Enhancing Multimodal Language Models with Olfactory Information

Murathan Kurfalı[1,*], Jonas K. Olofsson[1] and Thomas Hörberg[1]

[1]Sensory-Cognitive Interaction Lab, Department of Psychology, Stockholm University

Abstract

This paper explores the incorporation of olfactory data into multimodal language models, a relatively under-explored area in computational linguistics. We tackled the challenge of detecting olfactory stimuli in text and images, with a particular emphasis on multilingual contexts. Our approach involved enhancing the Large Language and Vision Assistant (LLaVA) model, through fine-tuning with a specialized dataset of around 2500 image-text pairs. By leveraging the open-source nature of LLaVA and the resource-efficient fine-tuning techniques such as Low-rank Adapter (LoRA), our study aims to contribute to the broader exploration of adapting language models to previously under-researched sensory modalities, such as olfaction.

1. Introduction

The field of multimodal machine learning has predominantly concentrated on text and image data. This focus is primarily because efficient representation techniques are readily available for these modalities. Conversely, other sensory dimensions, such as olfaction and gustation, have received less attention. This can be attributed partly to the challenge of incorporating their complex chemical structures into machine learning frameworks. However, it is important to note that these less-explored modalities are also implicitly present in both text and images, a facet that has been largely overlooked. The Multimodal Understanding of Smells in Texts and Images (MUSTI) addresses this gap by encouraging research in the detection of olfactory sources through texts and images in a multilingual context.

In this paper, we outline our contribution to MUSTI 2023 [1], which focuses on evaluating the capabilities of current open-source multimodal language models in identifying the sources of olfactory stimuli. To this end, we utilized the Large Language and Vision Assistant (LLaVA) model [2, 3], one of the most prominent multimodal open-source models available. Our research examines both the standard capabilities of the LLaVA model and its potential for fine-tuning with olfactory data. We observed significant performance improvements in the model when fine-tuned with a very limited dataset of approximately 2000 image-text pairs. We believe that the investigation of under-researched modalities in such models has great potential to advance the field. Successful results would showcase the effectiveness of currently available multimodal (text, image) language models in identifying olfactory sources and reveal the possibilities of using information from various senses to improve the comprehension of olfaction in these models. Therefore, in addition to developing more compact language models, such exploration is also related to broader questions of cognitive science, such as the interplay of different senses.

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online

*Corresponding author.

murathan.kurfali@su.se (M. Kurfalı)

Table 1
The prompts used in the fine-tuning step.

	Prompt
Subtask 1	Determine if the following text and image share common elements, with a specific focus on smell sources. Look for entities such as objects, animals, fruits, or any other elements that could be potential sources of smells. Answer YES or NO. Image: <image> Text: <text>
Subtask 2	Determine if the following text and image share common elements, with a specific focus on smell sources. Look for entities such as objects, animals, fruits, or any other elements that could be potential sources of smells. If you identify any such common elements, please list them in the same language as the provided text. If no common elements are found, simply respond with 'No common elements identified.' Image: <image> Text: <text>

2. Related Work

Prior research in olfaction through natural language processing is limited. [4] developed a FrameNet-like taxonomy to account for different aspects of the olfactory situations to facilitate more NLP-oriented research. Building on this, [5] developed a multilingual benchmark with manual annotations for these situations. [6] successfully trained a token classification model with this benchmark that can accurately identify olfactory elements even in modern out-of-domain texts like perfume reviews. [7] used word embeddings for an odor vocabulary in English, mapping odor descriptors and their olfactory-semantic organization. [8] further applied this to analyze sensory descriptors in wine, perfume, and food. Complementing text-based research, image-based olfactory reference extraction has advanced: [9] used CNNs for odor-object localization, [10] created an art dataset for olfactory recognition, and [11] developed a dataset for identifying smell gestures in historical artworks.

The closest line of research is last year's shared task [12]. [13] evaluated the then state-of-the-art multimodal models, ViLBERT and mUNITER, for detecting common olfactory references in multilingual text and images. The researchers formulated the task as a visual entailment problem and demonstrated significant performance improvements through model fine-tuning. [14] addressed the challenge by constructing a unified text-image object representation method for olfactory information where Yolov5[1] is used to represent image data and multilingual BERT for texts.

3. Approach

3.1. Model

Our approach employs LLaVA, which is a general-domain multimodal conversation model[2, 3], as the starting point and fine-tunes to the olfactory domain. LLaVA has a rather straightforward architecture, consisting of a vision encoder and a language model which are integrated through a linear projection layer. The various versions of LLaVA, namely 7B and 13B, are named based on the size of the Vicuna language model[15] used as the text encoder.

