

Faithful and Plausible Natural Language Explanations for Image Classification: A Pipeline Approach

Adam Wojciechowski^{1,3} and Mateusz Lango^{1,2} and Ondřej Dušek²

¹Poznan University of Technology, Faculty of Computing and Telecommunications, Poland

²Charles University, Faculty of Mathematics and Physics, Prague, Czechia

³Samsung AI Center Warsaw, Poland

a.wojciecho4@samsung.com, {lango, odusek}@ufal.mff.cuni.cz

Abstract

Existing explanation methods for image classification struggle to provide faithful and plausible explanations. This paper addresses this issue by proposing a post-hoc natural language explanation method that can be applied to any CNN-based classifier without altering its training process or affecting predictive performance. By analysing influential neurons and the corresponding activation maps, the method generates a faithful description of the classifier’s decision process in the form of a structured meaning representation, which is then converted into text by a language model. Through this pipeline approach, the generated explanations are grounded in the neural network architecture, providing accurate insight into the classification process while remaining accessible to non-experts. Experimental results show that the NLEs constructed by our method are significantly more plausible and faithful than baselines. In particular, user interventions in the neural network structure (masking of neurons) are three times more effective.

1 Introduction

Despite remarkable advances in computer vision, the deployment of image classification systems, especially in critical domains, poses significant challenges. One of them is the opacity of deep models and the difficulty of providing reliable explanations for their predictions (Doshi-Velez et al., 2017).

Therefore, several types of explanation methods have been proposed, including various forms of saliency maps (Selvaraju et al., 2017), feature importances (Ribeiro et al., 2016), concept-based explanations (Chen et al., 2019), counterfactual explanations (Vermeire et al., 2022), etc. A particularly interesting form of explaining predictions is offered by natural language explanation (NLE) techniques (Camburu et al., 2018; Wu and Mooney, 2019). Such explanations are not only understandable by non-expert users, but can also be used to

support conversations with the user in a dialogue system (Raczyński et al., 2023).

There are two critical properties of an explanation: *faithfulness* and *plausibility* (Jacovi and Goldberg, 2020; Atanasova et al., 2023). A faithful explanation should accurately reflect the inner workings of the system and provide information on the real reasons why the model reached a certain decision. Plausibility then refers to how convincing the explanation appears to the user.

In the case of NLE, obtaining high plausibility is straightforward, as textual explanations are usually human-friendly (Gurrapu et al., 2023), but achieving faithfulness is challenging. In the context of image classification, image captioning methods offer plausible but unfaithful NLEs (Xu et al., 2015; Kamakshi and Krishnan, 2023). Some methods try to improve faithfulness by conditioning generation on both the predicted class and image features (Hendricks et al., 2016; Kim et al., 2018; Marasović et al., 2020; Sammani et al., 2022), but the faithfulness provided is still limited as the model is not aware of the classifier’s decision process. Other methods train image classifiers to jointly predict the class and visual rationales, and generate explanations based on them (Wickramanayake et al., 2021; Kayser et al., 2022). However, the rationales are predicted independently of the class and do not participate in the classifier’s decision process. Most importantly, such methods change the training procedure and the architecture of the classifier, often affecting classification performance.

In this paper, we propose a post-hoc natural language explanation method for image classification that can be used with any standard convolutional neural network (CNN) classifier. To illustrate the classification decision process, the method analyses which neurons of the CNN were most influential in reaching a given decision, and which regions of the image caused them to activate. For each influential neuron, a neuron annotation method computes vi-

