A Text-Image Olfactory Matching Method Based on the Distribution of Real-World Data

Yi Shao[1], Yulong Sun[1], Wenbo Wan[1], Jing Li[1,*] and Jiande Sun[1,*]

[1]*Shandong Normal University, China*

Abstract

Correlation between olfactory information and human memory allows images and texts to rely on their content to separate from human olfactory cells and create imaginary olfactory experiences for humans. This means that images and text may contain equally rich olfactory information, and utilizing this olfactory information is inevitably limited by the distribution characteristics of image or text data, such as language gaps and long-tail distribution problems. To this end, this paper proposes a method based on target detection, which models similar olfactory information contained in images and texts into the same feature space, bridging the cross-language and cross-modal gaps, and adopts data augmentation and special sampling strategies respectively to alleviate the language imbalance of text data and the long-tail distribution of image objects.

1. Introduction

In this paper, we delve into the MUSTI task of MediaEval2023[1]. The MUSTI task is a text-image olfactory understanding challenge starting in 2022 [2], and existing works [3, 4] have already demonstrated the value and feasibility of this task. In MUSTI task of MediaEval2023, subtask 1 is to detect whether the image and text in each sample of the development set contain objects that cause the same olfactory experience. Further, subtask 2 is to point out what these objects are. Subtask 3 is to perform the above two subtasks on a zero-shot Slovenian dataset.

The texts in the development set are composed of English, French, German, and Italian, but the proportions of these four languages are uneven (en: 795, fr: 300, de: 480, it: 799). On the other hand, the images in the development set are mostly European medieval paintings, and the content contains a large number of objects of different classes, such as fruits, animals, portraits, jewelry decorations, etc. There is a long-tail distribution phenomenon which causes the model's detection performance for tail classes to drop sharply. Existing image and text retrieval research rarely involves olfactory information, so we build the model based on the characteristics of the data set and combined with traditional target detection algorithms.

2. Approach

2.1. Stage 1 for Coarse-Grained

As shown in Figure 1, our proposed method is divided into Stage 1 for coarse-grained matching and Stage 2 for fine-grained matching. The first stage is used for coarse-grained classification,

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

*Corresponding author.

†These authors contributed equally.

2021020981@stu.sdnu.edu.cn (Y. Shao); 2022020647@stu.sdnu.edu.cn (Y. Sun); ;wanwenbo@sdnu.edu.cn (W. Wan); lijingjdsun@hotmail.com (J. Li); jiandesun@hotmail.com (J. Sun)

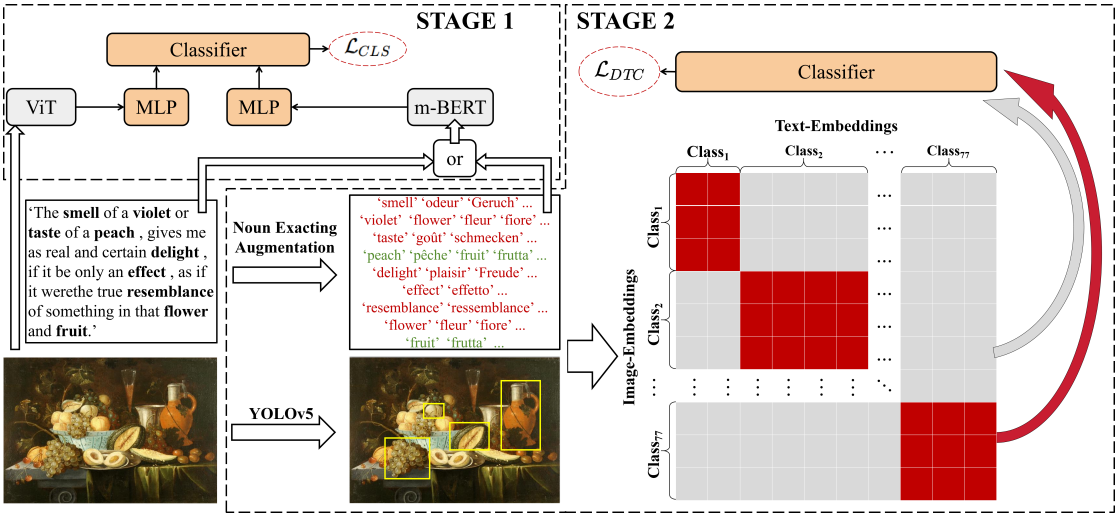

Figure 1: Overview of the two-stage model.

that is, to detect whether there are similar olfactory objects in images and texts, using multilingual BERT (m-BERT)[5] and ViT[6] respectively to extract text features and image features. For text features, we input complete sentences and nouns after data augmentation into m-BERT to obtain a fine-tuned model based on sentence features. The purpose of data augmentation is to alleviate the problem of cross-language gaps in the dataset. Specifically, we translated the nouns in each group into four languages and added superclasses which come from official data. After these two text feature selection strategies, the text features will be spliced with the image features, and the binary cross-entropy loss of coarse-grained stage $\mathcal{L}_{CLS}$ will be calculated.

