Explainable image classification based on word embeddings

Ing. Julija Jeršova, Prof. Ing. Tomáš Kliegr, Ph.D.

Faculty of Informatics and Statistics, Prague University of Economics and Business

Motivation

Modern AI image classifiers are powerful but often work as "black boxes," leaving users uncertain about why a prediction was made. Existing explainability methods, such as heatmaps, provide only low-level visual saliency and often lack semantic clarity. Our goal was to design an approach that makes AI reasoning fully transparent and human-understandable.

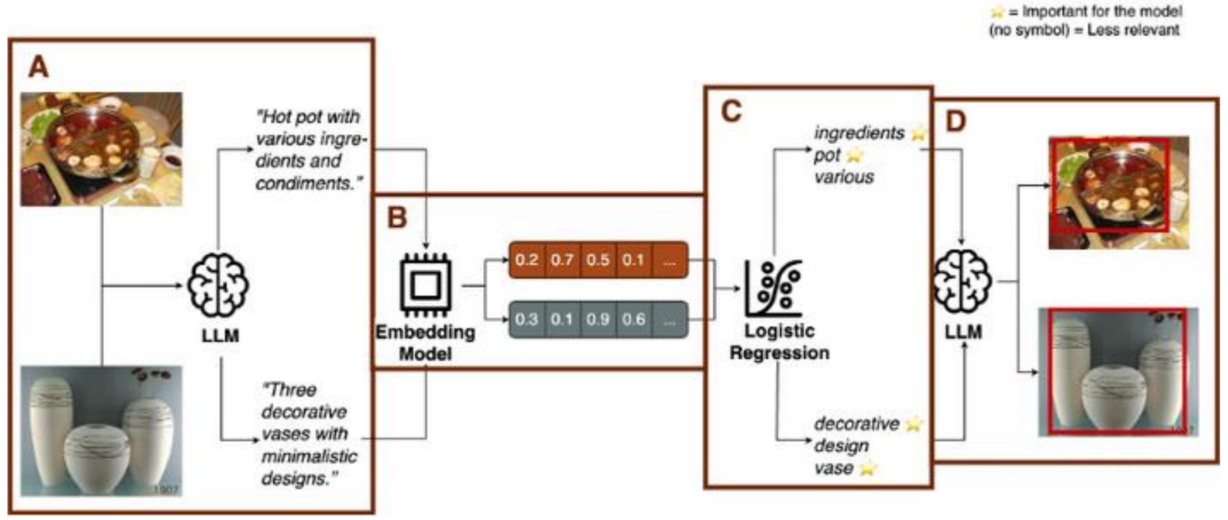


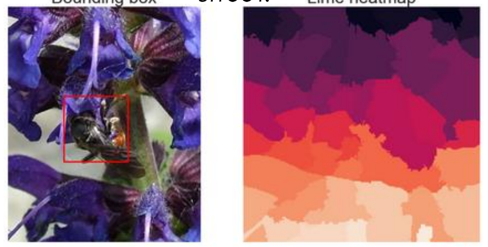
Figure 1- Workflow of the proposed method for simultaneous image classification and explanation using LLMs. In Phase A, training images are described using LLM-generated textual captions. In Phase B, the captions are represented as embeddings. In Phase C, important concepts are identified using logistic regression. In Phase D, bounding boxes are drawn using a multimodal LLM.

Methods

We developed a novel framework that explains image classification through semantic concepts instead of raw pixels:

- A multimodal Large Language Model (LLM) generates natural language descriptions of images.
- These descriptions are encoded into sentence embeddings.
- A logistic regression model is trained on averaged embeddings for classification.
- Explanations are provided via keyword-level attribution (SMER), highlighting the most influential words.
- The top-ranked terms are then localized in the image with bounding boxes, using targeted LLM prompting.

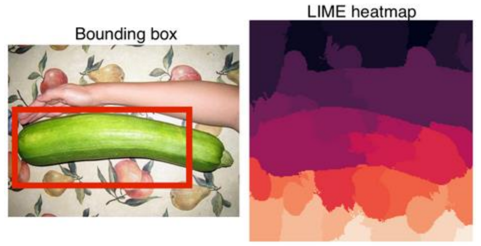
"White ambulance parked on a city street."



"Bee perched on thorny stem in sunlight."

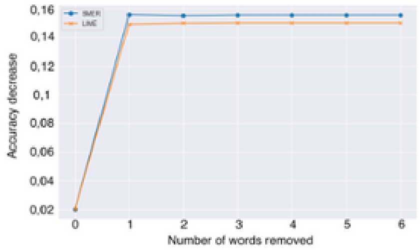
"Man playing violin with focus on strings."