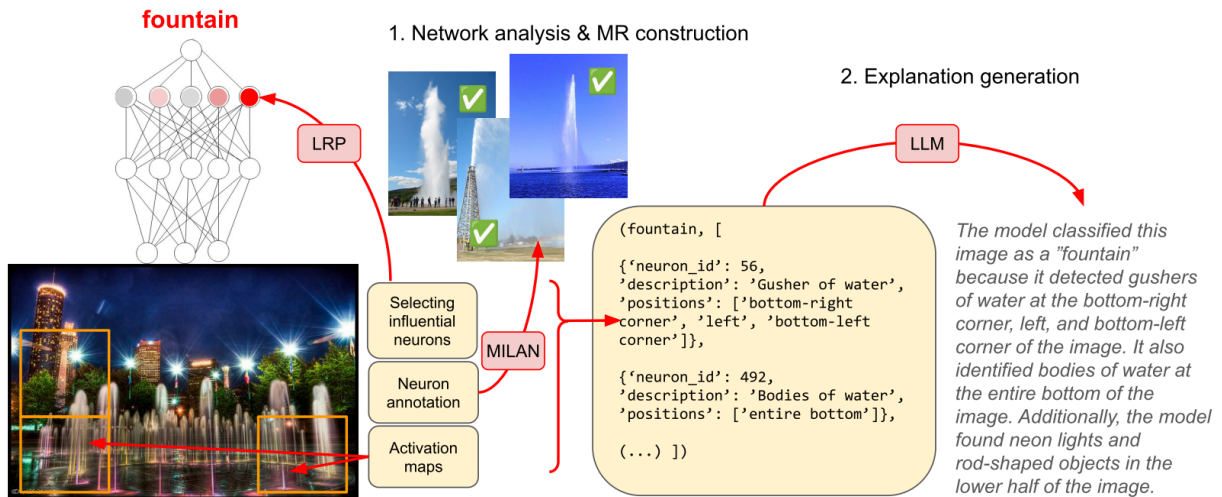


Figure 1: Overview of the presented approach. Note that the information provided in the text is supported by the model’s internal decision process, e.g. there is a convolutional filter specialized in the detection of "gushes of water", which was strongly activated at the mentioned positions of the image.

sual pattern exemplars and produces a short (one phrase) description of a pattern that the neuron detects. The information gathered in this process serves to create a simple meaning representation, which is converted to natural-language text by a large language model (LLM).

Our simple pipeline method (see Fig. 1 and Sect. 2) produces explanations that are directly grounded in the classifier’s network architecture, but without interfering with its training process or affecting its predictive performance. It also does not need any gold-standard training explanations. The provided NLEs reflect the process underlying classification by specifying the most influential neurons computed by well-established explainable AI methods. At the same time, the final text is fluent and easy to understand. This results in NLEs that are both plausible and highly faithful – significantly more so than baselines, as demonstrated in Sect. 3. Our experimental code is publicly available.¹

2 Method

Our NLE generation approach uses two processing steps, described below: meaning representation (MR) construction and MR-to-text conversion.

2.1 Meaning representation construction

We first produce a meaning representation in the form of a JSON object, containing information about the neurons responsible for a given classifier prediction (why?), what patterns those neurons detected (what?), and in which parts of the image

they were activated (where?). The MR includes the predicted class and the list of most influential neurons, each represented by (1) description – a phrase describing the pattern that most excites the neuron (convolutional filter), (2) positions – list of coarse-grained image positions (e.g., “bottom-right corner”) where the neuron was activated. An example of a MR is provided in Fig. 1.

MR construction starts by storing all neuron activations from the given CNN-based classifier for the prediction to explain. Next, the most important neurons are selected, annotated with a description, and tied to an image region, as follows:

Selecting the most influential neurons To pick the most relevant neurons, we apply the well-established Layer-wise Relevance Propagation (LRP) method (Bach et al., 2015). LRP performs a backward pass through the classifier network to establish the influence of each neuron to the final prediction (see App. H for formulas). We select k neurons with the highest LRP scores (with k being a parameter controlling the brevity-detail tradeoff).

Neuron annotation We adopt the MILAN neuron annotator (Hernandez et al., 2022) to generate descriptions of selected neurons. MILAN first finds images in the classifier training set that make a given neuron highly activated (Bau et al., 2017). These exemplar images are used to generate a description of the pattern that this neuron detects.

Note that although the last step of MILAN is essentially image captioning, it does not affect the faithfulness of NLEs produced by the pipeline, as

¹<https://github.com/wojciechowskiofficial/FLEX>

long as its output is of sufficient quality. The images that illustrate a neuron’s decision process are computed by analysing its activations, and the captioning is only used to convert the result into text.

Establishing image regions for neuron activation

The neuron’s raw activation map is divided into a 3×3 grid with manually assigned labels such as ‘top-left corner’, ‘top’, ‘top-right corner’, etc. We then select all grid cells where the neuron’s activation exceeded half of its maximum value. We apply several substitution rules (see App. B) to make the list of cells shorter and more human-readable.

2.2 Explanation generation

The second step of our method is converting the faithful MR created above into a user-friendly text. As we do not have any gold-standard explanation texts, the task is performed by prompting a large language model (LLM).

We instructed the model to (1) produce fluent text, (2) summarise the content of the MR (e.g. if two neurons detect similar patterns, they can be combined in the text), (3) prioritise readability, (4) come up with its own formulation of spatial positions to improve fluency. We also provide one handcrafted MR-to-text conversion example. The prompt is shown in App. C. LLMs could in theory hallucinate and thus reduce the explanations’ faithfulness. However, in Section 3 we show experimentally that current LLMs are reliable enough to produce useful explanations.

