

# Handle the problem of ample label space by using the Image-guided Feature Extractor on the MUSTI dataset

Le Ngoc-Duc<sup>1,†</sup>, Le Minh-Hung<sup>2,†</sup> and Dinh Quang-Vinh<sup>3,\*</sup>

<sup>1</sup>*Ho Chi Minh City University of Science, Vietnamese*

<sup>2</sup>*Hanoi University of Science and Technology, Vietnamese*

<sup>3</sup>*Vietnamese-German University, Vietnamese*

## Abstract

Among multimodal tasks, olfactory perception remains a largely unexplored field. The two most significant difficulties that need to be overcome are that the label space is ample while the data set size is generally of too small volume. The second is the imbalanced nature of labels in the data set. In this paper, we develop and evaluate our model in the task of predicting the congruence of olfactory experiences between an image and a corresponding text passage on the MUSTI dataset. To solve the label imbalance problem and optimize the process of extracting multimedia images and text with large feature spaces, we propose a model that selectively selects the text features based on image features. By selecting texts that need attention, our model outperforms existing baselines on training and testing data sets. Code available at: <https://github.com/Haru-Lab-Space/MMM2024.git>.

## 1. Introduction

This paper aims to enhance the process of multimodal text-image retrieval through ensemble learning with image extraction models that have been trained on large datasets. Besides, Image-guided Feature Extractor can help the model encode text more effectively by providing it with potential values that should give more attention. Thanks to the excellent and diverse results from combining image extraction models, text encoders can now receive the most optimal summary information. Then, the information extracted from images and text is brought into a shared space, and their similarity is evaluated based on cosine distance. We evaluate our model results in the task of predicting the congruence of olfactory experiences between an image and a corresponding text passage, on the MUSTI dataset and obtain superior results in all parameters compared to the baselines. Through models like these, museums and galleries can enhance the experience for users with special needs, such as those with visual impairments. Besides, it can also become a new future for the perfume industry when they can recreate the history of its formation and development [1].

In the remainder of this paper, we present a comprehensive exploration of our proposed model and its implications for the field. Section 3 delves into the intricacies of our proposed model and the underlying motivation for its development. Section 4 is dedicated to presenting and analyzing the results of our experiments, comparing the performance of our proposed model against established benchmarks. In Section 5, we further investigate by exploring various model variants, elucidating the nuanced differences in performance.

---

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


\*Corresponding author.

†These authors contributed equally.

✉ [ngocducdarkzonezero@gmail.com](mailto:ngocducdarkzonezero@gmail.com) (L. Ngoc-Duc); [leminhhung933@gmail.com](mailto:leminhhung933@gmail.com) (L. Minh-Hung); [vinh.dq2@vgu.edu.vn](mailto:vinh.dq2@vgu.edu.vn) (D. Quang-Vinh)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

## 2. Related Work

Multimodal image-text retrieval aims to enable efficient retrieval of image-text correlations. It is a difficult task because of the differences in the information representation space of the features. To overcome that, Y. He used two convolutional neural networks to adapt to learning features between images and text [2]. By bringing them to the same representation space, the model evaluates their similarity through cosine similarity. Y. Zhan constructs a multimodal projection and evaluates the difference between pairs of input images and text [3]. A. Baldrati synthesizes the characteristics of the two phases according to the features to find the differential features between query and target images [4]. H. Dong proposed a model that leverages text information to teach the image encoder to retrieve features based on the graph [5]. Although there have been many efforts to improve the ability to grasp image and text objects, previous research methods performed information extraction fragmentarily. The information generated from separate encoding and extraction units creates information-rich vectors, which unintentionally confuses the model when choosing where to pay attention among countless potential options.

## 3. Approach

### 3.1. Cross-Feature Encoder

Through problem analysis, we realized we could use some of the features found in images and text. Those information-rich feature embeddings sometimes complicate the inference process and confuse the model when concluding. Realizing this, we built a model consisting of three main components: Image Encoder, Image-guided Feature Extractor, and Text Encoder.