2.2. Stage 2 for Fine-Grained

The second stage is used for fine-grained classification, that is, detecting which objects cause similar olfactory experiences in images and texts. We counted the subtask 2 labels of all samples, and categorized objects observed across all images into 77 classes for marking bounding boxes for training YOLOv5. A hyperparameter k which means an image contains objects whose total occurrence frequency in all images is less than k was introduced to alleviate long-tail distribution problems in theses classes. The performance of YOLOv5 trained on training sets divided with different k values is shown in Section3. The text feature extractor and image feature extractor trained in the first stage will be used as the initial state of the second stage to extract noun features and image features in each bounding box respectively. A pair of text features and image features of the same category are regarded as Positive samples, otherwise they are regarded as negative samples. These feature pairs will then be fed into the classifier to calculate the binary cross-entropy loss $\mathcal{L}_{DTC}$.

3. Results and Analysis

3.1. Preliminary Experimental Works

Images in the dataset had several obvious content categories: food, flowers, animals (including prey and herds), water-related (such as the sea, ports, rivers, whales, fish, etc.), jewelry decoration, personal portraits, crowd gatherings, etc. Most of these categories have significant

Table 1
Comparison of accuracy of different text feature extractors in Stage 1

Model	similarity		classifier			
	sentence	nouns	sentence	nouns	sentence	nouns
BERT	bert-multilingual-cased		bert-multilingual-cased		bert-multilingual-uncased	
Acc.	0.7579	0.7558	0.8110	**0.8232**	0.8063	0.7979

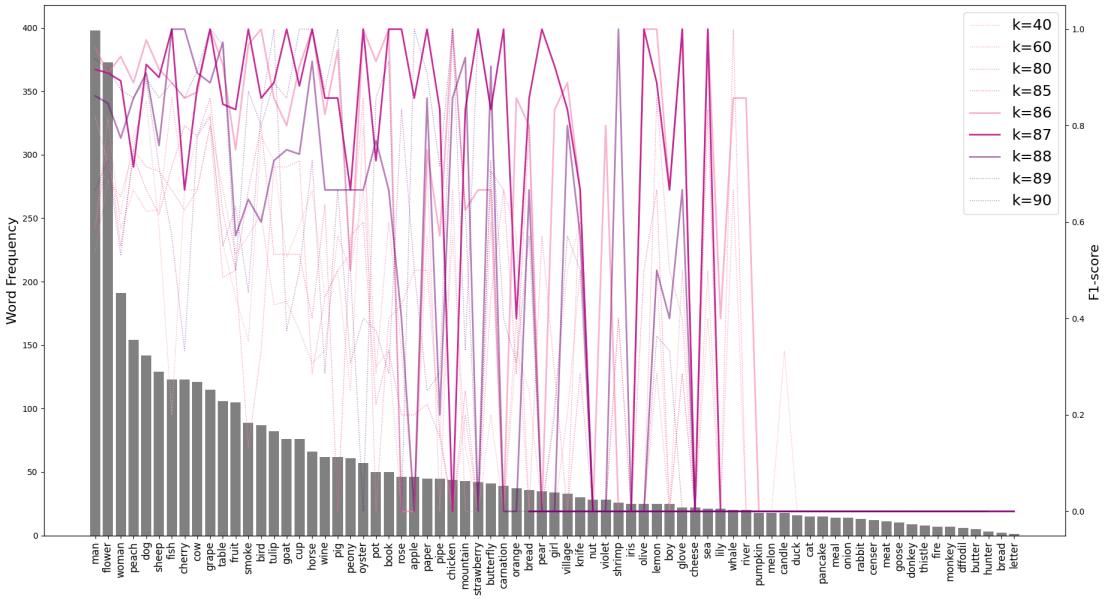

Figure 2: Performance of YOLOv5 trained on training sets divided with different k values.

frequency of specific objects and significant olfactory characteristics, so we respectively converted the images to gray scale images or retained the original color content, and respectively used fine-tuned ResNet50 as classifier and K-means algorithm for clustering, but all the performances were poor. In Stage 1, the text feature extractor of the model uses a variety of official HuggingFace official versions and the performance comparison is shown in Table 1.

