REASONING OR SPURIOUS CORRELATIONS? APPLYING TRANSFORMERS TO PROPOSITIONAL LOGIC

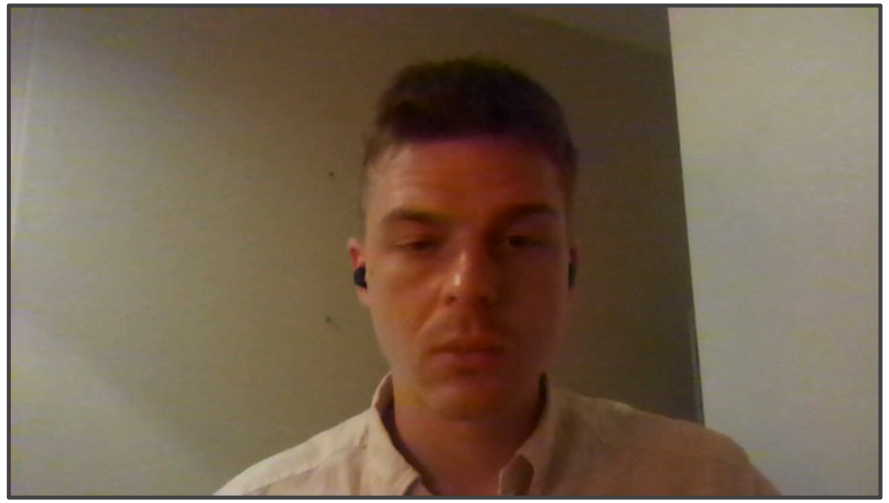
DANIEL ENSTRÖM, VIKTOR KJELLBERG, MOA JOHANSSON

UNIVERSITY OF GOTHENBURG / CHALMERS UNIVERSITY OF TECHNOLOGY

REASONING OR SPURIOUS CORRELATIONS?
APPLYING TRANSFORMERS TO PROPOSITIONAL LOGIC

Presentation overview

- The dataset - SimpleLogic extended with proofs
- Brief overview of the two architectures
- Results in terms of model accuracy and consistency
- Qualitative analysis of the errors of the best performing architecture