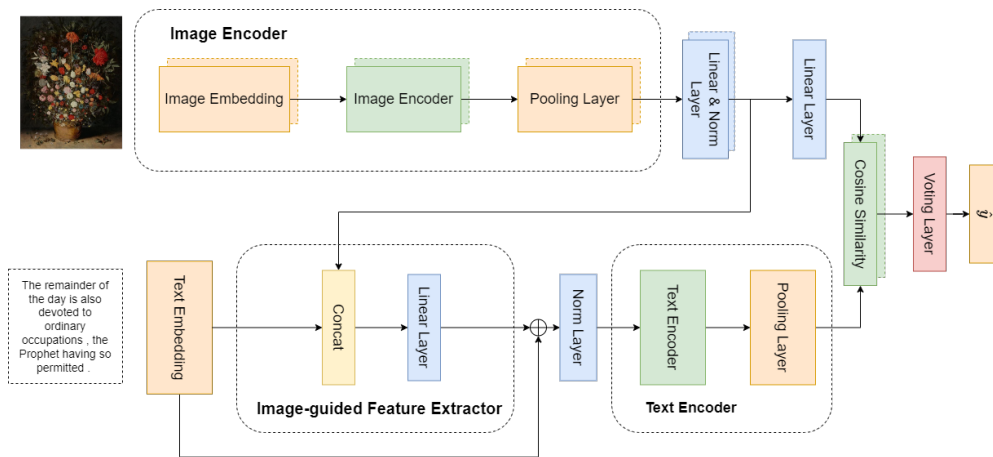


Figure 1: Overview of the proposed model.

#### 3.1.1. Image Encoder

The image encoder is a set of encoders of the Vision Transformer [6] and Resnet-34 [7] models. We use the image as input to all two of these encoding units. Because the output size of each model is different, we use linear layers to bring them to the same size as the text embedding.

### 3.1.2. Image-guided Feature Extractor (IGFE)

The idea behind this processing block is that we only need a minimal amount of features compared to what the actual text encoder can do. Therefore, we use the output of the image encoder to target the text embedding to the needed features. This helps the text encoding unit to pay more attention to essential features. We do this by successively appending the text embedding to the outputs of the image encoding units. A linear class here will decide how much of the image encoder’s information is retained to become the "instructor".

### 3.1.3. Text Encoder

The **instructors** and text embeddings are then added together and normalized before being fed into the text encoder. In this setting, we use BERT’s multilingual text encoder - a text processing model that consists only of Encoder classes [8].

## 3.2. Prediction

We use a cosine distance measure and a weighted soft aggregation method based on the weights of a linear layer. Thanks to the flexibility of the weights, we can allocate the decision influence of each model accordingly.

## 4. Results and Analysis

We found that our model outperformed every metric compared to the baseline when evaluating the model’s accuracy on the test dataset. We took the measurement data of the baseline models directly from the author’s article [9] and [10].

**Table 1**

Performance of different methods for data imbalance problem on test set.

Model	F1-score
Yolov5 + BERT [9]	0.6033
mUNITER-SNLI-MUSTI [10]	0.6176
Our ( $\gamma = 2, \alpha = 0.25$ )	0.7196
Our ( $\gamma = 2, \alpha = 0.3$ )	<b>0.7442</b>

Meanwhile, compared to the results on the test set, our model has outstanding improvements in accuracy of about 14% (74.41% on the proposed model compared to 60.33% on the Yolov5 + BERT [9] and 61.76% on the mUNITER-SNLI-MUSTI [10]). This superiority comes from four main reasons. First, we use Local Loss [11] as the loss function of the model. This makes it possible for us to overcome the problem of imbalanced datasets. We tested the data augmentation method and removed NO labels until their numbers equalized. However, this method produces a large dataset and requires ten times more training time. Besides, its accuracy has not improved at all. Second, we use a feature selection unit, which allows us to help the model focus on what is needed and reduces the model’s confusion when predicting. Third, we apply the ensemble model to take advantage of the good points of each model. At the final voting step, we use a linear layer as a soft voting method instead of allocating equal attention to the prediction results of the two individual models. This way, we can take advantage of all two models without being wholly affected by each model’s adverse effects or suboptimal corners. Finally, we use BERT’s multilingual text encoding model, which allows us to learn better text embeddings.

Based on the achieved results described in Tables 1 and 2, we realize that choosing a combination of Ensemble learning methods, Focal loss, and Image-guided Feature Extractor can help the model achieve certain successes. However, choosing a suboptimal set of coefficients (in this setting  $\gamma$  and  $\alpha$  of focal loss) and overusing dropout layers can degrade model performance.

