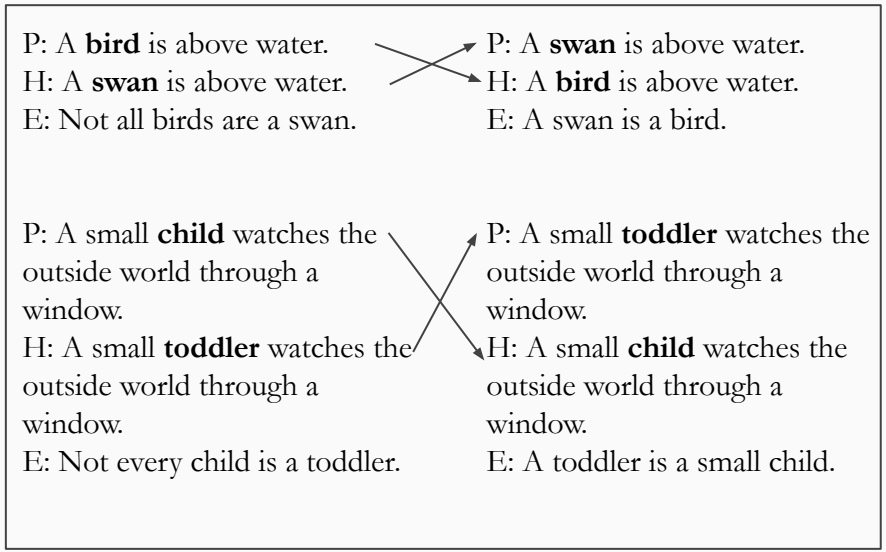
Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations

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

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.**

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

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



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

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

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

Inconsistencies are mostly due to **lack of common sense**.

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.**

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

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

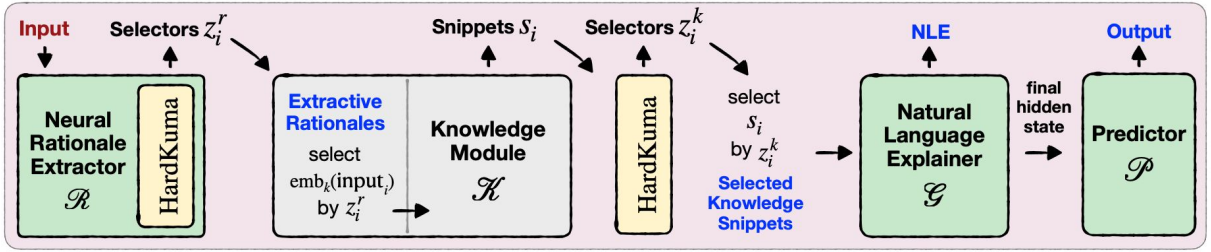
Goal: **knowledge grounding** for NLEs-generating models

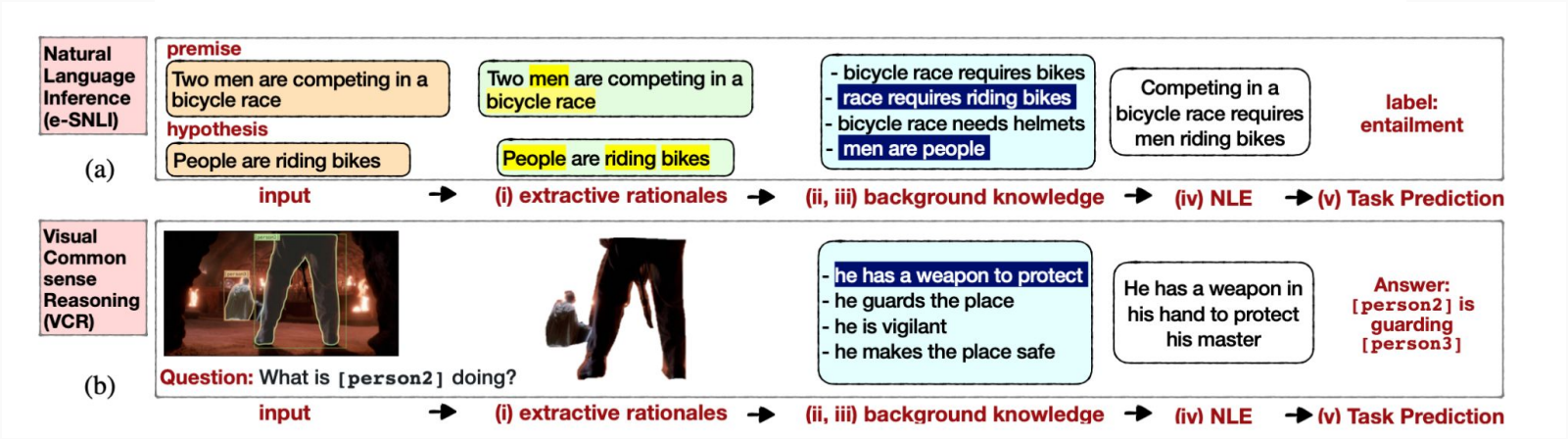
Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