3 Experimental evaluation

3.1 Experimental setup

Dataset All experiments were performed on the ImageNet dataset. The classifier was trained on the training set and our explanation method was run to explain predictions made on the validation data.²

Models We experiment with explaining the predictions of the smallest CNN classifier from the popular ResNet family: ResNet18 (He et al., 2015).³ We fill our MRs with $k = 10$ top neurons indicated by LRP from the Captum library (Kokhlikyan et al., 2020) and annotate them using MILAN’s original implementation. As the LLM for the MR-to-text conversion, we employ GPT-4 (gpt-4-0613; Achiam et al., 2023).

²Annotated ImageNet test set is not publicly available.

³It reaches only a 41.4% accuracy on the validation set, but high classification performance is not the goal of our study.

Baselines We compare to the following methods:

- Show, attend and tell (SAT) by Xu et al. (2015) is an image captioning method used for explaining predictions (Kamakshi and Krishnan, 2023).
- NLX-GPT (Sammani et al., 2022) is an explainable visual question-answering method that produces NLEs with an encoder-decoder architecture that combines CNN with a transformer-based language model.

3.2 Are the output explanations plausible?

To assess the plausibility of generated explanations, we conducted a small-scale manual annotation experiment. We recruited ten annotators: five non-experts hired on the Prolific platform and five experts with at least one published paper on explainable AI. Each annotator was presented with 30 image-explanation pairs (300 in total) and asked to rate on a scale of 1-5 whether the explanations were (1) fluent, (2) easy to understand (comprehensible), (3) convincing, and (4) insightful for the underlying decision process.⁴ The overall quality of the explanations was also rated (see App. D).

The results are presented in Tab. 1 and examples of generated NLEs can be found in Tab. 3 (see App. A for more). Our method obtains the highest overall quality according to both experts and non-experts. It also produces the most plausible explanations (most convincing and insightful). Since the baselines produce much shorter explanations, it is not surprising that our longer explanations are a bit more difficult to understand. Interestingly, experts generally give higher ratings than non-experts for all methods and all factors except for providing insight into the decision process. Here, experts rate the baselines lower than non-experts, but they consistently rate the explanations provided by our pipeline higher. The improvements of our method over baselines are statistically significant on both plausibility measures and overall quality. For fluency, our method is indistinguishable from SAT (see details in App. I).

3.3 Are the output explanations faithful?

The faithfulness of the generated explanations is assessed through two intervention experiments: (1) checking if rationales from NLEs change the prediction by masking parts of input images, (2) in-

⁴Note that the question on understanding the decision process does not measure faithfulness, but the user’s subjective opinion on whether they understand how the model works.

	Experts			Non-experts			Overall		
	SAT	NLX-GPT	Ours	SAT	NLX-GPT	Ours	SAT	NLX-GPT	Ours
Fluency	4.70	4.12	4.64	3.64	2.80	3.70	4.17	3.46	4.17
Comprehensibility	4.94	4.42	4.18	3.70	2.88	3.24	4.32	3.65	3.71
Plausibility (convincing)	2.16	2.28	3.44	2.00	2.22	2.70	2.08	2.25	3.07
Plausibility (explanatory)	1.74	2.14	3.40	2.14	2.28	2.94	1.94	2.21	3.17
Overall quality	2.12	2.40	3.46	1.94	2.02	2.54	2.03	2.21	3.00

Table 1: The results of a human evaluation experiment in which NLEs provided by different methods were evaluated on 5 factors. The overall inner-annotation agreement is 0.53 as measured by Krippendorff’s alpha.

	Covering		Highlighting		Neuron mask.	
	c.f.↑	Δp ↑	c.f.↓	Δp ↓	c.f.↑	Δp ↑
SAT	0.50	0.26	0.80	0.38	0.20	0.06
NLX-GPT	0.60	0.30	0.84	0.40	0.19	0.07
Ours	0.88	0.46	0.66	0.26	0.66	0.34

Table 2: The results of three intervention experiments: percentage of examples for which user intervention resulted in a class flip (c.f.) and the average drop of probability of the predicted class (Δp).

	BLEU	MET.	c.f.	Δp .
Intra-set stability (5% noise)	41.33	0.610	0.32	0.153
Intra-set stability (20% noise)	30.87	0.521	0.81	0.255
Inter-set stability	26.01	0.469	n/a	n/a

Table 3: The results of stability analysis experiment: BLEU, METEOR (MET.), frequency of the class flip (c.f.) and drop of predicted class probability (Δp).

fluencing the network prediction by masking influential neurons. We further assess the stability and diversity of the explanations for our method, and we directly evaluate the reliability of our MR-to-text conversion.

Masking input image We asked annotators to *cover* with white rectangles parts of images that contained the decision rationale indicated in the NLE, 50% area at most (see App. G for details). We re-classified covered images and measured changes in prediction and the average decrease in the probability of the originally predicted class. We also performed an opposite experiment, with the annotators *highlighting* only parts of image mentioned in the explanation and covering the rest.

