Multimodal Learning for Image-Text Matching: A Blip-Based Approach

Dhanya Srinivasan[1], Subhashree M[1,*], Mirunalini P[1] and Jaisakthi S M[2]

[1]Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai - 603110, Tamil Nadu, India

[2] School of Computer Science & Engineering, Vellore Institute of Technology, Chennai Campus, Chennai-600127, Tamil Nadu, India

Abstract

This study delves into the domain of multimodal learning, focusing on image-text alignment to discern common olfactory references within multilingual content. The task aims for the Multimodal Understanding of Smells in Texts and Images (MUSTI) at MediaEval '23. The goal of this task is to address the gap between multimedia analysis and multimedia representation. The task aims to predict whether a text passage and an image evoke the same smell source or not. Our research employs the Bootstrapping Language Image Pre-training (BLIP) model which is a Visual Language Pre-training (VLP) model and capable of both vision-language understanding and generative tasks. We particularly engaged the BlipForConditionalGeneration model, a variant of BLIP, for image captioning to generate textual descriptions for the input images. These generated captions are matched with the corresponding text of the images based on the similarity score. Using the obtained similarity score a binary classification is performed using a multinomial Naive Bayes classifier. Our objective is to evaluate the effectiveness of amalgamating image captioning and text classification for this task. We employed a base model using BLIP and fine-tuned the same model and achieved an F1 score of 48.93% and 55.91% respectively.

Keywords: Deep learning model, Smells, BLIP

1. Introduction

Exploring olfactory information in images and text is crucial for historical and interdisciplinary research, shedding light on nuanced cultural contexts. Museums and galleries globally pioneer olfactory enrichments for immersive experiences, emphasizing the interdisciplinary potential and the importance of historically accurate olfactory settings. Automating olfactory information extraction has not gained much importance among the researchers since it is a challenging task to identify them in texts or images because of rare linguistic evidence in texts and its implicit representations in images [1]. Motivated by the profound impact of scent on emotions and memories, the MUSTI challenge at Mediaeval'23 explores the olfactory dimension in digital collections.

This paper focuses on the MUSTI subtask 1, expediting the understanding of olfactory references in multilingual text and images and forging connections between modalities. This task is a binary classification of whether an image and a text passage evoke the same smell source or not.

In this study we evaluate the effectiveness of amalgamating image captioning and text classification for this task. We assess the performance of the state-of-the-art model, BLIP, for

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

*Corresponding author.

†These authors contributed equally.

dhanya2010903@ssn.edu.in (D. Srinivasan); subhashree2010066@ssn.edu.in (S. M); miruna@ssn.edu.in (M. P); jaisakthi.murugaiyan@vit.ac.in (J. S. M)

image captioning, followed by the Multinomial Naive Bayes classifier to predict the classification labels on the test data of the MUSTI challenge. We provide insights into the performances of both the base and fine-tuned versions of this model.

In Section 2, we cite related work and references. Next, we provide a detailed explanation of our approach in Section 3. Following this, the results of the models in various configurations are reported in Section 4. Finally, a summary of our evaluation and an outlook conclude this paper in Section 5.

2. Related Work

The detection of smells, or olfaction, has traditionally been associated with human senses, but recent interdisciplinary research has extended this concept to image and text analysis. In image analysis, convolutional neural networks (CNNs), as demonstrated by [2], have shown promise in correlating visual patterns with specific smells. On the textual front, Natural Language Processing (NLP) techniques, as explored by [3], utilize word embeddings and semantic analysis to infer olfactory attributes from textual descriptions. A recent trend involves combining both modalities, as seen in the work of [1], where a multimodal deep learning architecture jointly analyzes images and textual descriptions for improved olfaction detection. Challenges include the subjective nature of olfactory perception and the need for large-scale annotated datasets, but ongoing research aims to refine multimodal models, explore transfer learning techniques, and address ethical considerations related to olfaction data in image and text. This nascent field holds promise for applications ranging from environmental monitoring to sentiment analysis of product reviews. A multilingual benchmark annotated with smell-related information which covers six languages are made available to the research community and they also discussed olfactory information extraction [4]. The performance of two state-of-the-art models, VilBERT and mUNITER on the MUSTI challenge test data and present the performances of base and fine-tuned versions of these models [1]. In [5] studies the relatedness of evocation of smells between texts and images generated was given as a task and overview of the participants model. performance analysis and dataset were also discussed. Shoa et al. [6] proposed a object detection based method for matching olfactory information in text and images. But this work faces a problem of data imbalance since the authors extract both positive and negative objects from the image. ICPR2022 ODeuropa Challenge [7] focused on recognizing odor-active objects in historical artworks. The winning team used PPYOLO-E [8] object detector with CSP-Resnet [9] backbone, trying grayscale image augmentation and style transfer for training, but found a strong object detection model to be most effective.