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

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

RExC

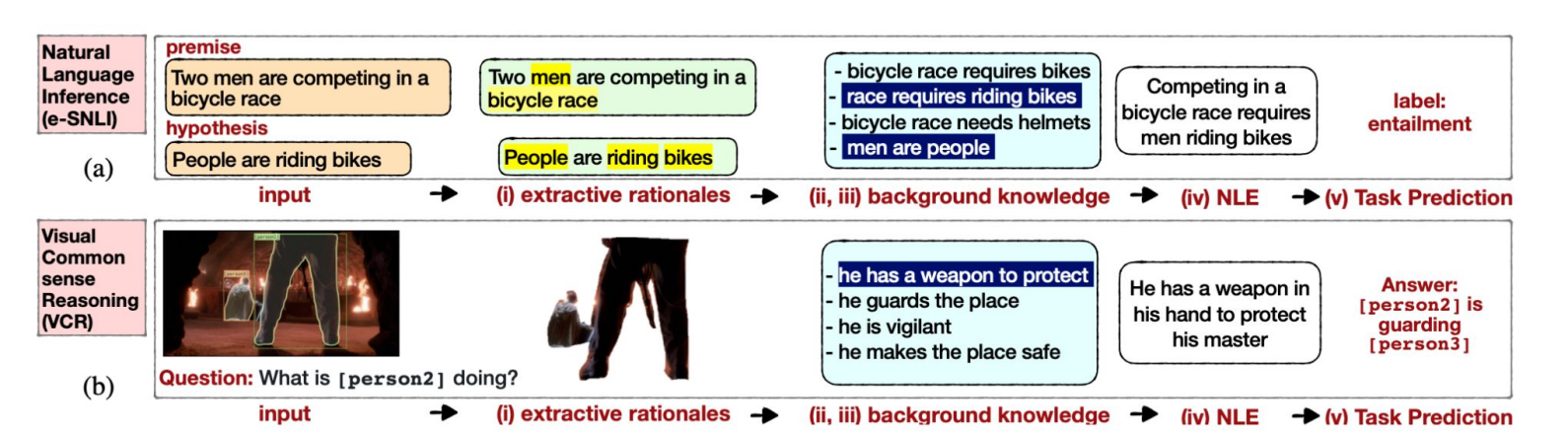




Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

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

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



Other advantages:

- **two complementary types of explanations**: extractive rationales and NLEs
- selected background knowledge can act as **additional** explanations (RExC+) or as **sufficient** explanations (RExC-ZS) in a **zero-shot** setup

Experiments

Natural Language Tasks

Natural Language Inference

premise: Two men are competing in a bicycle race
hypothesis: People are riding bikes

label: *entailment*

e-SNLI (Camburu et al., 2018)

Commonsense Validation

A: Coffee stimulates people
B: Coffee depresses people

label: *B is invalid*

ComVE (Wang et al., 2019)

Commonsense QA

Q: Where does a wild bird usually live?
A: a) cage, b) sky, c) countryside, d) desert, e) windowsill

label: *sky*

CoS-E (Rajani et al., 2019)

Vision Language Tasks

Visual Entailment

Hypothesis: Some tennis players pose

label: *entailment*

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

Visual Commonsense Reasoning

Q: What is the place?

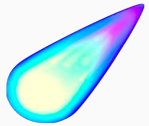
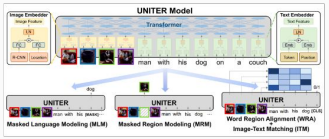
label: *They are in a hospital room*

VCR (Zellers et al., 2019)

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

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

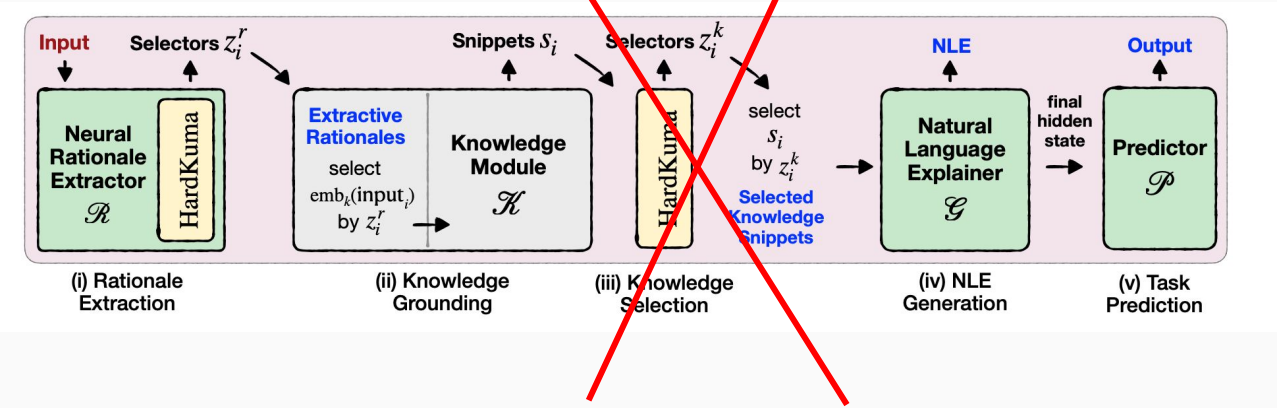
@ICML'22 B. Majumder, O. Camburu, T. Lukasiewicz, J. McAuley.

NLP	**BART**: a Seq2Seq pretrained transformer with a MLP prediction head (Lewis et al., 2020) 	**COMET**: Commonsense Transformer trained on ConceptNet (Bosselut et al., 2019) 	**BART**: a Seq2Seq pretrained transformer with a Language Model head (Lewis et al., 2020) 
VL	**UNITER**: a Seq2Seq pretrained transformer for text and images with a MLP prediction head (Chen et al., 2020) 	**Visual-COMET**: Commonsense Transformer trained on Visual Commonsense Graph (Park et al., 2020) 	**GPT2**: a pretrained transformer-based Language Model (Radford et al., 2020) 