For covering, the use of our NLEs resulted in the highest average probability decrease and the change of the original prediction for 88% of examples (see Table 2). Our method reached the best results in the highlighting experiment as well, producing the least amount of changes.

We also re-ran our NLE pipeline with parts of the input image covered. This led to significant

changes: on average, 78% (median 90%) of the neurons indicated in MRs were different.

Masking influential neurons To show the NLEs’ ability to reflect classifier decisions, we asked the annotators to read the NLEs and select up to five most influential neurons from a MILAN-annotated list. The classifier was then re-run with the selected neurons masked. The results in Table 2 reveal that masking neurons suggested by our NLEs led to a five times higher decrease in the predicted class probability and over three times higher class flip rate than baselines.

More detailed results are presented in Fig. 2. Masking neurons in the order indicated by the annotators using our method leads to an increasing change in the classifier’s prediction and a gradual decrease in the predicted class probability value. In contrast, masking the neurons indicated using other methods leads to a small decrease in class probability for one masked neuron and almost no further decrease for more masked neurons. As we attribute the effect of the first masked neuron to examples of classes that are highly related to a singular pattern (e.g., a neuron annotated “water” for the class “sea”), this indicates that annotators gain very little insight into how the neural network made a decision from the explanations provided by the baselines.

Explanation stability analysis Following Wiegrefe et al. (2021), we measure explanation robustness against adding random noise to the input image (intra-set stability, see App. F for details) by comparing BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) as well as class flip frequency and class probability change against the original predictions. We also check for outputs’ diversity (inter-set stability) using BLEU and METEOR overlap against explanations for other classes. The results in Table 3 show that the explanations are both distinct for different classes and highly sensitive to noise: As we add more noise

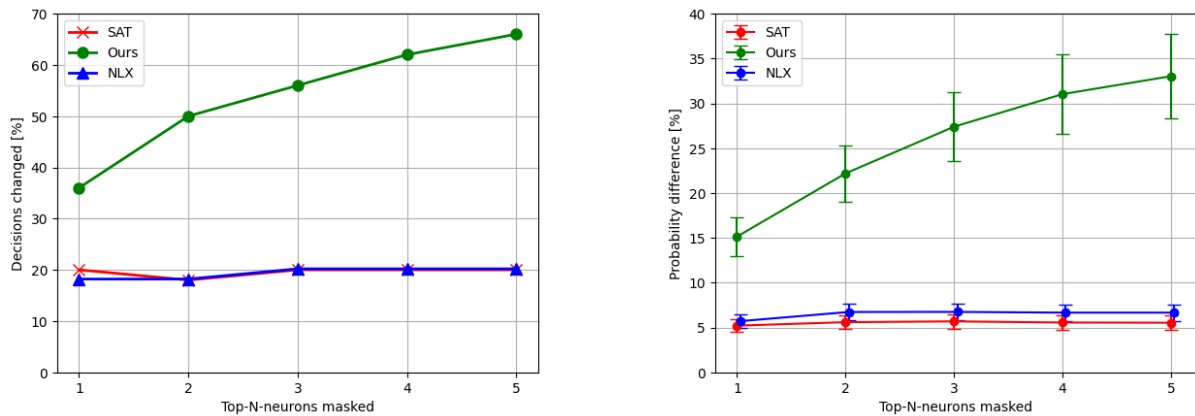


Figure 2: Results of neuron masking experiments when varying the number of masked neurons. Left: percentage of changed predictions after masking selected neurons; right: decrease of the probability of the predicted class after masking selected neurons.


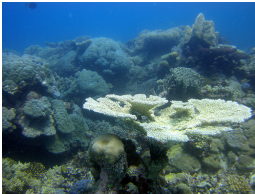

Image	Class	Explanation
	Target: "lakeside" Prediction: "lakeside"	The model classified this image as 'lakeside' because it detected water in the lower half and the bottom-right corner of the image. It also noticed volcanoes and mountains in the upper half, and walkways at the bottom-left corner and bottom. Additionally, the model found items with straight features at the left, bottom-left corner, and bottom of the image.
	Target: "coral reef" Prediction: "coral reef"	The model classified this image as a "coral reef" due to the detection of aquatic life throughout the image, particularly at the bottom, left, and right. The presence of water covering the entire image and nature-related elements further supported this classification. The model also noticed items that are connected in the lower half of the image and similar color patterns, which are common in coral reef environments.
	Target: "wall clock" Prediction: "wall clock"	The model classified this image as a "wall clock" because it detected clocks and other gauges, circular, round objects, and black and white objects at the bottom-right corner, center, and bottom of the image. It also noticed rounded edges in the right half of the image and the indent in an hourglass shape at the bottom-right corner, center, right, and bottom.