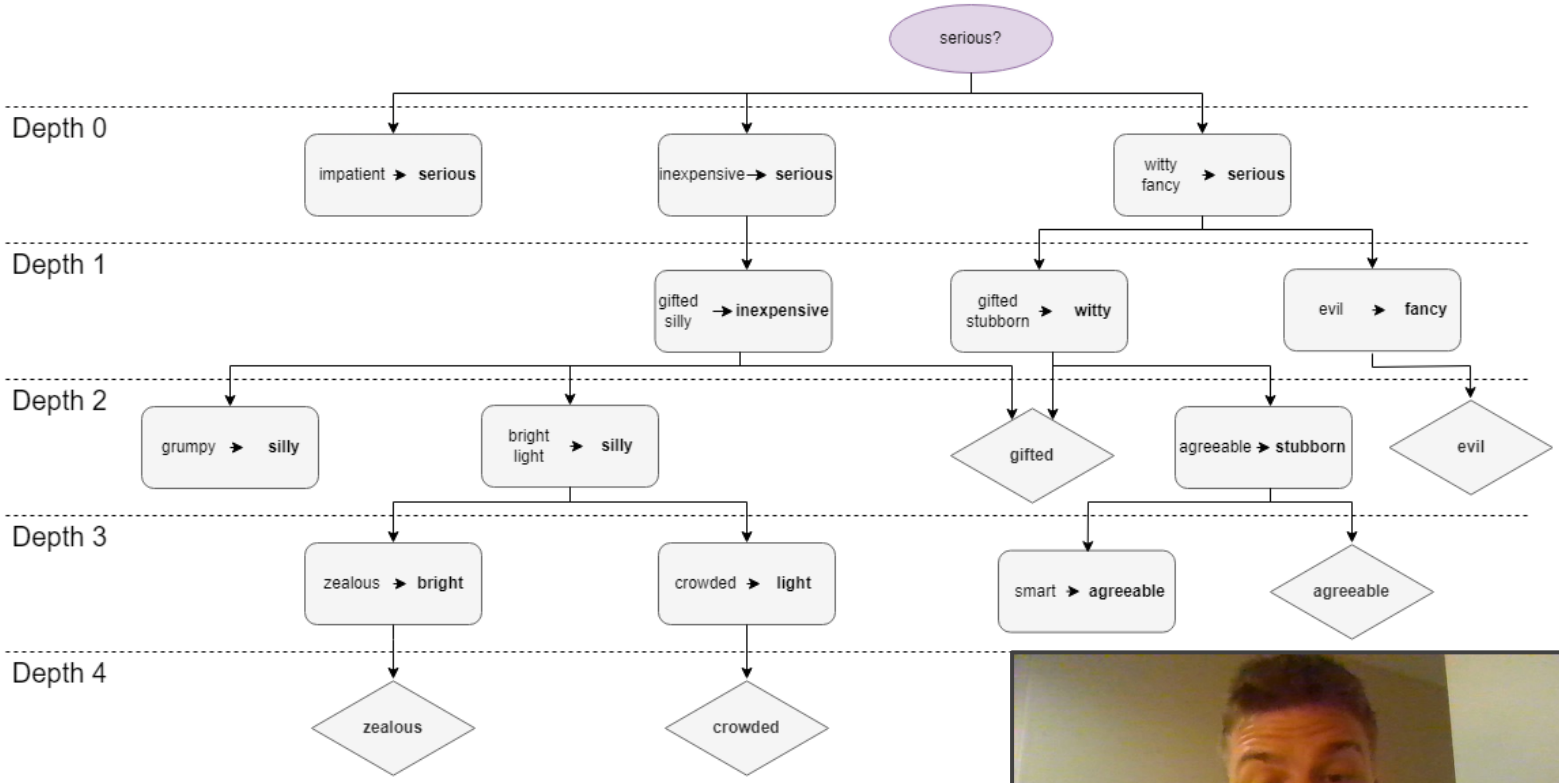
Daniel Enström

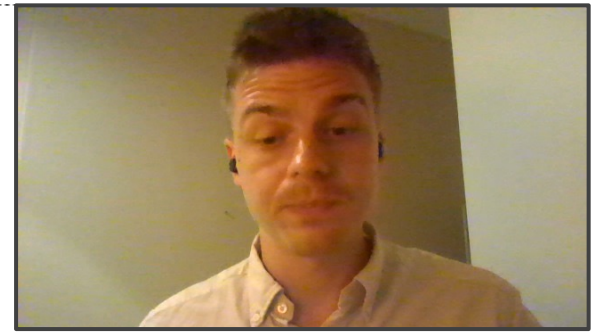
SIMPLELOGIC – A DATASET OF SATISIFIABILITY PROBLEMS

QUERY: serious?

RULES:
impatient ⟶ serious
gifted ∧ silly ⟶ inexpensive
bright ∧ light ⟶ silly
crowded ⟶ light
.....
evil ⟶ fancy