Avoid no-hit issue of indexed KBs

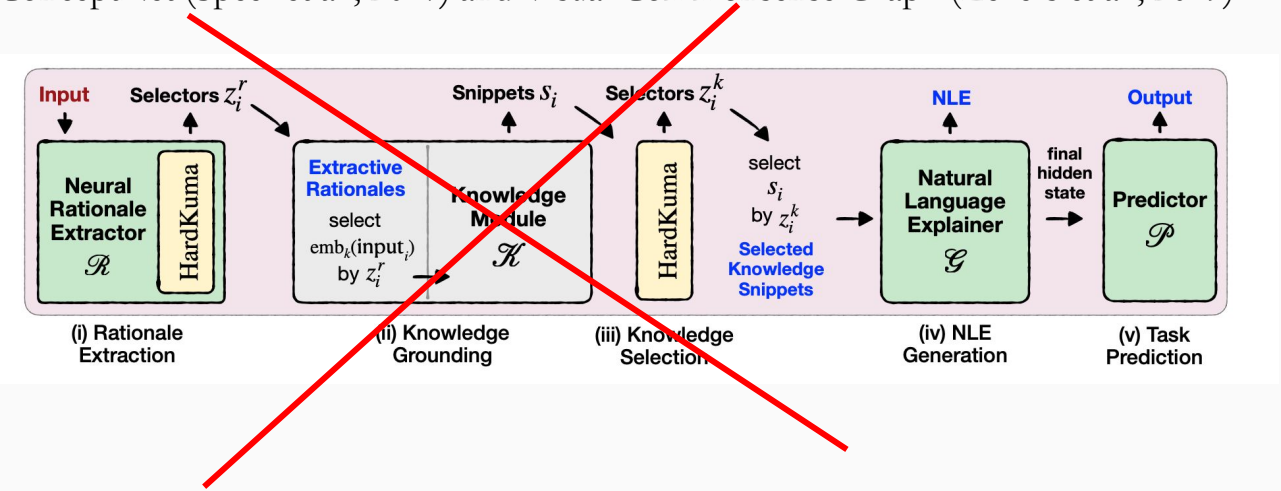
Ablations

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



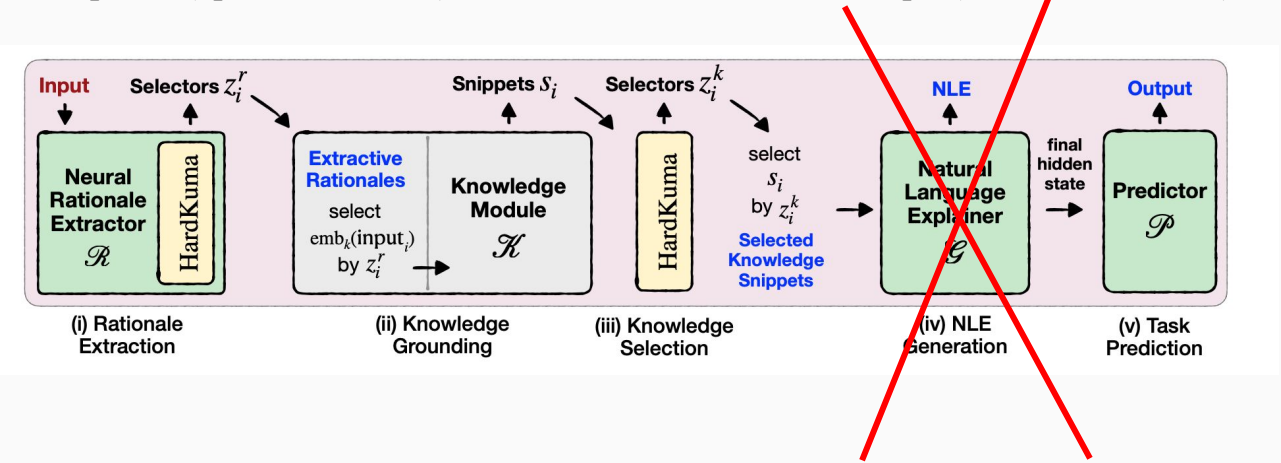
Ablations

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



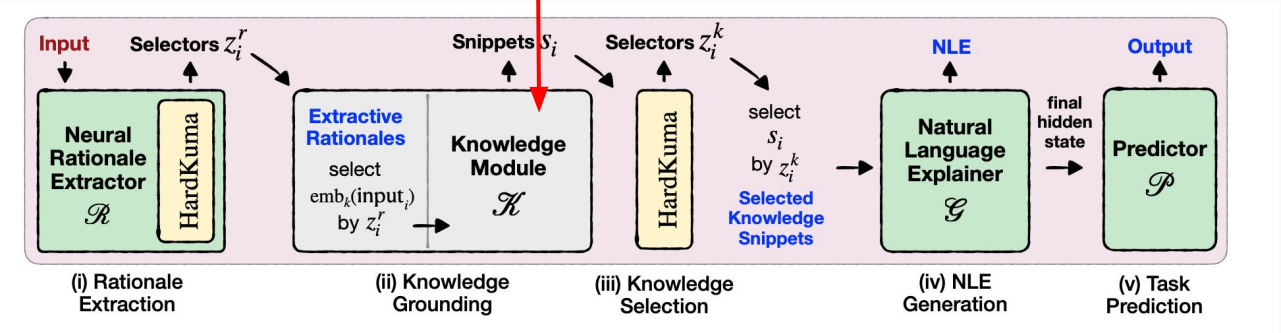
Ablations

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



Ablations