Figure 3: Examples of explanations provided by our method.

resulting in increased classification changes, the explanation BLEU and METEOR gradually drop, but they are at their lowest when comparing between different classes.

Reliability of the MR-to-text transform The human evaluation of our approach's MR-to-text reliability was similar to plausibility evaluation, but limited to non-expert Prolific annotators. We asked five yes-no questions on information in the text not grounded in the MR (i.e., hallucinations), omission of MR information, fluency, spatial information fidelity, and overall correctness (see App. E).

The results show the MR-to-text conversion as

highly reliable, as only 8% texts contain hallucinations. Omissions are more frequent (44%), but this is expected as the LLM is instructed to summarise the MR and prioritise readability. This factor most likely affected the overall score (58%). The explanations are mostly fluent (96%), with correct spatial information (82%).

Additionally, we repeated this experiment for explanations generated by an open LLM (Llama 3 70B) instead of GPT-4. The results presented in App. E show that open LLM generated NLEs with a significantly higher number of hallucinations and omissions, but this did not affect the overall quality score given by the annotators.

Acknowledgements

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

Limitations

This paper produces a new method for plausible and more faithful natural language explanations for image classification. Although we believe that the method provides significantly better faithfulness than the previously proposed methods, it does not obtain completely faithful explanations. The faithfulness of the explanations provided by our method depends on the quality of the neuron annotations produced by MILAN and the neurons indicated by LRP. Both techniques can be considered as state of the art, but they still occasionally produce incorrect results. Therefore, the results of NLE methods should be treated with caution. Additionally, this work uses pre-trained language models, which are known to expose certain social biases reflected in their training data.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse

Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. *GPT-4 technical report*.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen,

- and Isabelle Augenstein. 2023. [Faithfulness tests for natural language explanations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.
- Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. [On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation](#). *PLoS ONE*, 10.
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. [Network dissection: Quantifying interpretability of deep visual representations](#). *arXiv preprint arXiv:1704.05796*.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. [e-snli: Natural language inference with natural language explanations](#). In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.
- Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. [This looks like that: deep learning for interpretable image recognition](#). *Advances in neural information processing systems*, 32.
- Janez Demšar. 2006. [Statistical comparisons of classifiers over multiple data sets](#). *Journal of Machine Learning Research*, 7(1):1–30.
- Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O’Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. 2017. [Accountability of AI under the law: The role of explanation](#). *arXiv preprint arXiv:1711.01134*.
- Sai Gurrupu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A Batareseh. 2023. [Rationalization for explainable NLP: A survey](#). *Frontiers in Artificial Intelligence*, 6.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. [Deep residual learning for image recognition](#). *arXiv preprint arXiv:1512.03385*.
- Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. [Generating visual explanations](#). In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 3–19. Springer.
- Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. 2022. [Natural language descriptions of deep visual features](#). In *International Conference on Learning Representations*.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Vidhya Kamakshi and Narayanan C. Krishnan. 2023. [Explainable image classification: The journey so far and the road ahead](#). *AI*, 4(3):620–651.
- Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartłomiej Papież, and Thomas Lukasiewicz. 2022. [Explaining chest x-ray pathologies in natural language](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 701–713, Cham. Springer Nature Switzerland.
- Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. [Textual explanations for self-driving vehicles](#). In *Computer Vision – ECCV 2018*, pages 577–593, Cham. Springer International Publishing.
- Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. [Captum: A unified and generic model interpretability library for pytorch](#). *arXiv preprint arXiv:2009.07896*.
- Alon Lavie and Abhaya Agarwal. 2007. [METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.
- Scott M Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. [Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2810–2829, Online. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. [Rise: Randomized input sampling for explanation of black-box models](#). *arXiv preprint arXiv:1806.07421*.

Jakub Raczynski, Mateusz Lango, and Jerzy Stefanowski. 2023. [The problem of coherence in natural language explanations of recommendations](#). In *26th European Conference on Artificial Intelligence (ECAI)*, pages 1922–1929, Kraków, Poland.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. [“Why should I trust you?”: Explaining the predictions of any classifier](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. [NLX-GPT: A model for natural language explanations in vision and vision-language tasks](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8322–8332.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. [Grad-CAM: Visual explanations from deep networks via gradient-based localization](#). In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbosa de Oliveira, and David Martens. 2022. [Explainable image classification with evidence counterfactual](#). *Pattern Analysis and Applications*, 25(2):315–335.

Sandareka Wickramanayake, Wynne Hsu, and Mong Li Lee. 2021. [Comprehensible convolutional neural networks via guided concept learning](#). In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Sarah Wiegrefe, Ana Marasović, and Noah A. Smith. 2021. [Measuring association between labels and free-text rationales](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jialin Wu and Raymond Mooney. 2019. [Faithful multimodal explanation for visual question answering](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, Florence, Italy. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. [Show, attend and tell: Neural image caption generation with visual attention](#). In *International conference on machine learning*, pages 2048–2057. PMLR.