FACTS:
zealous
crowded
gifted
agreeable
evil

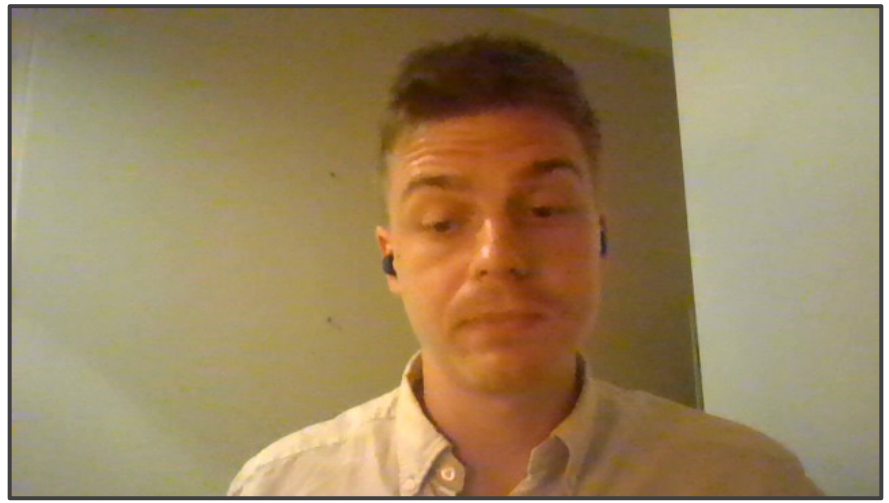




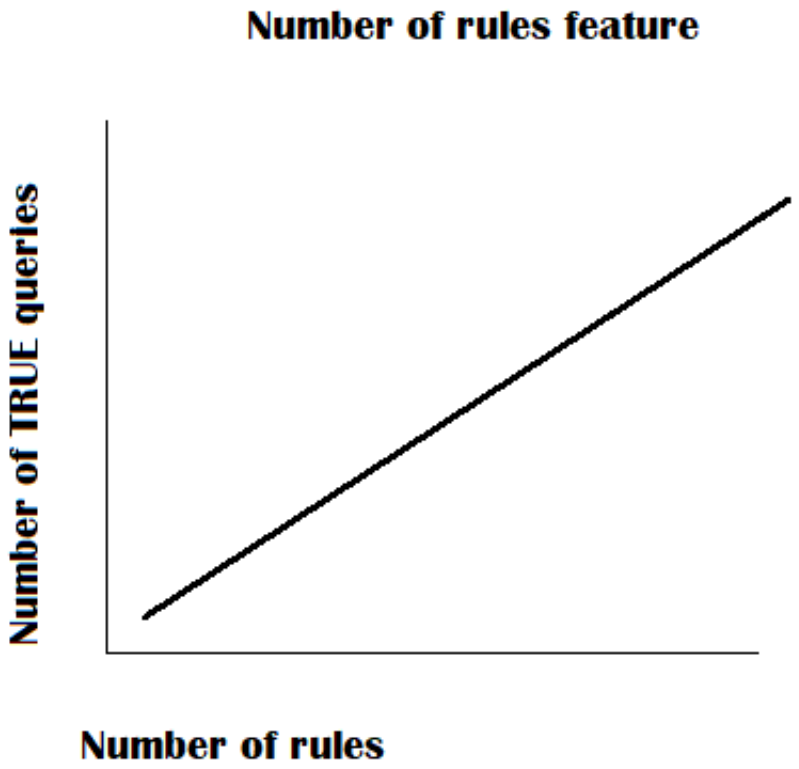
Daniel Enström

SimpleLogic

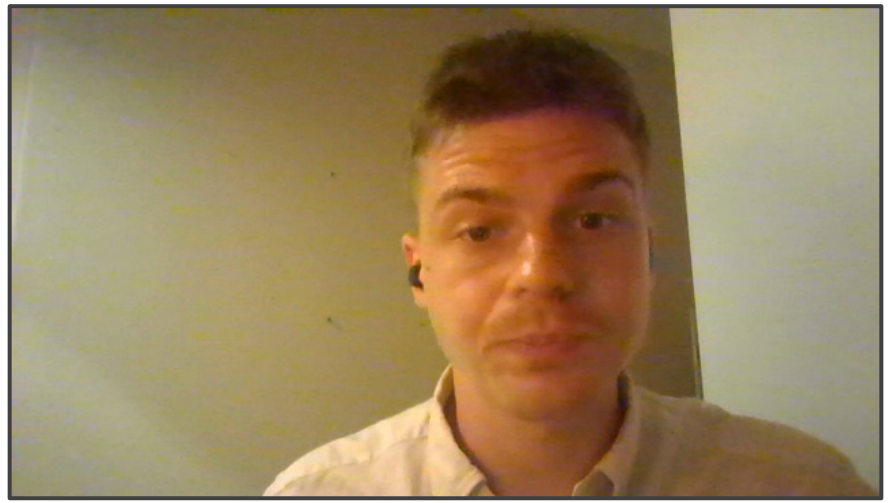
- The three datasets are generated in different ways
  - Label-Priority (LP)
    - Truth-value of literals sampled first
  - Rule-Priority (RP)
    - Rules sampled first
  - Rule-Priority Balanced (RP_b)
    - Like RP, but without the ”number of rules” statistical feature



SimpleLogic

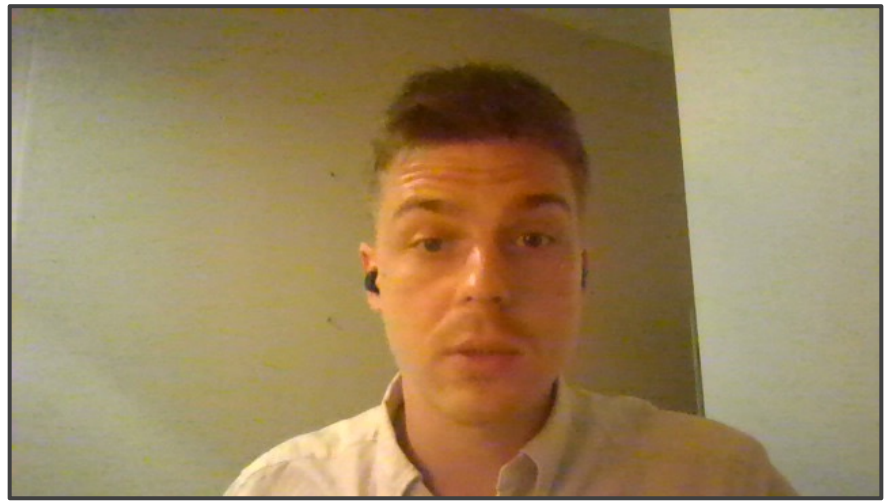


- An example of a statistical feature is the ”number of rules” feature
- This feature is present in RP but removed in RP_b