"Three decorative vases with minimalistic designs."

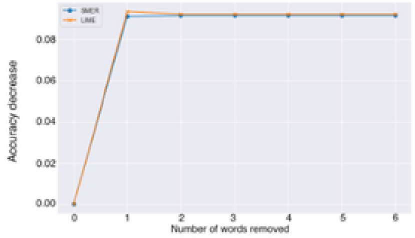


"Hand next to a large green zucchini."

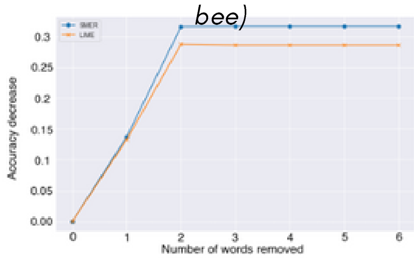
Figure 2:-Examples of a bounding-box explanation generated by our TEBOL method (left) and LIME-based heatmap explanation (right)



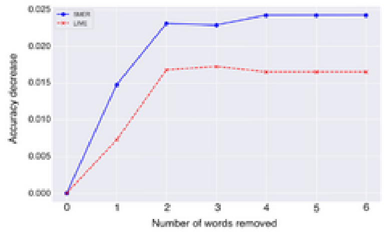
(a) (ambulance, fire truck)



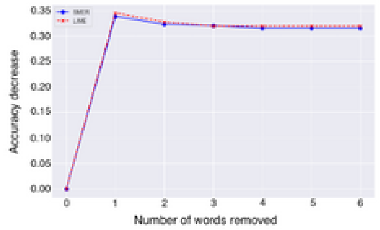
(b) (ant, bee)



(c) (acoustic guitar, violin)



(d) (hotpot, vase)



(e) (zucchini, cucumber)

Figure 3 - Comparison of SMER and LIME via AOPC curves across five binary classification tasks.

Results

- Evaluated on **14,351 ImageNet images (10 object classes).**
- **Achieved 91-98% accuracy, compared to 54-76%** for a VGG16 feature-based baseline.
- AOPC metric (Area Over the Perturbation Curve) – measures how much performance drops when important words are removed. **SMER consistently outperformed LIME** (e.g., 1.57 vs. 1.43 on violin vs. guitar).
- **User study with 270 participants**: <u>Bounding-box explanations were rated significantly more interpretable</u> than Grad-CAM heatmaps ($p < 0.001$).

Evaluation Aspect	Bounding Box	Heatmap	Significance (p-value)
Interpretability rating (Wilcoxon test)	**Very easy: 89** **Easy: 112** Comprehensible: 66 **Difficult: 3** **Very difficult: 0**	Very easy: 35 Easy: 91 Comprehensible: 116 Difficult: 27 Very difficult: 1	Yes ($p < 0.001$)
Interpretability (t-test)	**Mean = 4.06**	Mean = 3.48	Yes ($p < 0.001$)
Interpretability by image class (Wilcoxon test)	**Ambulance: 4.06** **Firetruck: 3.62** Ant: 2.84 Bee: 2.76 **Acoustic guitar: 3.33** Violin: 2.96 **Hotpot: 3.10** **Vase: 3.23** **Cucumber: 2.76** **Zucchini: 2.73**	Ambulance: 3.36 Firetruck: 3.49 **Ant: 3.48** **Bee: 3.68** Acoustic guitar: 2.97 **Violin: 3.37** Hotpot: 2.70 Vase: 3.02 Cucumber: 2.71 Zucchini: 2.53	Yes ($p < 0.001$) No Yes ($p < 0.001$) Yes ($p < 0.001$) Yes ($p < 0.001$) Yes ($p < 0.001$) Yes ($p < 0.001$) No No No
Open feedback sentiment (Chi-square test)	**Positive: 174** Neutral: 9 **Negative: 54**	Positive: 42 Neutral: 5 Negative: 55	Yes ($p < 0.001$)
Open feedback themes	Clarity, precision, ease	Visibility, overlap, color confusion	Yes (qualitative)

Table 1- Summary of evaluation results from the user study (N=270) comparing bounding boxes and heatmaps across interpretability and performance aspects.The p-value in the last column corresponds to the test type listed in the first column; better results for each pair are highlighted in bold.

Impact

Our findings demonstrate that language-aligned explanations enhance trust, clarity, and usability of AI predictions. The method bridges the gap between machine reasoning and human understanding, with potential applications in:

- Healthcare diagnostics (clearer AI-supported decisions for doctors)
- Education (teaching AI reasoning in an intuitive way)
- Everyday AI tools (improving transparency for non-experts)