A Examples of explanations

Several examples of explanations provided by the methods under study are given in Tab. 4.

B Simplification rules for spacial positions

The spatial information about neuron activation is learned by inspecting activation maps of the chosen neurons. The activation map is converted to a binary image, such that if a given pixel of an activation map exceeds half of a maximal value in that activation map it is converted to the value of one, and the rest of the pixels are assigned with the value zero. Then the binary activation map is divided into a 3x3 grid of congruent squares oriented such that the ordering begins with a 0 in the top-left corner, progressing sequentially across each row from left to right and top to bottom, culminating with the number 8 in the bottom-right position. The basic positions names are given below.

```
0 : "top-left corner",
1 : "top",
2 : "top-right corner",
3 : "left",
4 : "center",
5 : "right",
6 : "bottom-left corner",
7 : "bottom",
8 : "bottom-right corner"
```

The final appearance of the positions appended to the meaning representation is governed by a set of rules, where position names are assigned to sets of basic positions. If all basic positions from the set are present in the MR, we replace them with a corresponding compound position. These sets and position names are given below.

```
"entire top": {0, 1, 2},
"entire bottom": {6, 7, 8},
"entire left": {0, 3, 6},
"entire right": {2, 5, 8},
"perimeter": {0, 1, 2, 5, 8, 7, 6, 3},
"center cross": {1, 3, 4, 5, 7},
"upper half": {0, 1, 2, 3, 4, 5},
"lower half": {3, 4, 5, 6, 7, 8},
"left half": {0, 1, 3, 4, 6, 7},
"right half": {1, 2, 4, 5, 7, 8}
```

Additionally if "entire *" and "*" half" coexist in the meaning representation simultaneously, where "*" represents one of {"left", "right", "top", "bottom"}, in both compound positions names, the "entire *" is deleted from the meaning representation, since it is comprised of the subset of the set, which makes up corresponding "*" half". An example would be if "entire top" and "upper half" are both in meaning representation, the "entire top" would be deleted since it is comprised of {0, 1, 2} tiles, which also partially make up the "upper half" ({0, 1, 2, 3, 4, 5}). Finally, if more than any, distinct, 6 out of 9 tiles are active, every position of a given neuron

	Our Method	NLX-GPT	Show, Attend and Tell
1	The model classified this image as a "cradle" due to the detection of items with circular features on the right half of the image, items with straight features in the center and on the entire right side, and rounded edges in pictures at the center, right, and bottom of the image. It also noticed the indent in an hourglass shape in the center, entire right, and bottom of the image. Additionally, human hands were detected at the center, right, and bottom of the image.	There is bathroom in the image because there is a sink and a toilet.	A black and white photo of a guitar case.
2	The model classified this image as "volcano" because it detected elements of nature in the lower half of the image. It also identified a gusher of water, which could be interpreted as lava, across various parts of the image including the left, center, right, and bottom. Additionally, the model found items with both curved and straight features in the lower half of the image.	There is mountain in the image because there is a large mountain in the background.	A view of a mountain range in a cloudy sky.
3	The model classified this image as a "library" because it detected shelves and books in the lower half of the image. It also noticed objects with led, text, and circular objects, rectangular objects, and cubed objects throughout the entire image. Additionally, items with straight features were found in the left half of the image, and grids were seen in the bottom-right corner and the right side of the image.	There is library in the image because there are bookshelves full of books.	A bookshelf filled with lots of books.

Table 4: Examples of image classification explanations provided by method under study.

is deleted and the spatial information placeholder is set to "entire image". This whole approach significantly reduces the length of the positional representation, while simultaneously making the position names more plausible to the system user, since we believe the final space of possible positions is natural and intuitive for humans to grasp quickly and effectively.

C Prompt for MR-to-text conversion

To convert meaning representations into text, the following prompt was applied to the language model:

You are given a problem of creating textual explanation of an image classification performed by neural network. You will be given a Python object representing network output in the form of ``image class', [(detected object', 'position'), ...]``. I want you to convert this object into a textual explanation. You should:

1. Create a grammatically correct sentence which will explain the model's decision.
2. Decide which detected objects do not fit with image class and do not include them in the explanation. For instance, 'dentist' class and 'animal heads' objects are completely unrelated. However, the descriptions that aren't directly related to image class, but can be indirectly correlated, especially in terms of shape, color, or texture resemblance should be included (like 'fountain' and 'sea' because of the water they

have in common or 'brick wall' and 'grid' because the texture is similar). Never mention that you chose neuron descriptions and do not talk about the neuron descriptions that were discarded.