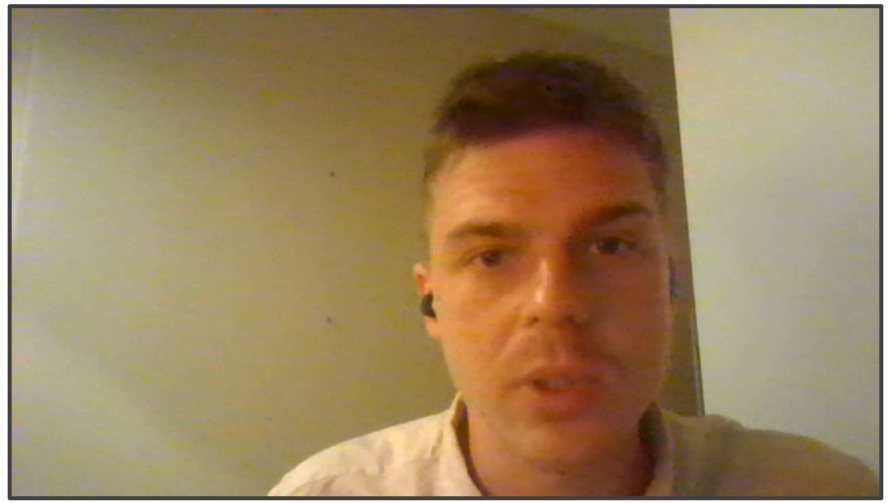
Replication of Zhang et. al (2022)

TRAIN	TEST	0	1	2	3	4	5	6	TOTAL
RP	RP	99.8	100.0	99.4	98.9	98.6	96.9	95.9	98.5
RP	RP_b	99.2	99.2	98.6	98.0	96.6	93.9	89.1	96.4
RP	LP	99.9	99.9	99.0	94.3	83.8	65.6	50.0	84.7
RP_b	RP	99.8	99.9	99.5	98.9	98.6	97.9	96.9	98.8
RP_b	RP_b	99.6	99.5	99.0	98.4	98.0	96.7	94.1	97.9
RP_b	LP	99.7	99.4	99.3	96.4	87.6	72.6	57.2	87.5