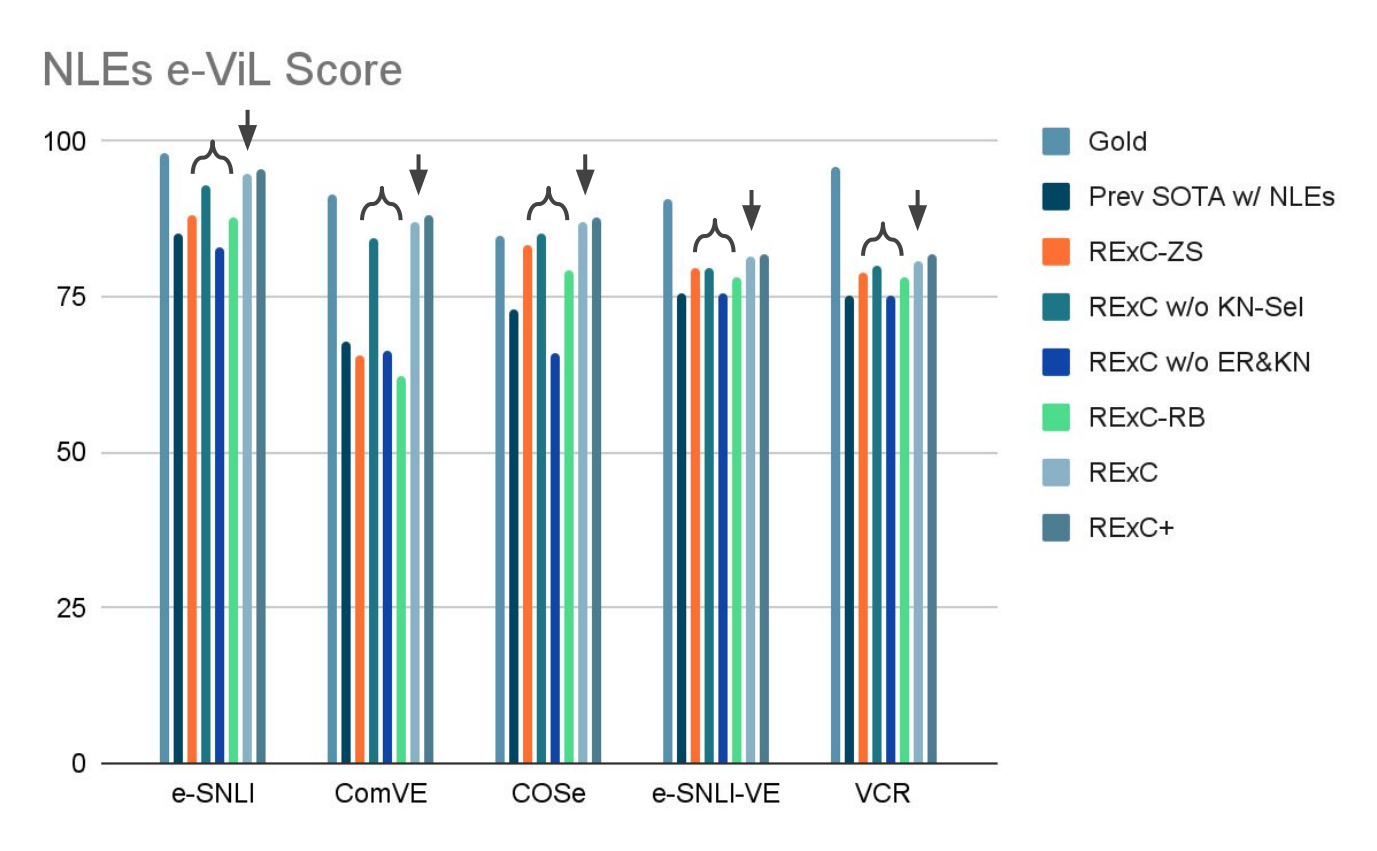
- knowledge selection (w/o KN-Sel)
- ER and knowledge selectors (w/o KN & ER)
- NLE generator (RExC-ZS) – supervision only from the output and selected knowledge snippets as NLEs
- generative knowledge module replaced with a retrieval-based knowledge source (RExC-RB)
  - ConceptNet (Speer et al., 2017)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



