Grounded NLI update

Utrecht University

February 20, 2025

Universiteit Utrecht

Method

- Draws inspiration from V-SNLI (Vu 2018): image-based NLI.
- Address NLI through images only? Text $->$ image $->$ text.
- Cognitive perspectives: realistic interpretation.
- Computational perspectives: no new/finetuned models required.

Method (illustrated)

- TTI: SD2, DALL-E-3
- Similarity scores: BLIP-1
- VQA: GPT-4 (BLIP-2, InstructBLIP discarded).
- Addressing neutral class problems: aggregation of multiple predictions.

Universiteit Utrecht

Experiment 1: results

- 100 premises & 300 hypotheses from SNLI.
- Competitive prediction quality.
- Struggle with neutral.

Task	Method	Aggregation	Score
Neut. vs Ent.	CSS	Oracle	81%
		Greatest Abs. Val.	73%
		Av. Val.	71%
		Maj. Class	70%
	VQA	Oracle	73%
		Av. Val.	69%
		Maj. Class	65.5%
Ent. vs Contr.	CSS	Oracle	97%
		Greatest Abs. Val.	95%
		Av. Val.	95%
		Maj. Class	95%
	VQA	Oracle	97%
		Average Val.	90%
		Maj. Class	93.5%

Universiteit Utrecht

Experiment 2: data

- Validate the SNLI experiment: consider the hypothesis-side bias (Gururangan et al. 2018).
- Sample 100 premises and 300 hypotheses from the hard and from the easy subset of SNLI (200 and 600 in total).
- Test against a fine-tuned model, supposedly prone to that bias (NLI-RoBERTa).

Experiment 2: results

- Decline switching from easy to hard also for VQA models.
- Decline is different for SD and DALL-E, image features matter.
- Hypothesis-side bias inversely correlated with textual overlap.

Task	Method	Easy	Hard
NLI	BLEU	39	45 (+6)
	RoBERTa	98.9	83.1 (-15.8)
	CSS DALL-E	70.2	69.1 (-1.1)
	VQA DALL-E	90	74.7 (-15.3)
	VQA SD	89.4	70.1 (-19.3)
Neut. vs Ent.	BLEU	36	40 (+4)
	RoBERTa	99.4	81.5 (-17.9)
	CSS DALL-E	61.7	61.9 (+0.2)
	VQA DALL-E	86.8	64.6 (-22.2)
	VQA SD	85	58 (-27)
Ent. vs Contr.	BLEU	43	55 (+12)
	RoBERTa	98.4	83.7 (-14.7)
	CS DALL-E	75.5	72 (-3.5)
	VQA DALL-E	95.7	85.1 (-10.6)
	VQA SD	92	81 (-11)

Universiteit Utrecht

Experiment 3: data

- Additional validation: consider the textual overlap heuristic.
- HANS: abstract terms, quasi-adversarial attacks (see discussion)
- Too much data in HANS given the pipeline throughput.

Universiteit Utrecht

Experiment 3: data

- Own data template, 100 points.
- *The [noun1] who [transitive_verb] the [noun2] [intransitive_verb].* (e.g. *The girl who greets the dog laughs.*)
- *The [noun1] [intransitive_verb].* (e.g. *The girl laughs*)

Universiteit Utrecht

Experiment 3: results

- Considerable advantage over RoBERTa.
- Different results with the two TTI models, i.e. text of hypotheses plays a limited role.

Task	Method	Adversarial
Ent. vs Contr.	RoBERTa	65.5%
	CSS DALL-E	56%
	VQA DALL-E	85%
	VQA SD	74%

Universiteit Utrecht

Specific issues

- Experiment 2 outcomes: does it prove the point?
- Experiment 3 dataset: replaceable by Winoground?
- Grounded image generation to do grounding: Lian 2023. LLM-generated layouts for stable diffusion.

Fundamental issues

- Does NLI need an ad hoc solution? Drifting away from the NLU paradigm.
- Impossible to work with abstract sentences.
- Vulnerability to adversarial attacks.
- New perspectives: aligning grounded and text-based output of VLMs. Possible image generation refinement.