In Stage 2, we select training samples through k each time and then randomly select samples to ensure that the training set:validation set is 8:2, and there is at least one sample for each category in the validation set. When k=0, the training set is completely randomly selected. The proportion of selected samples in the development set is 50.92% when k=40, while it is 80.63% when k=90. The impact of different values of super parameter k on the performance of YOLOv5 in Stage 2 is shown in Figure 2. The larger the k value, the more categories can be ensured to appear stably in the training set. However, a too large k value will also cause some tail categories to appear too few times in the validation set, causing the randomness of the validation results to increase. The overall F1 score of YOLO on the head categories gradually increases as the k value increases, and roughly converges when k=86. But as the k value increases to 90, the F1-score of the head class drops significantly. The overall detection F1 score on the tail category first increases and then decreases as the k value increases, reaching the overall optimum when k=87. For extreme tail categories, a smaller k value can result in a lower F1 score, while a slightly larger k value will cause YOLO to no longer be able to detect extreme tail categories. Therefore, we finally select k=87 to take into account the performance of all head and tail categories. In addition to text data, we also tried data augmentation on image data in

Table 2
Comparison of F1-score of images and texts in different languages on Stage 2

Subtask	Models	F1-score				
		en	de	fr	it	micro avg.
Subtask1	dummy baseline [3]	0.4285	0.4289	0.3333	0.4273	0.4075
	mUniter finetuned [3]	0.4473	0.4644	0.3605	0.5020	0.4473
	mUniter-MUSTI [3]	0.6965	0.4579	0.5022	0.6535	0.6011
	mUniter-SNLI-MUSTI [3]	0.7482	**0.5014**	**0.5053**	0.6850	**0.6176**
	Yi et al. [4]	**0.7867**	0.4568	0.3743	**0.7501**	0.6033
	ours	**0.7829**	**0.4845**	**0.5133**	**0.7074**	**0.6198**
Subtask2	Yi et al. [4]	0.7427	0.7276	0.4599	0.7487	0.6708
	ours	-	-	-	-	0.0572
Subtask3	**ours** (Subtask1)	-	-	-	-	0.3845
	ours (Subtask2)	-	-	-	-	0.0258

Stage 2. We used Stable Diffusion to try to generate some medieval-style meals and portraits, but they all have visible differences from the real medieval paintings in the development set. Besides, it is difficult to establish effective tail data augmentation for paintings with various styles and clutter based on affine transformation. For examples, an apple may look like a peach after the color is changed, while a candle is difficult to detect after rotation because the candles in other samples are vertical and the pipes are usually in the form of smoking slanted strips.

3.2. Comparison Experiments

The comparisons of the F1 score of three subtasks are shown in Table 2. The overall structure of Yi et al. [4] on subtask1 is consistent with the proposed method, but the data augmentation method is not used for cross-language bridging. On German and French samples that number is small, the proposed method has obvious improvements compared with [4], which proves that the text data enhancement we use has a strong ability to bridge the language gap.

On subtask 2, the proposed method is far inferior to [4] because it does not treat different words representing the same category as the same word, unlike [4], but directly performs binary classification calculations on all different words and image regions. Our original intention was to build a model that can directly map the textual word features and image region features to the same feature space, thus eliminating the process of manually organizing different words representing the same category. However, based on performance comparison, the proposed method significantly degrades performance compared to [4] due to the proposed method requires a small feature differences between each image region and each text word of the same category. Compared with the manually compiled list of approximate olfactory nouns in [4], the proposed model obviously saves labor costs significantly, but based on the huge drop in performance, this method is not available.

Acknowledgments

Thanks to the organizers of the MediaEval2023, especially to those organizers for MUSTI. This work was supported in part by the Joint Project for Innovation and Development of Shandong Natural Science Foundation (ZR2022LZH012) and Joint Project for Smart Computing of Shandong Natural Science Foundation (ZR2020LZH015).

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] 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.

[3] K. Akdemir, A. Hürriyetoglu, R. Troncy, T. Paccosi, S. Menini, M. Zinnen, V. Christlein, Multimodal and multilingual understanding of smells using vilbert and muniter, 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/paper36.pdf.

[4] Y. Shao, Y. Zhang, W. Wan, J. Li, J. Sun, Multilingual text-image olfactory object matching based on object detection, 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/paper15.pdf.

[5] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).