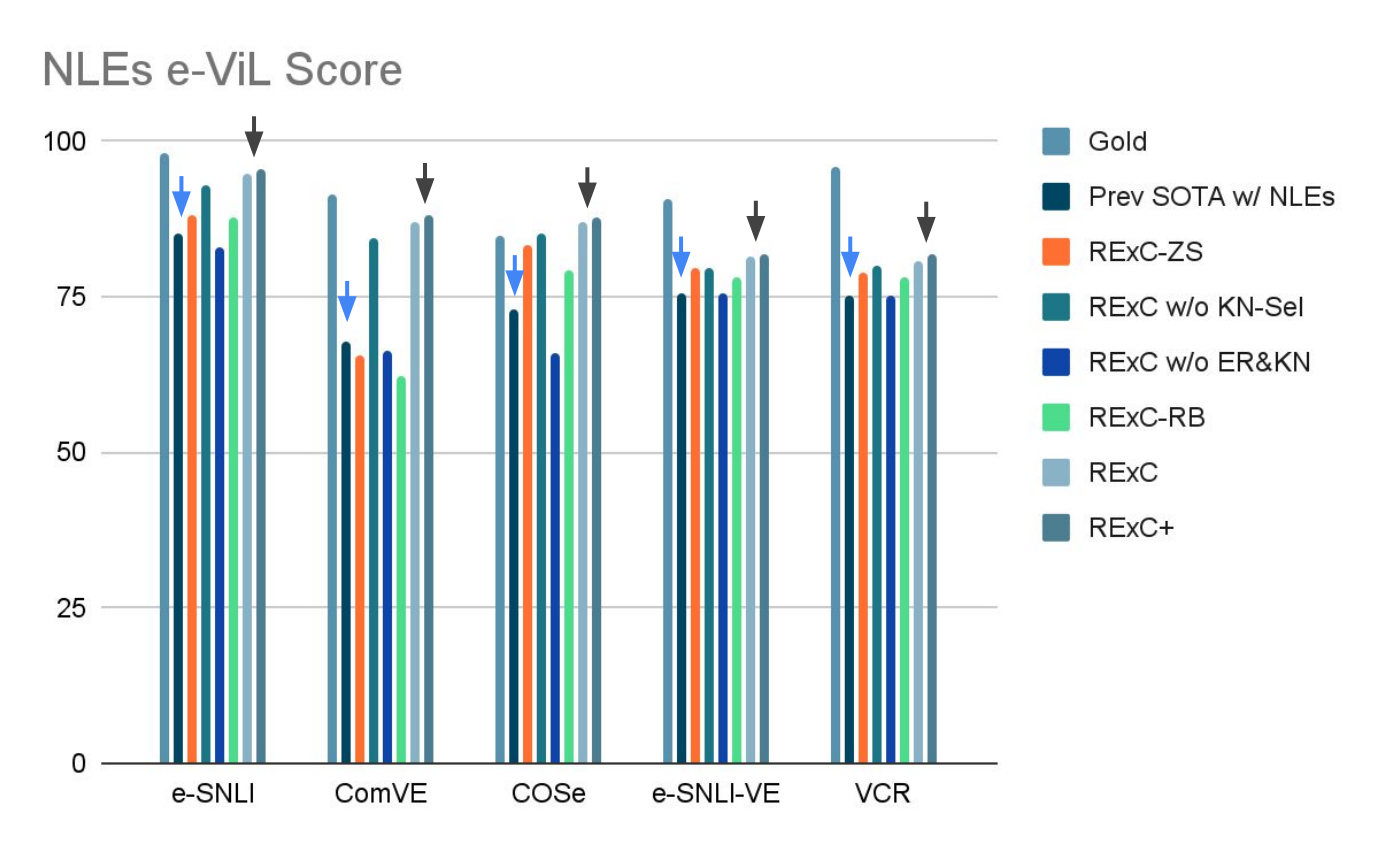
Results


NLEs e-ViL Score
100
75
50
25
0
e-SNLI
ComVE
COSe
e-SNLI-VE
VCR
Gold
Prev SOTA w/ NLEs
RExC-ZS
RExC w/o KN-Sel
RExC w/o ER&KN
RExC-RB
RExC
RExC+

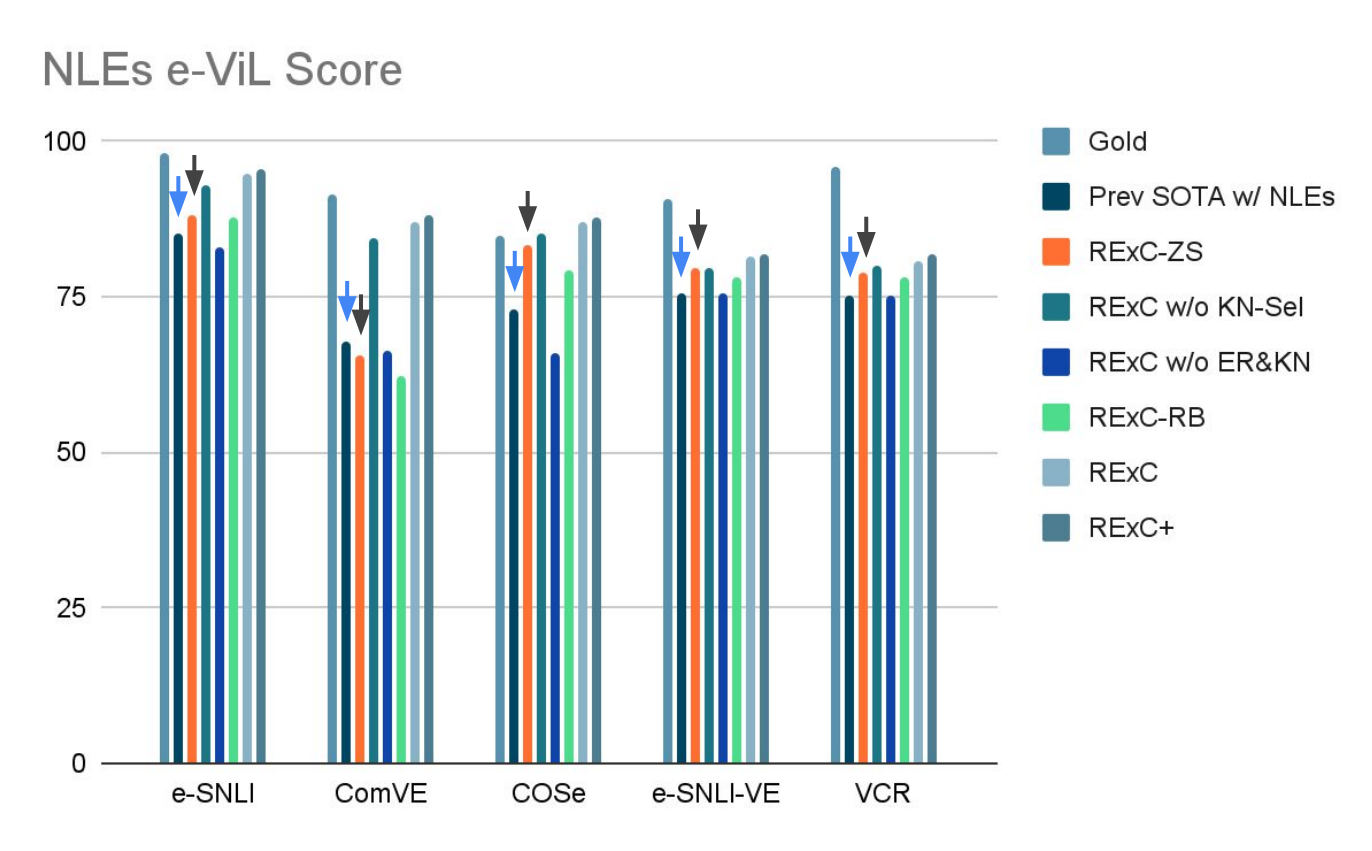
Results


NLEs e-ViL Score
100
75
50
25
0
e-SNLI
ComVE
COSe
e-SNLI-VE
VCR
Gold
Prev SOTA w/ NLEs
RExC-ZS
RExC w/o KN-Sel
RExC w/o ER&KN
RExC-RB
RExC
RExC+

Results


NLEs e-ViL Score
100
75
50
25
0
e-SNLI
ComVE
COSe
e-SNLI-VE
VCR
Gold
Prev SOTA w/ NLEs
RExC-ZS
RExC w/o KN-Sel
RExC w/o ER&KN
RExC-RB
RExC
RExC+

Results


NLEs e-ViL Score
100
75
50
25
0
e-SNLI
ComVE
COSe
e-SNLI-VE
VCR
Gold
Prev SOTA w/ NLEs
RExC-ZS
RExC w/o KN-Sel
RExC w/o ER&KN
RExC-RB
RExC
RExC+