EVALUATION

- Accuracy
- Consistency
  - I.e. combination of soundness and completeness

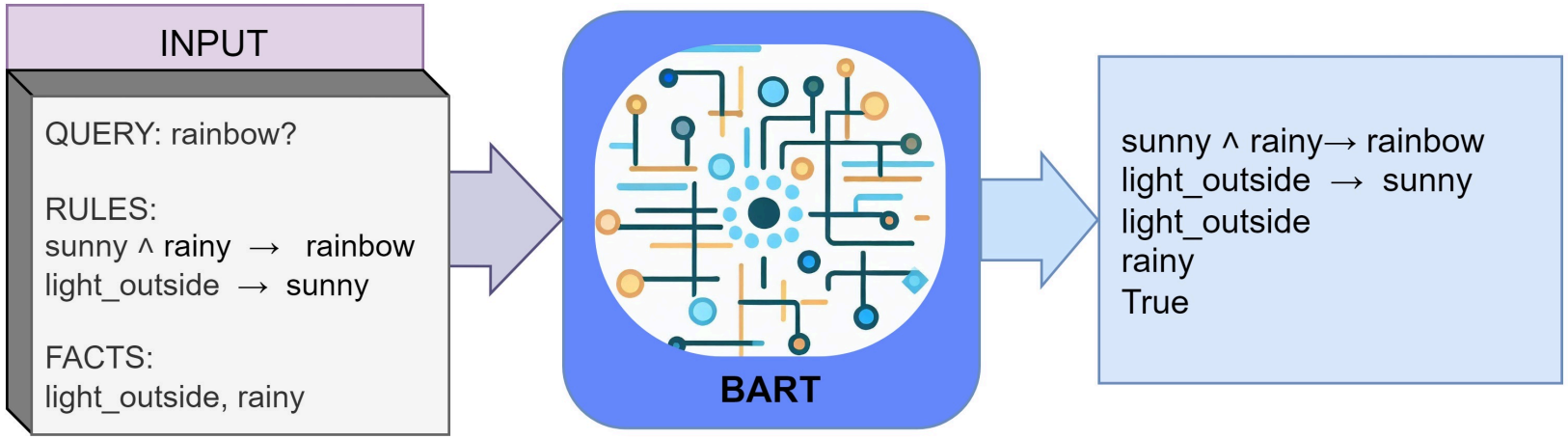


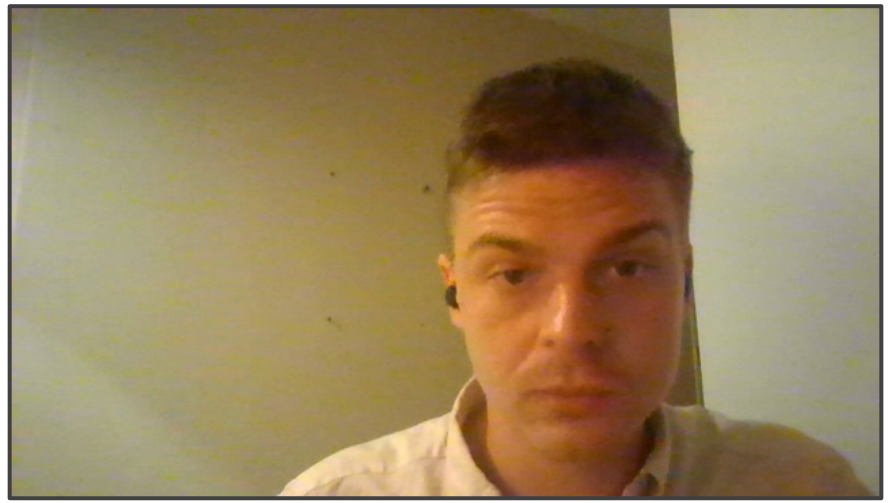
MODELS USED

- Whole Proof BART (WP-BART)
  - Generates the proof as ONE SINGLE output-sequence from BART
- Symbolic Iterative Proof BART (SIP-BART)
  - Generates the proof step by step
    - The final proof is an aggregation of all generated steps

WP-BART

- A generative BART architecture
- Three models were trained
  - One for each of the respective datasets
- Training proofs generated with backward chaining
  - i.e. starting from the query
- Evaluated primarily on accuracy





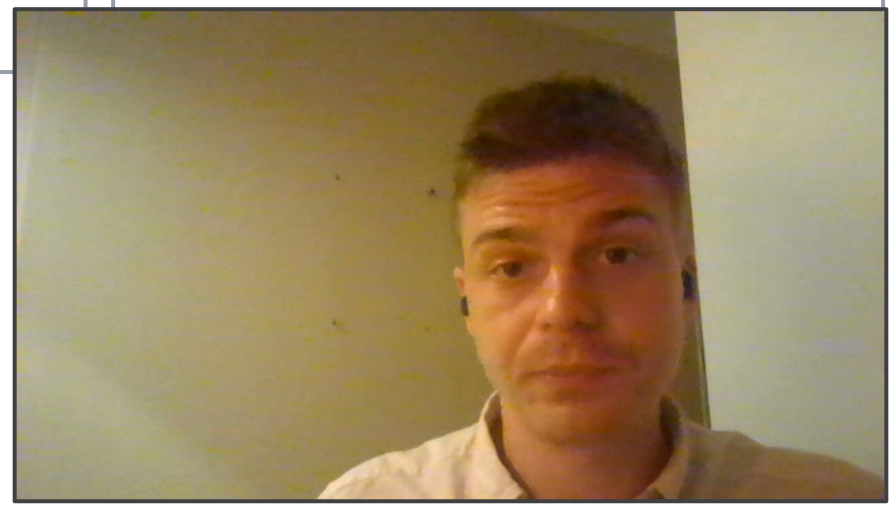
WP-BART RESULTS

WP-BART

TRAIN	TEST	TOTAL
LP	RP	80.6
LP	RP_b	81.4
LP	LP	93.9
RP	RP	84.5
RP	RP_b	85.5
RP	LP	75.2
RP_b	RP	88.7
RP_b	RP_b	89.9
RP_b	LP	83.2