To optimize LLaVA for the designated subtasks, we transform the existing data into a format suitable for instruction-tuning. The specific prompts used in our final model is provided in Table 1. Despite the related nature of these subtasks, we ensured each prompt was self-contained. Therefore, from each text-image pair, two training instances were created. An important finding

[1]https://github.com/ultralytics/yolov5

Table 2
MUSTI train and development set data statistics. The positive and negative instances are based on the Subtask 1 labels.

Language	Train		Development		Total
	Positive	Negative	Positive	Negative	
English	179	538	19	59	795
Italian	179	541	19	60	799
French	92	179	10	19	300
German	86	347	9	38	480

during this phase was the need to guide the model to list objects in the same language as the input text, as it tended to default to English otherwise.

During the fine-tuning, instead of updating the entire model, we used LoRA which greatly reduces the number of learnable parameters by freezing the model and learning much smaller projection matrices between layers [16].

For our study, we followed the LLaVA model's official GitHub hyperparameters[2], experimenting with various LoRA settings but found no significant performance differences. Consequently, we chose a rank and alpha value of 16. We fine-tuned the LLaVA-13B model for three epochs on our dataset, noting no further gains with extended training, likely due to dataset size limitations. The fine-tuning was completed in under 3 hours using a 40GB A100 GPU at a batch size of 4.

3.2. Data

The MUSTI 2023 dataset comprises pairs of texts and images, selected to evoke olfactory experiences and sourced from historical archives. The dataset encompasses four languages: English (EN), German (DE), French (FR), and Italian (IT), and contains a total of 2,374 image-text pairs. However, the data is unbalanced, with only 593 pairs annotated as positive, meaning the existence of at least one common smell source. We created an in-house development set, constituting 10% of the total data, ensuring a similar distribution of positive and negative examples across all languages. The development set is primarily used for hyperparameter tuning and prompt tuning. Table 2 details the distribution of these pairs in the training and development sets, highlighting the number of positive and negative samples in each language.

4. Results and Analysis

In this section, we present and discuss the results of our models on the in-house development set allocated during the training phase, as well as the official results on the test sets. We participated in all three subtasks, namely i) identification of whether or not text passages and images evoke the same smell source, ii) listing them, and also iii) performing the same tasks for another language in a zero-shot setting.

During the development phase, we evaluated both the LLaVA-7B and LLaVA-13B models, before and after fine-tuning, as shown in Table 3. These models were assessed across multiple languages. The F1-macro scores revealed significant enhancements following fine-tuning. Notably, the fine-tuned LLaVA-13B model achieved a remarkable overall F1-macro score of 0.882, with particular improvements in recognizing positive samples. This suggests that the base models initially lacked the necessary knowledge for detecting potential smell sources. The LLaVA-7B model also showed competitive performance, especially considering its smaller size.

[2]https://github.com/haotian-liu/LLaVA/blob/main/scripts/v1_5/finetune_task_lora.sh

Table 3
The performance of the plain and fine-tuned LLaVA models on the development set. The Overall score is the F1-macro on all predictions on the entire development set.

Model	English		Italian		French		German		Overall
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Macro
LLaVA-7B	0.491	0.738	0.550	0.847	0.667	0.811	0.240	0.725	0.636
LLaVA-13B	0.581	0.896	0.240	0.857	0.167	0.783	0.308	0.889	0.619
Fine-tuned LLaVA 7B	0.865	0.958	0.743	0.927	0.700	0.842	0.800	0.962	0.860
Fine-tuned LLaVA-13B	0.833	0.950	0.824	0.952	0.706	0.878	0.875	0.974	0.882

Thanks to LoRA, the computational cost of fine-tuning was minimal, making the switch to a larger model have almost no tangible effect on the resources needed.

However, the most significant improvement was observed in subtask 2. On the development data, we noted that the plain LLaVA models scored almost 0 F-score for positive examples, i.e. when the evaluation ignored the correct classification of no common objects. This was because the models either failed to provide any list or listed all objects in the images. After fine-tuning, the performance of LLaVA-13B drastically improved to an F-score of 0.61, indicating that the model learned to discern which objects needed to be detected.

In the official results (Table 4), the fine-tuned LLaVA-13B model showed balanced macro precision and recall on both test and test-zero sets, performing well in identifying negative samples. The F1-Scores reached 0.776 for Subtask 1 and 0.698 for Subtask 2 in the test set. In the zero-shot scenario, performance dropped to 0.65 and 0.538 for Subtask 2, respectively. However, this is still promising, especially considering that Slovenian was not included in the pre-training or fine-tuning phases. These results highlight the fine-tuned LLaVA model's efficacy in recognizing olfactory data, marking a notable advancement in the capabilities of multimodal language models.