Results

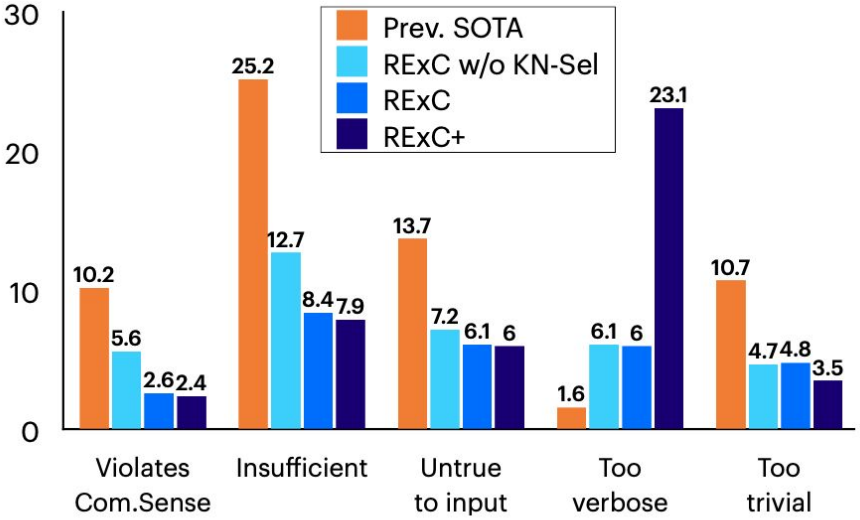
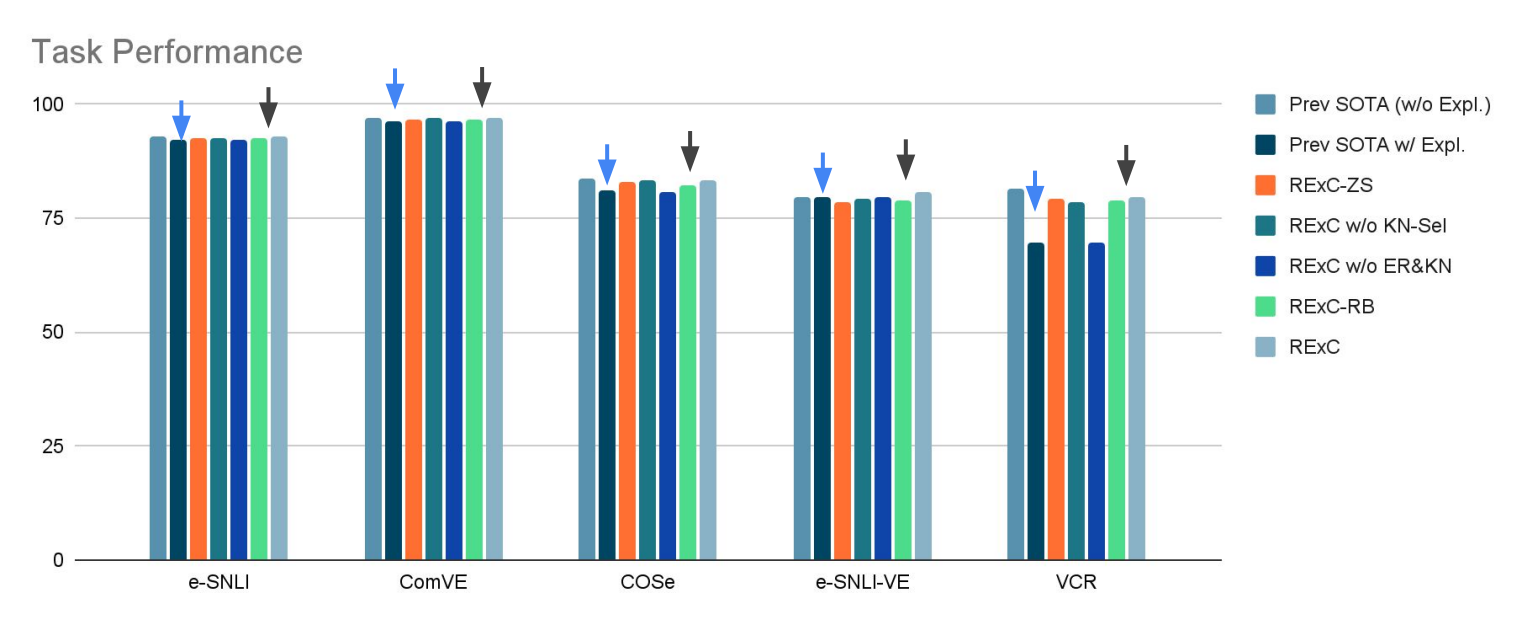


Figure 6. **Main limitations of the generated NLEs** obtained from user study. All numbers are in % and are averaged by systems and datasets for both NL and VL tasks. Human annotators could choose multiple limitations for an NLE.

Results


Task Performance
100
75
50
25
0
e-SNLI
ComVE
COSe
e-SNLI-VE
VCR
Prev SOTA (w/o Expl.)
Prev SOTA w/ Expl.
RExC-ZS
RExC w/o KN-Sel
RExC w/o ER&KN
RExC-RB
RExC

Results

RExC also outperforms the previous SOTA for extractive rationales

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

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

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

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

Results

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

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

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

First knowledge-grounded model with NLEs

Two complementary types of explanations

SOTA on NLEs quality over 5 tasks

SOTA on extractive rationales on 2 tasks (only ones with gold extractive rationales)

Promising zero-shot NLEs

Explaining Chest X-ray Pathologies in Natural Language

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

MIMIC-NLE: the first dataset of NLEs for a medical task

Explaining Chest X-ray Pathologies in Natural Language

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

MIMIC-NLE: the first dataset of NLEs for a medical task

Extract **diagnoses** and **NLEs for the diagnoses** from the radiology reports in MIMIC-CXR (Johnson et al., 2019) by applying keyword filters, the CheXbert labeler (Smit et al., 2020), and label hierarchies

A.E. Johnson et al., MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019.
A. Smit et al., Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT, EMNLP, 2020.