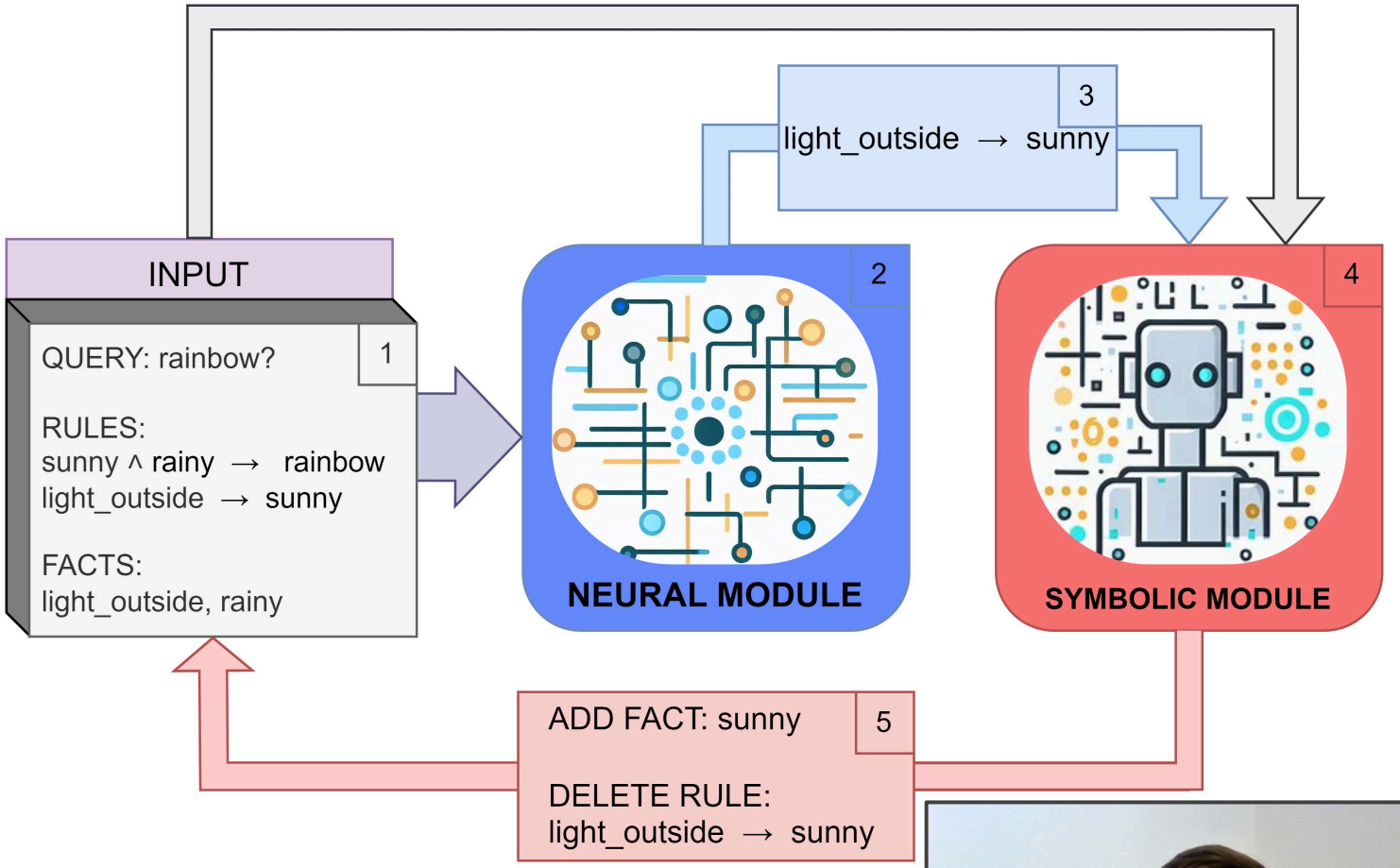
CLASSIFIER BENCHMARK

TRAIN	TEST	TOTAL
LP	RP	75.0
LP	RP_b	72.7
LP	LP	98.6
RP	RP	98.5
RP	RP_b	96.4
RP	LP	84.7
RP_b	RP	98.8
RP_b	RP_b	97.9
RP_b	LP	87.5



SIP-BART

- Symbolic Iterative Proof BART (SIP-BART)
- A combination of a neural module with a symbolic (rule-based) module
- The neural module is responsible to find the next applicable rule
- The symbolic module is just a short program that processes the output from the neural module and updates the input accordingly



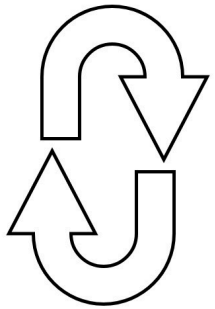
Viktor Kjellberg

INPUT

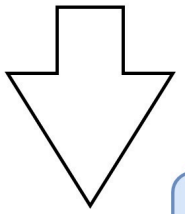
QUERY:
rainbow?

RULES:
cloudy ∧ droplets → rainy
light outside → sunny
sunny ∧ rainy → rainbow
dark outside → night
blue sky → sunny
rainy → wet grass

FACTS:
droplets
cloudy
light outside



GENERATED STEP



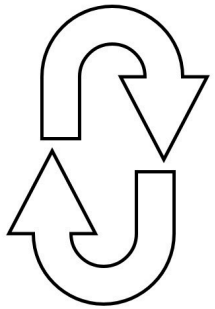
PROOF

INPUT

QUERY:
rainbow?

RULES:
cloudy ∧ droplets → rainy
light outside → sunny
sunny ∧ rainy → rainbow
dark outside → night
blue sky → sunny
rainy → wet grass

FACTS:
droplets
cloudy
light outside



GENERATED STEP

light outside → sunny

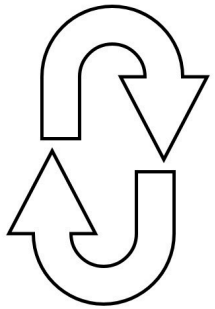
PROOF

INPUT

QUERY:
rainbow?