3. Prioritize the readability of the explanation. Include only essential detected objects and aggregate information to shorten the explanation.
4. Aggregate positions if possible, for example ['bottom-left corner', 'bottom', 'bottom-right corner'] should be aggregated into 'bottom'. If the positions list is too long or too ambiguous do not include them in the explanation.

Here is an example.

Python object:

```

"lakeside, [{"description": 'Nature', 'positions':
['left', 'right', 'bottom']}, {"description":
'The sky', 'positions': ['top-right corner',
'bottom-left corner', 'bottom']}, {"description":
'Red and white colored objects', 'positions':
['left', 'right', 'bottom']}, {"description":
'The ocean', 'positions': ['left', 'right',
'bottom']}, {"description": 'Animal heads',
'positions': [], 'id'}, {"description": 'The color
red', 'positions': ['left', 'right', 'bottom']},
{'description': 'White backgrounds', 'positions':
['bottom', 'left']}, {"description": 'Grass',
'positions': ['left', 'bottom-left corner',
'bottom', 'bottom-right corner']}, {"description":
'Dogs and guinea pig', 'positions': ['center']},
{'description': 'The color green', 'positions':
['top-right corner', 'right', 'bottom-right
corner']}]"
```

Answer: "The model assigned this image to the "lakeside" class because in the last layer it discovered nature, and the ocean at the left,

right, and bottom of the image. It also detected grass at the left, bottom-left corner, bottom, and bottom-right corner and the color green at the right of the image.
 (...)

D Human evaluation of explanation plausibility

The annotators are presented with an image, model’s prediction and a natural language explanation. Each question is answered on a scale from 1 (low) to 5 (high). The following questions are asked:

- How fluent (linguistically correct) the text is?
- How easy to understand the text is?
- How convincing do you find the explanation of the decision made by the model?
- After reading the explanation, how well do you understand how the decision of the model was taken?
- How would you rate the overall quality of the explanation?

The annotation instructions are provided in the code repository.

E Human evaluation of MR-to-text transformation

The annotators are presented with a meaning representation in the form of formatted JSON without a given image, since it should not influence the assessment of MR-to-text transformation. The following binary questions are asked:

- Does the text contain information that was not present in the meaning representation?
- Is there any important information from meaning representation omitted in the text?
- Is the text linguistically correct?
- If any spatial compression occurred between explained neurons, is the said compression correct?
- Overall, do you find this meaning representation to text transformation acceptable, i.e. sufficiently good for explanation purposes?

Question	GPT-4	Llama 3 70B
Hallucinations ↓	0.08	0.46
Omissions ↓	0.44	0.64
Fluency ↑	0.96	0.86
Spacial compression ↑	0.82	0.82
Overall correctness ↑	0.58	0.58

Table 5: Human evaluation of MR to text conversion with GPT-4 and Llama 3 70B as backbone LLMs. The percentage of "yes" answers is reported.

The annotation instructions are provided in the code repository.

The experiment was carried out on explanations generated with two LLMs: GPT-4 (used in all other experiments) and an open-weight alternative, namely Llama 3 70B from ollama library⁵. Both sets of explanations were generated using the same prompt and for the same MRs.

The results are presented in Tab. 5. Both GPT-4 and Llama appear to have a similar ability to perform spatial compressions, but for the other factors examined, the NLEs generated by Llama fall short of those generated by GPT-4. Llama’s explanations are almost 6 times more likely to contain hallucinations. They also have more omissions and lower fluency. Nevertheless, the overall correctness of the NLE’s generated by both LLMs is the same and not overly high. We think that this result is influenced by the fact that some annotators penalise the correctness of NLE’s too much due to omitted information from the meaning representation. Note that the MRs are quite long and the generated NLEs are supposed to summarise them and shorten them to improve readability, thus omitting some information.

F Details on stability experiments

Let us assume that an image is a matrix X , such that $X \in [0, 1]$. We model input perturbations by adding random noise to the images, sampled from the standard normal distribution. Since an interval of possible pixel values is $[0, 1]$, to account for the unboundedness of a standard normal distribution we use a clipping operation defined as follows:

$$\text{clip}(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \leq x \leq 1 \\ 1, & \text{if } x > 1 \end{cases}$$

⁵<https://ollama.com/library/llama3>

The formula for the function ζ that adds a perturbation to the image is given below.

$$\zeta(X) = \text{clip}(\mathcal{N}(0, 1) \cdot i + X)$$

where i is the noise intensity.

We perform two types of stability experiments:

1. *intra-set stability*: We compare pairs of unaltered images (1st set of images) and corresponding images mapped by the ζ function (2nd set of images). After the ζ mapping is performed, we run the proposed method on both images, yielding two explanations. We then compute various language similarity metrics between the two explanations.

