Faithful and Plausible Natural Language Explanations for Image Classification

A Pipeline Approach

Adam Wojciechowski
a.wojciecho4@samsung.com

Mateusz Lango
lango@ufal.mff.cuni.cz

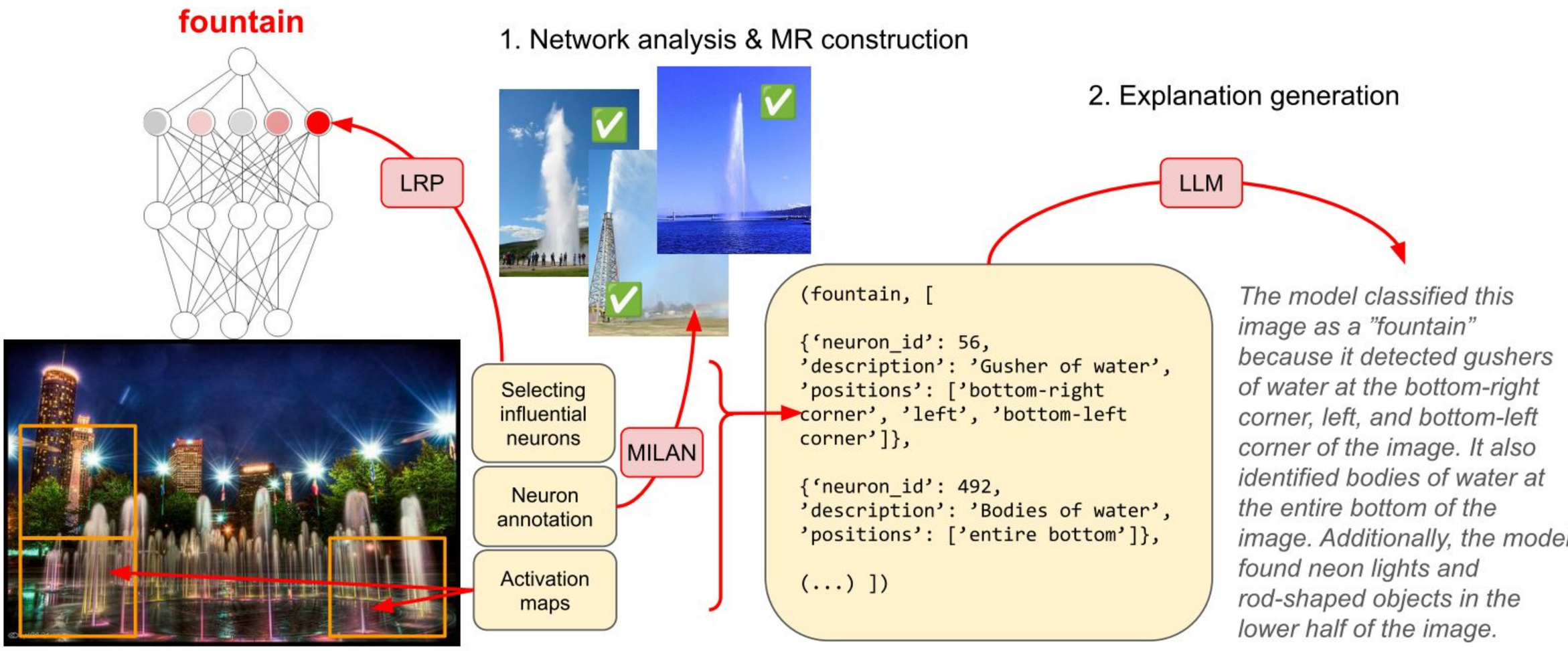
Ondřej Dušek
odusek@ufal.mff.cuni.cz

Charles University

Politechnika Poznańska – Poznan University of Technology

We turn **faithful** CNN neuron attribution results into **plausible** easy-to-read text using GPT-4

How it works?