RULES:
cloudy ∧ droplets → rainy
light outside → sunny
sunny ∧ rainy → rainbow
dark outside → night
blue sky → sunny
rainy → wet grass

FACTS:
droplets
cloudy
light outside



GENERATED STEP

light outside → sunny

PROOF

1. light outside → sunny

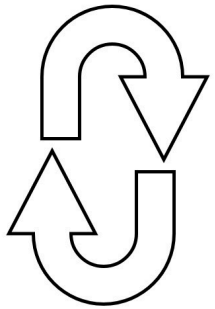
INPUT

QUERY:
rainbow?

RULES:
cloudy ∧ droplets → rainy

sunny ∧ rainy → rainbow
dark outside → night
blue sky → sunny
rainy → wet grass

FACTS:
droplets
cloudy
light outside



GENERATED STEP

light outside → sunny

PROOF

1. light outside → sunny

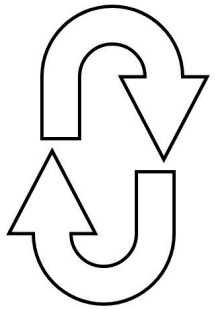
INPUT

QUERY:
rainbow?

RULES:
cloudy ∧ droplets → rainy

sunny ∧ rainy → rainbow
dark outside → night
blue sky → sunny
rainy → wet grass

FACTS:
droplets
cloudy
light outside
sunny



GENERATED STEP

light outside → sunny

PROOF

1. light outside → sunny

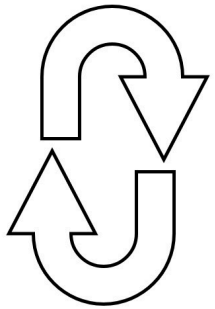
INPUT

QUERY:
rainbow?

RULES:

dark outside → night
blue sky → sunny

FACTS:
droplets
cloudy
light outside
sunny
rainy
wet grass
rainbow



GENERATED STEP

TRUE

PROOF

1. light outside → sunny
2. cloudy ∧ droplets → rainy
3. rainy → wet grass
4. sunny ∧ rainy → rainbow
5. TRUE

SIP-BART

- Evaluated by accuracy,
  - On the predicted truth-value only (not the entire proof)
  - Each dataset and all problem depths
- As well as consistency of proofs

Viktor Kjellberg

TRAIN	TEST	0	1	2	3	4	5	6	TOTAL
LP	LP	99.94	100.	99.97	100.	100.	100.	100.	**99.98**
LP	RP	99.97	99.97	100.	99.92	99.90	99.872	99.49	**99.87**
LP	RP_b	100.	99.97	99.97	99.89	99.94	99.74	99.25	**99.81**
RP	LP	99.97	100.	100.	100.	100.	100.	100.	**99.99**
RP	RP	100.	100.	100.	100.	100.	100.	99.97	**99.99**
RP	RP_b	100.	100.	100.	100.	100.	100.	100.	**100.**
RP_b	LP	99.97	100.	99.97	100.	100.	100.	100.	**99.99**
RP_b	RP	99.94	100.	100.	100.	99.97	99.97	99.97	**99.98**
RP_b	RP_b	100.	100.	100.	100.	100.	99.97	100.	**99.99**

TRAIN	TEST	0	1	2	3	4	5	6	TOTAL
LP	RP	97.4	92.5	64.5	60.2	67.6	72.6	69.9	**75.0**
LP	RP_b	97.7	93.3	60.2	56.7	63.9	68.7	68.5	**72.7**
LP	LP	99.8	99.8	99.8	99.6	98.8	97.2	95.4	**98.6**
RP	RP	99.8	100.0	99.4	98.9	98.6	96.9	95.9	**98.5**
RP	RP_b	99.2	99.2	98.6	98.0	96.6	93.9	89.1	**96.4**
RP	LP	99.9	99.9	99.0	94.3	83.8	65.6	50.0	**84.7**
RP_b	RP	99.8	99.9	99.5	98.9	98.6	97.9	96.9	**98.8**
RP_b	RP_b	99.6	99.5	99.0	98.4	98.0	96.7	94.1	**97.9**
RP_b	LP	99.7	99.4	99.3	96.4	87.6	72.6	57.2	**87.5**

Benchmark

SIP-BART

All model are able to achieve an almost perfect accuracy across all test data.