Explaining Chest X-ray Pathologies in Natural Language

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

MIMIC-NLE: the first dataset of NLEs for a medical task

Extract **diagnoses** and **NLEs for the diagnoses** from the radiology reports in MIMIC-CXR (Johnson et al., 2019) by applying keyword filters, the CheXbert labeler (Smit et al., 2020), and label hierarchies

- Divide the findings from CheXbert between evidence and diagnosis

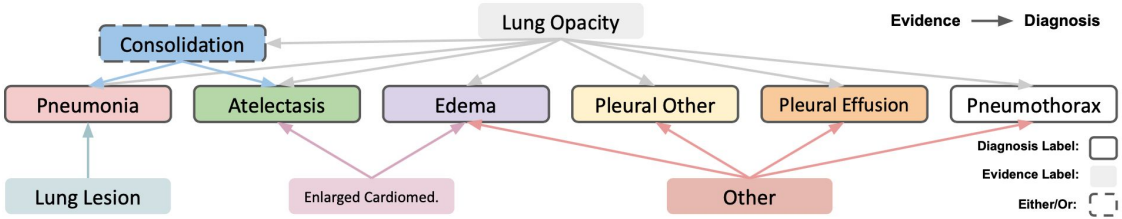


Fig. 2. Our evidence graph that visualizes which of the labels can act as evidence for which diagnosis labels.

A.E. Johnson et al., MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019.
A. Smit et al., Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT, EMNLP, 2020.

Explaining Chest X-ray Pathologies in Natural Language

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

MIMIC-NLE: the first dataset of NLEs for a medical task

Extract **diagnoses** and **NLEs for the diagnoses** from the radiology reports in MIMIC-CXR (Johnson et al., 2019) by applying keyword filters, the CheXbert labeler (Smit et al., 2020), and label hierarchies

- Divide the findings from CheXbert between evidence and diagnosis
- Identify a set of rules that mark a sentence from the radiology report as a valid NLE

Table 1. This table denotes all the included label combinations for NLEs, including which of the labels are being explained and which are the evidence. The column "*kw req.*" specifies which label combinations additionally require the presence of an explanation keyword to be considered an NLE. "*Other / misc.*" refers to evidence that has not been picked up by the CheXbert labeler. If not denoted by ^{U} or ^{P}, all labels can be either positive or uncertain. A^U and B^U are the sets A and B, where all labels are given as uncertain. $\mathcal{P}_{\geq 2}(A^U)$ is the power set of A^U, where each set has at least two labels (i.e., any combination of at least two labels from A^U).

MIMIC-NLE Label Combinations		
Evidence	**Diagnosis Label(s)**	**kw req.**
Other / misc.	$d \in A = \{\text{Pleural Eff., Edema, Pleural Other, Pneumoth.}\}$	yes
Other / misc.	$s \in \mathcal{P}_{\geq 2}(A^U)$	yes
Lung Opacity	$d \in B = A \cup \{\text{Pneumonia, Atelectasis}\}$	no
Lung Opacity	$s \in \mathcal{P}_{\geq 2}(B^U)$	no
Lung Opacity	Consolidation	no
Consolidation	Pneumonia	no
{Lung Op., Cons.}	Pneumonia	no
Lung Lesion	Pneumonia	yes
Lung Opacity	$\{\text{Atelectasis}^P, \text{Pneumonia}^U\}$	no
Consolidation	$\{\text{Atelectasis}^U, \text{Pneumonia}^U\}$	no
Enlarged Card.	Edema	yes
Enlarged Card.	Atelectasis	yes

A.E. Johnson et al., MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019.
A. Smit et al., Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT, EMNLP, 2020.

Explaining Chest X-ray Pathologies in Natural Language

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

MIMIC-NLE: the first dataset of NLEs for a medical task

Extract **diagnoses** and **NLEs for the diagnoses** from the radiology reports in MIMIC-CXR (Johnson et al., 2019) by applying keyword filters, the CheXbert labeler (Smit et al., 2020), and label hierarchies

- Divide the findings from CheXbert between evidence and diagnosis
- Identify a set of rules that mark a sentence from the radiology report as a valid NLE
- **44,935 image-diagnosis-NLE** triplets (38,003 NLEs: some NLEs explain multiple diagnoses)

A.E. Johnson et al., MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019.
A. Smit et al., Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT, EMNLP, 2020.

Explaining Chest X-ray Pathologies in Natural Language

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

Models

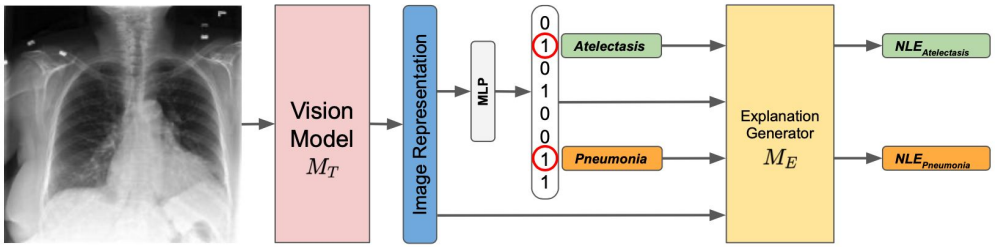


Fig. 3. The model pipeline to provide an NLE for a prediction.

Three baselines

- adapt two SOTA for chest X-ray captioning: **TieNet** (Wang et al., 2018) and **RATCHET** (Hou et al., 2021) with DenseNet-12 from TorchXRayVision (Cohen et al., 2020) as the vision model
- new baseline inspired by our e-UG: **DPT** = DenseNet-121 + GPT-2 (Radford et al, 2019)

New automatic metric

- CLEV (CLinical EVidence) score = accuracy with which an NLE refers to all the clinical evidence from the GT NLE

X. Wang et al., Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, 2018.
B. Hou et al., RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting, MICCAI 2021.
J.P. Cohen et al., TorchXRayVision: A library of chest X-ray datasets and models, 2020.
A. Radford, Language Models are Unsupervised Multitask Learners, 2019.