## 5. Ablation Study

In addition to models incapable of capturing imbalanced data, we find that there is no absolute superiority in any specific model. However, models that used an Image-guided Feature Extractor consistently recorded better results than baseline models that did not use proposed units.

**Table 2**

Performance of different methods for data imbalance problem on test set.

Model	Negative Samples			Positive Samples			avg
	precision	recall	F1-score	precision	recall	F1-score	
ViT+SwinTransformer	0.6876	1.0000	0.8149	0.0000	0.0000	0.0000	0.6876
Resnet+SwinTransformer	0.7538	0.8766	0.8106	0.5767	0.3701	0.4508	0.7183
ViT+Resnet+SwinTransformer	0.7598	0.8998	0.8239	0.6291	0.3740	0.4691	0.7355
ViT+Resnet	0.7624	0.9070	0.8284	0.6486	0.3780	0.4776	0.7417
ViT, IGFE	0.7147	0.8784	0.7881	0.4603	0.2283	0.3053	0.6753
Resnet, IGFE	0.7515	0.8819	0.8115	0.5796	0.3583	0.4428	0.7183
SwinTransformer, IGFE	0.6876	1.0000	0.8149	0.0000	0.0000	0.0000	0.6876
ViT+SwinTransformer, IGFE	0.6876	1.0000	0.8149	0.0000	0.0000	0.0000	0.6876
Resnet+SwinTransformer, IGFE	<b>0.7626</b>	0.8390	0.7990	0.5455	<b>0.4252</b>	<b>0.4779</b>	0.7097
ViT+Resnet+SwinTransformer, IGFE	0.7557	0.8909	0.8177	0.6039	0.3661	0.4559	0.7269
ViT+Resnet, IGFE	0.7600	<b>0.9177</b>	<b>0.8314</b>	<b>0.6667</b>	0.3622	0.4694	<b>0.7442</b>

Furthermore, in terms of ensemble learning, we can partly predict the results of the combination model based on its components. Specifically, we use combined variants of Vision Transformer, Resnet-34, and SwinTransformer in this study. If evaluated on each model, we can see that Resnet and SwinTransformer are models for best and worst results. While SwinTransformer has demonstrated excellent results in various tasks [12], it falls short in this particular challenge. Its performance does not reach its full potential in this task, resulting in model combinations built on the SwinTransformer foundation consistently underperforming compared to variants with the same number of sub-components. Furthermore, the combined model based on all three image encoders gives lower results than the model, including only Vision Transformer and Resnet-34. From this, combining models in ensemble learning does not improve accuracy based on the number of components it covers. Instead, it requires more testing and experience to choose suitable combinations. Besides, through testing, we found that Focal Loss’s gamma and alpha values are 2 and 0.3, respectively, which will help the model achieve the best accuracy on the test set and maintain the necessary simplicity.

## 6. Discussion and Outlook

We have proposed a methodology for feature encoding in text-image integration, utilizing the Image-guided Feature Extractor. The incorporation of this component empowers the model to concentrate its attention on discerned objects within the input image. The model has demonstrated noteworthy efficacy in the task of predicting the congruence of olfactory experiences between an image and a corresponding text passage, assessed on the MUSTI dataset. Nonetheless, the dissimilarity in embedding spaces between images and text persists as an unresolved challenge in the meticulous selection of congruent features.

## 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] Y. He, S. Xiang, C. Kang, J. Wang, C. Pan, Cross-modal retrieval via deep and bidirectional representation learning, *IEEE Transactions on Multimedia* 18 (2016) 1363–1377. doi:10.1109/TMM.2016.2558463.
- [3] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.
- [4] A. Baldrati, M. Bertini, T. Uricchio, A. del Bimbo, Composed image retrieval using contrastive learning and task-oriented clip-based features, 2023. arXiv:2308.11485.
- [5] H. Dong, Z. Wang, Q. Qiu, G. Sapiro, Using text to teach image retrieval, 2020. arXiv:2011.09928.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, *ArXiv abs/2010.11929* (2020). URL: <https://api.semanticscholar.org/CorpusID:225039882>.
- [7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.
- [8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
- [9] 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>.
- [10] 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>.
- [11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. arXiv:1708.02002.
- [12] D. C. Bui, T. V. Le, B. H. Ngo, C2t-net: Channel-aware cross-fused transformer-style networks for pedestrian attribute recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, 2024, pp. 351–358.