Viktor Kjellberg

CONSISTENCY
(PROOF-CORRECTNESS)

- Nonexisting Rule
- Inapplicable Rule
- Spurious Match
- Unexhausted Search Space

Train	Test	Error Rate	Total Consis-tency
LP	LP	0.046	99.954
LP	RP	0.686	99.314
LP	RP_b	0.889	99.111
RP	LP	0.025	99.975
RP	RP	0.029	99.971
RP	RP_b	0.039	99.961
RP_b	LP	0.018	99.982
RP_b	RP	0.061	99.939
RP_b	RP_b	0.039	99.961

NONEXISTING RULE

The generated rule does not exist

- I.e. the rule in the BART output does not exist in the list of rules in the original problem description

Reasons:

- The model has commited a synonym-error. A similar word in meaning or a synonym was generated instead of the literal used in the original problem
- Only part of a rule was generated - one or more conditions are missing

Generated Inference Steps

pleasant → condemned
outstanding → average
courageous, dishonest → quaint
quaint, disgusted → hypocritical
True

Generated Inference Steps

courageous → rude
False

Model Input

QUERY:
hypocritical?

RULES AND FACTS:
old-fashioned → intellectual
long, pleasant, victorious → intellectual
fearless, dishonest → quaint
courageous, dishonest → supportive
quaint, disgusted → hypocritical
intellectual
fearless
pleasant
unpleasant

Model Input

QUERY:
ugliest?

RULES AND FACTS:
famous → adorable
selfish, confused, gifted → famous
fine, stormy → helpful
charming, courageous → rude
cute → amused
adorable, famous, fine → stormy
adorable, elated → helpful
frightened, rude → stormy
courageous

INAPPLICABLE RULE

- A generated rule can not be applied
  - If a generated rule does exist in the list of rules but the conditions for this rule can't be satisfied
- Reasons:
  - A part of a rule has been confused as a fact. This happens because the input is a string of the query, rules and facts

Model Input

quaint?, [….], outrageous ^ vivacious → inquisitive, cute, aggressive

Query | Rules | Facts

Generated Inference Steps

inquisitive → comfortable:
comfortable → quaint:
True

Generated Inference Steps

uptight → calm:
True

Model Input

QUERY:
quaint?

RULES AND FACTS:
inquisitive → comfortable:
comfortable → quaint:
....
serious → aggressive:
thoughtless, tidy, lonely → messy:
outrageous, vivacious → inquisitive:
cute
aggressive

Model Input

QUERY:
calm?

RULES AND FACTS:
attractive → since
uptight → calm
talkative → di
....
impatient →
horrible, dish
popular
curious
worried
blushing
innocent

SPURIOUS MATCH

- If the query does not exist in the facts and therefore has not been satisfied
- Reasons:
  - A synonym from the facts or the generated proof steps are confused as the query. A consequence of using a pre-trained model

Generated Inference Steps

True

Generated Inference Steps

busy → discusted
....
aggressive, attentive → adorable
True

Model Input

QUERY:
gorgeous?

RULES AND FACTS:
toughtless → intellectual
long, pleasant → average
fearless, dishonest → quaint
intellectual
unpleasant
glamorous

Model Input

QUERY:
cute?

RULES AND FACTS:
busy → disgusted
aggressive, attentive → adorable
fine, stormy → helpful
....
fine
busy
aggerssive
attentive

UNEXHAUSED SEARCH SPACE

- A consistent proof for a False problem that reaches the end is per definition complete
- If the whole search space has been exhausted and there is no longer any rules that can be applied
- No clear patterns why this appears more that it seems to miss one applicable rule

CONSISTENCY
(PROOF-CORRECTNESS)

- Nonexisting rule is the most common error for all models
- Nonexisting Rule and Inapplicable Rule are relevant for both True and False predicted proofs.
- Spurious Match is only relevant for proofs predicted True
- Unexhausted Search is only relevant for proofs predicted False

Train	Test	Non-existing Rule	Inapp-licable Rule	Spur-ious Match	Unex-hausted Search	Error Rate	Total Consis-tency
LP	LP	0.036	0.	0.007	0.004	0.046	99.954
LP	RP	0.661	0.	0.004	0.021	0.686	99.314
LP	RP_b	0.868	0.	0.	0.021	0.889	99.111
RP	LP	0.018	0.004	0.004	0.	0.025	99.975
RP	RP	0.007	0.018	0.	0.004	0.029	99.971
RP	RP_b	0.021	0.014	0.004	0.	0.039	99.961
RP_b	LP	0.011	0.	0.004	0.004	0.018	99.982
RP_b	RP	0.032	0.011	0.007	0.011	0.061	99.939
RP_b	RP_b	0.025	0.007	0.004	0.004	0.039	99.961

SUMMARY

- The SIP-BART models were able to achieve a high accuracy and consistency score.
- Almost all errors can be contributed to the use of a pretrained BART model and the inherent attention mechanism of Transformers.