3. Approach

In this study, the research methodology unfolds through a systematic approach, addressing the complexities inherent in matching language descriptions with visual stimuli in the context of olfactory experiences. The methodology followed is an amalgamation of image captioning using BLIP model and text classification using a Naive Bayes classifier.

3.1. Data Collection and Preprocessing

The study employs a dataset comprising of image-text pairs sourced from [10]. The dataset includes information such as image filenames, text descriptions, language labels, and labels of objects invoking the smell if present. A metadata file was prepared using the image filenames

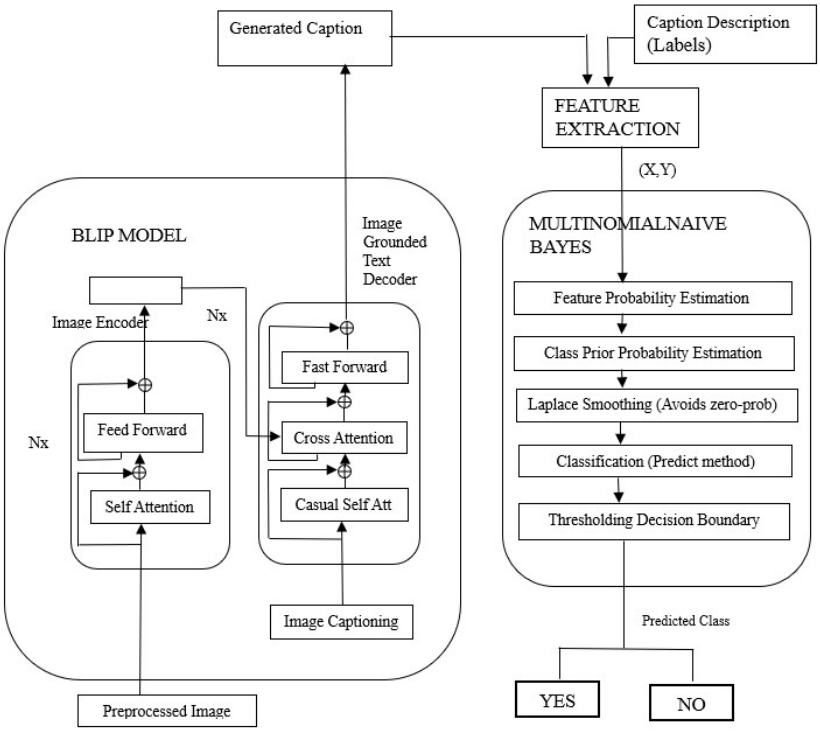

Figure 1: Proposed Architecture for Understanding Smells in Texts and Images

and their corresponding captions as given in the dataset. This was stored for the fine-tuning of the captioning model.

3.2. Image Captioning

For understanding smells in texts and images we propose to use the BLIP model [11] which is a VLP framework, basically used for vision language objectives such as: image-text contrastive learning, image-text matching, and image-conditioned language modeling. We have used two models, the first model used a pre-trained architecture (baseline model) and the latter one was a fine-tuned model where the hyperparameters were set.

The baseline model was directly used to generate image captions and the obtained captions were used further to understand the correlation between the images and texts. In the fine-tuned model, the processor functions were used as a wrapper to combine the two processors - BERT tokenizer and BLIP image processor into a single interface, allowing the model to handle both text and image inputs seamlessly during inference and training. It applies WordPiece tokenization on text, simultaneously resizing and preprocessing raw images into the format required by the model.

In the case of fine-tuned model, the same baseline model was tuned by loading the images as batches of 17 and then the model undergoes a rigorous fine-tuning regimen of 20 epochs, facilitated by the Adam with weight decay optimizer (AdamW) with a learning rate of 5e-5.

Short captions are then decoded from the output of the model for all images obtained using the URLs. These captions are then stored and used for subsequent classification.

3.3. Text Similarity Classifier

The Multinomial Naive Bayes classifier is chosen for its aptitude in discerning binary relationships, undergoing training for 3 epochs on the transformed training dataset, transforming texts using a CountVectorizer. The classifier is then used for binary classification based on the similarity of original text and generated captions. By ultimately assigning binary labels using a threshold, we attempt the classification task of identifying the correlation of smells across different modalities.