We compare two kinds of pairs of images: unaltered - lightly perturbed ($i = 0.05$) and unaltered - heavily perturbed ($i = 0.2$).

2. *inter-set stability*: We compare explanations pairs produced for different images to verify the diversity of generated explanations.

The computations were conducted on explanations produced by the proposed method for a 500-element subset of validation ImageNet data. The subset was constructed by randomly picking 50 examples from 10 selected, diverse classes (library, over skirt, palace, prison, wall clock, lakeside, coral reef, volcano, fountain, basset) in a stratified manner.

G Details on the covering experiment

Given an image and explanation, we instructed the annotator to cover the decision rationales with white rectangles. The annotator instruction is given below:

Based on the following explanation of the classifier’s prediction, cover the reasons for its decision with white rectangles. You can use as many rectangles as you like, but the total area of the covered image cannot be larger than half of the image. If it is not possible to cover the mentioned reasons by covering only 50% of the image, please do your best to cover the most important information.

An example of annotation provided is given in Fig 4.



Figure 4: Example image with and without human annotation in covering experiment.

H Layer-wise Relevance Propagation

To choose the most relevant neurons, the well-established Layer-wise Relevance Propagation (LRP) method is applied (Bach et al., 2015). LRP performs a special backward pass through the neural network to establish the influence of each neuron to the final prediction. Starting with the predicted value, LRP distributes it among the neurons in each layer, assigning them relevance scores. The following rule for relevance reallocation is used:

$$R_i = \sum_j \frac{z_{ij}}{\sum_i z_{ij}} R_j$$

where R_i is the i -th neuron relevance score, z_{ij} express how much i -th neuron has contributed to make j -th neuron relevant (calculated as the product of the neuron’s activation and the corresponding weight), the sums \sum_i (\sum_j) iterate over all neurons in a given (next) layer.

Choice of LRP as an explanation method Although we present a pipeline approach and there is some variability in how it can be implemented, we believe there are important reasons for using our pipeline with LRP.

First, unlike many other methods, LRP works at the neuron level, which is strictly required by our method. Therefore, methods that provide pixel-level importance scores such as RISE (Petsiuk et al., 2018) or Grad-CAM (Selvaraju et al., 2017), methods that typically work on image segments such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017) are not suitable for our approach.

Second, LRP has been shown to achieve high faithfulness in many studies and can be considered state of the art in this respect. Note that achieving high faithfulness is the main goal of our approach.

Finally, LRP is theoretically motivated, has been shown to be useful in many applications, and has stable open source implementations.

I Statistical analysis of the human evaluation results

For the results of human evaluation of plausibility, we performed the non-parametric global Friedman test followed by Nemenyi post-hoc analysis (as recommended in (Demšar, 2006)). We were able to reject the null hypothesis of the Friedman test for all the measures with $p < 0.001$. The Nemenyi post-hoc analysis with $\alpha = 5\%$ confirmed that our method obtains statistically significant improvements over other compared methods on both plausibility measures and the overall quality measure. On the fluency measure, our method is undistinguishable from SAT.

The critical distance plots from Nemenyi post-hoc analysis are provided in Figure 5. The lower result, the better. If the difference between the methods is not statistically significant, their results are connected with a thick horizontal line. More details on these plots can be found in (Demšar, 2006).

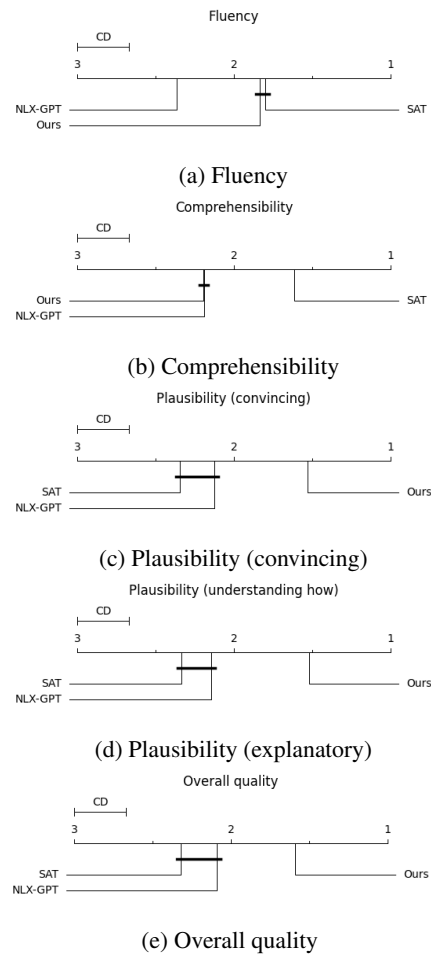


Figure 5: The results of Nemenyi post-hoc analysis for different aspects of evaluated explanations.