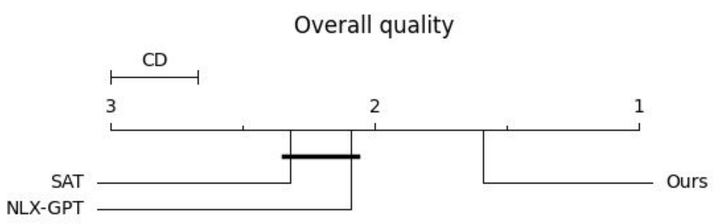


Is it plausible?

- **300 images** from ImageNet dataset, ResNet classifier
- **10 human evaluators**: 5 xAI experts and 5 non-experts
- Compared systems:
  - **SAT** - baseline image captioning method
  - **NLX-GPT** - baseline explainable visual QA method
  - **FLEX - our method** with neuron attribution + GPT-4

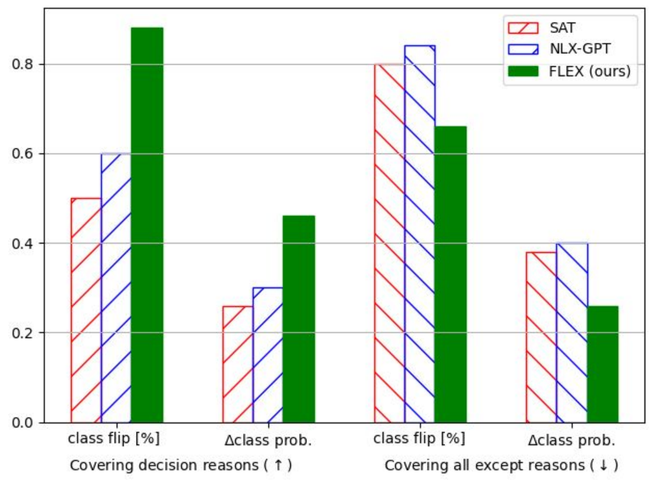
Evaluated aspect	SAT	NLX-GPT	FLEX (ours)
Fluency	**4.17**	3.46	**4.17**
Comprehensibility	**4.32**	3.65	3.71
Plausibility (convincing)	2.08	2.25	**3.07**
Plausibility (explanatory)	1.94	2.21	**3.17**
Overall quality	2.03	2.21	**3.00**

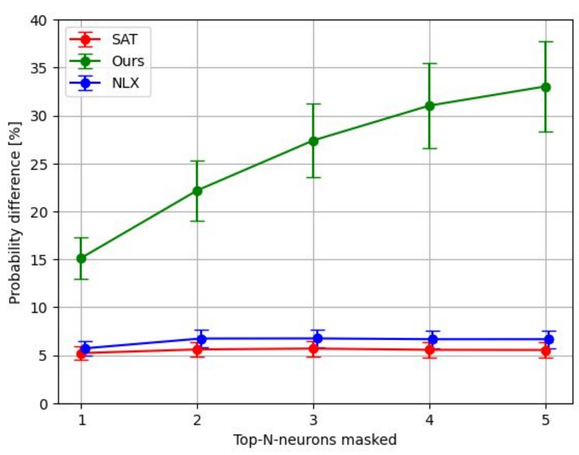
- **Statistically significant improvements** on plausibility and overall quality (Friedman & post-hoc Nemenyi test)



Is it faithful?

- **Covering areas** mentioned in FLEX's explanation **changes the original prediction in 88% of cases**.
- Covering produces significantly different explanations (90% different selected neurons)
- Humans using FLEX were able to **change 66% of predictions by masking only 5 neurons**
- Explanations are sensitive to noise (BLEU/METEOR ↓)
- Final LLM step **rarely introduces hallucinations** (8%).





Presented at EMNLP 2024, Miami, USA

Co-funded by the European Union (ERC, NG-NLG, 101039303) and National Science Centre, Poland (Grant No.~2022/47/D/ST6/01770). Using resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No.~LM2018101).

European Research Council
Established by the European Commission