Table 4
Official results of our submission on the test sets. The overall column reports the macro average.

	Test				**Test (zero-shot)**			
Metric	**Neg**	**Pos**	**Overall**	**Subtask2**	**Neg**	**Pos**	**Overall**	**Subtask2**
Precision	0.819	0.676	0.774	-	0.662	0.706	0.684	-
Recall	0.874	0.575	0.781	-	0.848	0.458	0.653	-
F1-Score	0.846	0.621	0.776	0.698	0.743	0.556	0.65	0.538
Accuracy	-	-	0.781	-	-	-	0.675	-

5. Conclusion

In our research, we delved into the less-explored territory of integrating olfactory data into multimodal language models. Using the LLaVA model, we focused on recognizing olfactory cues in a diverse range of texts and images. By fine-tuning LLaVA with around 2500 image-text pairs and employing the Low-rank Adapter (LoRA) method, we achieved notable enhancements in the model's ability to detect olfactory stimuli. We believe that our findings highlight the potential of multimodal language models in processing sensory information beyond conventional texts and visuals.

References

[1] A. Hürriyetoglu, I. Novalija, M. Zinnen, V. Christlein, P. Lisena, S. Menini, M. van Erp, R. Troncy, The MUSTI challenge @ MediaEval 2023 - multimodal understanding of smells in texts and images with zero-shot evaluation, in: Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, the Netherlands and Online, 1-2 February 2024, 2023.

[2] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, arXiv preprint arXiv:2304.08485 (2023).

[3] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, arXiv preprint arXiv:2310.03744 (2023).

[4] S. Tonelli, S. Menini, FrameNet-like annotation of olfactory information in texts, in: S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, S. Szpakowicz (Eds.), Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, Punta Cana, Dominican Republic (online), 2021, pp. 11–20. URL: https://aclanthology.org/2021.latechclfl-1.2. doi:10.18653/v1/2021.latechclfl-1.2.

[5] S. Menini, T. Paccosi, S. S. Tekiroglu, S. Tonelli, Building a multilingual taxonomy of olfactory terms with timestamps, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, 2022, pp. 4030–4039.

[6] M. Kurfalı, T. Hörberg, J. K. Olofsson, Automatic detection of olfactory context elements, in: 15TH PANGBORN SENSORY SCIENCE SYMPOSIUM-MEETING NEW CHALLENGES IN A CHANGING WORLD (PSSS 2023), volume 6, 2023.

[7] T. Hörberg, M. Larsson, J. K. Olofsson, The semantic organization of the english odor vocabulary, Cognitive science 46 (2022) e13205.

[8] T. Hörberg, M. Kurfalı, J. K. Olofsson, Odor and flavor vocabulary in wine, perfume and food product reviews: insights from language modeling, Food Quality and Preference (under review).

[9] S. Kim, J. Park, J. Bang, H. Lee, Seeing is smelling: Localizing odor-related objects in images, in: Proceedings of the 9th Augmented Human International Conference, 2018, pp. 1–9.

[10] M. Zinnen, P. Madhu, R. Kosti, P. Bell, A. Maier, V. Christlein, Odor: The icpr2022 odeuropa challenge on olfactory object recognition, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 4989–4994.

[11] M. Zinnen, A. Hussian, H. Tran, P. Madhu, A. Maier, V. Christlein, Sniffyart: The dataset of smelling persons, in: Proceedings of the 5th Workshop on analySis, Understanding and proMotion of heritAge Contents, 2023, pp. 49–58.

[12] A. Hürriyetoglu, T. Paccosi, S. Menini, M. Zinnen, P. Lisena, K. Akdemir, R. Troncy, M. van Erp, MUSTI - multimodal understanding of smells in texts and images at mediaeval 2022, in: S. Hicks, A. G. S. de Herrera, J. Langguth, A. Lommatzsch, S. Andreadis, M. Dao, P. Martin, A. Hürriyetoglu, V. Thambawita, T. S. Nordmo, R. Vuillemot, M. A. Larson (Eds.), Working Notes Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023, volume 3583 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3583/paper50.pdf.

[13] K. Akdemir, A. Hürriyetoğlu, R. Troncy, T. Paccosi, S. Menini, M. Zinnen, V. Christlein, Multimodal and Multilingual Understanding of Smells using VilBERT and mUNITER, in: MediaEval Benchmarking Initiative for Multimedia Evaluation, 2022.

[14] Y. Shao, Y. Zhang, W. Wan, J. Li, J. Sun, Multilingual Text-Image Olfactory Object Matching Based on Object Detection, in: MediaEval Benchmarking Initiative for Multimedia Evaluation, 2022.

[15] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al., Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, See https://vicuna. lmsys. org (accessed 14 April 2023) (2023).

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).