Explaining Chest X-ray Pathologies in Natural Language

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

Results

Table 2. The S_T score, clinical evaluation, and NLG scores for our baselines on the MIMIC-NLE test set. $\geq$GT reflects the share of generated NLEs that received a rating on-par or better than the GT. Clin.Sc. reflects the average rating of 1 (lowest) to 5 (highest) that was given to the NLEs by a clinician. R-L refers to Rouge-L, and Bn to the n-gram BLEU scores. Best results are in bold. As we only evaluate NLEs for correctly predicted diagnoses, our NLG metrics cover 534, 560, and 490 explanations for RATCHET, TieNet, and DPT, respectively.

	AUC	$\geq$GT	Clin.Sc.	CLEV	BERTS.	MET.	B1	B4	R-L	CIDEr	SPICE
GT	-	-	3.20	-	-	-	-	-	-	-	-
DenseNet-121	65.2	-	-	-	-	-	-	-	-	-	-
RATCHET	**66.4**	**48%**	**2.90**	74.7	77.6	**14.1**	**22.5**	**4.7**	**22.2**	**37.9**	**20.0**
TieNet	64.6	40%	2.60	**78.0**	**78.0**	12.4	17.3	3.5	19.4	33.9	17.2
DPT	62.5	**48%**	2.66	74.9	77.3	11.3	17.5	2.4	15.4	17.4	13.7

Explaining Chest X-ray Pathologies in Natural Language

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

Results

Table 2. The S_T score, clinical evaluation, and NLG scores for our baselines on the MIMIC-NLE test set. $\geq$GT reflects the share of generated NLEs that received a rating on-par or better than the GT. Clin.Sc. reflects the average rating of 1 (lowest) to 5 (highest) that was given to the NLEs by a clinician. R-L refers to Rouge-L, and Bn to the n-gram BLEU scores. Best results are in bold. As we only evaluate NLEs for correctly predicted diagnoses, our NLG metrics cover 534, 560, and 490 explanations for RATCHET, TieNet, and DPT, respectively.

	AUC	$\geq$GT	Clin.Sc.	CLEV	BERTS.	MET.	B1	B4	R-L	CIDEr	SPICE
GT	-	-	3.20	-	-	-	-	-	-	-	-
DenseNet-121	65.2	-	-	-	-	-	-	-	-	-	-
RATCHET	**66.4**	**48%**	**2.90**	74.7	77.6	**14.1**	**22.5**	**4.7**	**22.2**	**37.9**	**20.0**
TieNet	64.6	40%	2.60	**78.0**	**78.0**	12.4	17.3	3.5	19.4	33.9	17.2
DPT	62.5	**48%**	2.66	74.9	77.3	11.3	17.5	2.4	15.4	17.4	13.7

Explaining Chest X-ray Pathologies in Natural Language

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

Results

Table 2. The S_T score, clinical evaluation, and NLG scores for our baselines on the MIMIC-NLE test set. $\geq$GT reflects the share of generated NLEs that received a rating on-par or better than the GT. Clin.Sc. reflects the average rating of 1 (lowest) to 5 (highest) that was given to the NLEs by a clinician. R-L refers to Rouge-L, and Bn to the n-gram BLEU scores. Best results are in bold. As we only evaluate NLEs for correctly predicted diagnoses, our NLG metrics cover 534, 560, and 490 explanations for RATCHET, TieNet, and DPT, respectively.

	AUC	$\geq$GT	Clin.Sc.	CLEV	BERTS.	MET.	B1	B4	R-L	CIDEr	SPICE
GT	-	-	3.20	-	-	-	-	-	-	-	-
DenseNet-121	65.2	-	-	-	-	-	-	-	-	-	-
RATCHET	**66.4**	**48%**	**2.90**	74.7	77.6	**14.1**	**22.5**	**4.7**	**22.2**	**37.9**	**20.0**
TieNet	64.6	40%	2.60	**78.0**	**78.0**	12.4	17.3	3.5	19.4	33.9	17.2
DPT	62.5	**48%**	2.66	74.9	77.3	11.3	17.5	2.4	15.4	17.4	13.7

Explaining Chest X-ray Pathologies in Natural Language

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

Results

Table 2. The S_T score, clinical evaluation, and NLG scores for our baselines on the MIMIC-NLE test set. $\geq$GT reflects the share of generated NLEs that received a rating on-par or better than the GT. Clin.Sc. reflects the average rating of 1 (lowest) to 5 (highest) that was given to the NLEs by a clinician. R-L refers to Rouge-L, and Bn to the n-gram BLEU scores. Best results are in bold. As we only evaluate NLEs for correctly predicted diagnoses, our NLG metrics cover 534, 560, and 490 explanations for RATCHET, TieNet, and DPT, respectively.

	AUC	$\geq$GT	Clin.Sc.	CLEV	BERTS.	MET.	B1	B4	R-L	CIDEr	SPICE
GT	-	-	3.20	-	-	-	-	-	-	-	-
DenseNet-121	65.2	-	-	-	-	-	-	-	-	-	-
RATCHET	**66.4**	**48%**	**2.90**	74.7	77.6	**14.1**	**22.5**	**4.7**	**22.2**	**37.9**	**20.0**
TieNet	64.6	40%	2.60	**78.0**	**78.0**	12.4	17.3	3.5	19.4	33.9	17.2
DPT	62.5	**48%**	2.66	74.9	77.3	11.3	17.5	2.4	15.4	17.4	13.7

Explaining Chest X-ray Pathologies in Natural Language

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

Results

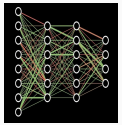
LABELS: Atelectasis (Positive)

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

Future Directions

Datasets of NLEs

- more challenging, e.g., multi-hop
- dialog
- domain specific



Models and benchmarks for high-quality NLEs

- grounding
- automatic metrics
- human in the loop

Faithfulness

- evaluation
- architectures

Improve task performance

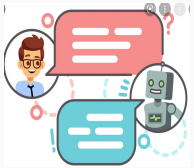
- regularizers
- active learning

Zero/Few-shot

- prompting
- transfer

Usefulness for users

- user-studies
- complementary explanations
- dialog



Dialog XAI

- prompting
- dialog architectures

Personalized XAI

- few-shot

. . .

Thank you!

Questions

@oanacamb