4. Results and Analysis

The state-of-the-art BLIP model demonstrated promising results in image captioning, generating textual descriptions for images. Subsequent integration with the Multinomial Naive Bayes classifier allowed us to construct a binary image-text similarity classifier. The performance metrics, including accuracy, precision, recall, and F1 score, were computed for both base and fine-tuned versions of the model. These metrics provide insights into the model's ability to correctly identify positive instances while minimizing false positives and false negatives.

Our analysis reveals the efficacy of combining image captioning with text classification for the MUSTI task. The utilization of a binary classifier helps determine the co-relation of smells across different modalities.

The model seems to perform significantly better when fine-tuned to the MUSTI dataset, as shown in Table 1. It achieves a precision of 67.42% and displays a moderate level of success in the classification task, in contrast to the 62.2% precision of the base model. The corresponding F1-scores of 55.91% and 48.93% for the fine-tuned and base models, respectively, further highlight the nuanced performance difference.

The variation of the metrics for each class "YES" and "NO" highlights the difference in the ability of the model to predict positive and negative classes. While the model is fairly successful in identifying and predicting negative classes in both the fine-tuned and base versions, it struggles in positive class prediction.

	Fine-tuned model				**Base model**			
Metric	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
NO	0.7895	0.7000	0.7420	150	0.7480	0.6333	0.6859	150
YES	0.3284	0.4400	0.3761	50	0.2466	0.3600	0.2927	50
Accuracy			0.6350	200			0.5650	200
Macro Avg	0.5589	0.5700	0.5591	200	0.4973	0.4967	0.4893	200
Weighted Avg	0.6742	0.6350	0.6506	200	0.6227	0.5650	0.5876	200

Table 1
Comparison of Classification Metrics for the English Language

In conclusion, the fine-tuned model, with a precision of 78.9%, recall of 70%, accuracy of 74.2%, and F1-score of 55.91% on the test data, demonstrates its superior performance in the given task.

5. Conclusion and Future Directions

In this study, we employ the VLP framework BLIP coupled with a Multinomial Naive Bayes classifier, achieving promising results through fine-tuning MUSTI data for English. Challenges

persist in automating olfactory information extraction due to limited linguistic evidence and implicit image representation. While BLIP excels for English, it falls short for multilingual data, and the use of Naive Bayes captures semantic similarity but struggles with detecting similar olfactory sources. Text tokenization, crucial for semantic understanding, may lead to information loss. As part of future work, we would like to further fine-tune the model in order to improve accuracy for the positive class and we may also explore advanced multimodal architectures and incorporate additional contextual cues to improve the model's grasp of olfactory references.

References

[1] A. Kiymet, H. Ali, T. Raphaël, T. Paccosi, S. Menini, Z. Mathias, C. Vincent, Multimodal and multilingual understanding of smells using vilbert and muniter, in: Proceedings of MediaEval 2022 CEUR Workshop, 2022.

[2] 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.

[3] S. Menini, T. Paccosi, S. S. Tekiroğlu, S. Tonelli, Scent mining: Extracting olfactory events, smell sources and qualities, in: S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, S. Szpakowicz (Eds.), Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 135–140. URL: https://aclanthology.org/2023.latechclfl-1.15. doi:10.18653/v1/2023.latechclfl-1.15.

[4] S. Menini, T. Paccosi, S. Tonelli, M. Van Erp, I. Leemans, P. Lisena, R. Troncy, W. Tullett, A. Hürriyetoğlu, G. Dijkstra, F. Gordijn, E. Jürgens, J. Koopman, A. Ouwerkerk, S. Steen, I. Novalija, J. Brank, D. Mladenic, A. Zidar, A multilingual benchmark to capture olfactory situations over time, in: N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, L. Borin (Eds.), Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1–10. URL: https://aclanthology.org/2022.lchange-1.1. doi:10.18653/v1/2022.lchange-1.1.

[5] 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.

[6] Y. Shao, Y. Zhang, W. Wan, J. Li, J. Sun, Multilingual text-image olfactory object matching based on object detection, in: Proceedings of MediaEval 2023 CEUR Workshop, 2022.

[7] 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.

[8] X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. Shen, J. Ren, S. Han, E. Ding, et al., Pp-yolo: An effective and efficient implementation of object detector, arXiv preprint arXiv:2007.12099 (2020).

[9] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, I.-H. Yeh, Cspnet: A new backbone that can enhance learning capability of cnn, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 390–391.

[10] Mediaeval 2023, https://multimediaeval.github.io/editions/2023/tasks/musti/, 2023.

[11] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.