Overview of ImageCLEFfusion 2023 Task - Testing Ensembling Methods in Diverse Scenarios

Liviu-Daniel Ştefan[1], Mihai Gabriel Constantin[1], Mihai Dogariu[1] and Bogdan Ionescu[1]

[1]*AI Multimedia Lab, Politehnica University of Bucharest, Romania*

Abstract

This paper presents a comprehensive overview of the second edition of the ImageCLEFfusion task, held in 2023. The primary goal of this endeavor is to facilitate the advancement of late fusion or ensembling methodologies, which possess the capability to leverage prediction outcomes derived from pre-computed inducers to generate superior and enhanced prediction outputs. The present iteration of this task encompasses three distinct challenges: the continuation of the previous year's regression challenge utilizing media interestingness data, where performance is measured via the mAP at 10 metric; the continuation of the retrieval challenge involving image search result diversification data, where performance is measured via the F1-score and Cluster Recall at 20; and the addition of a new multi-label classification task focused on concepts detection in medical data, where performance is measured via the F1-score. Participants were provided with a predetermined set of pre-computed inducers and were strictly prohibited from incorporating external inducers during the competition. This ensured a fair and standardized playing field for all participants. A total of 23 runs were received and the analysis of the proposed methods shows diversity among them ranging from machine learning approaches that join the inducer predictions to ensemble schemes that learn the results of other methods.

Keywords

Late fusion, Ensembling, Fusion benchmarking, Visual interestingness prediction, Image search results diversification, Caption detection

1. Introduction

The fusion task, part of ImageCLEF [1, 2], was first proposed in 2022 [3] comprising of two subtasks: a regression challenge utilizing media interestingness data (ImageCLEFfusion-int) and a retrieval challenge involving image search result diversification data (ImageCLEFfusion-div). In 2023 [2], both subtasks, ImageCLEFfusion-int and ImageCLEFfusion-div, were running again with the addition of a multi-label classification task focused on concepts detection in medical data (ImageCLEFfusion-cap). These type of tasks typically exhibit inferior performance in end-to-end systems when juxtaposed with conventional computer vision tasks. This phenomenon is frequently ascribed to their intrinsic subjectivity and multi-modality, compounded by challenges

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.
liviu-daniel.stefan@upb.ro (L. Ştefan); mihai.constantin84@upb.ro (M. G. Constantin); mihai.dogariu@upb.ro (M. Dogariu); bogdan.ionescu@upb.ro (B. Ionescu)
https://lstefan.aimultimedialab.ro/ (L. Ştefan)
0000-0001-9174-3923 (L. Ştefan); 0000-0002-2312-6672 (M. G. Constantin); 0000-0002-8189-8566 (M. Dogariu); 0000-0002-8189-8566 (B. Ionescu)

associated with establishing dependable ground-truth annotations [4, 5]. To address these limitations, researchers have turned to late fusion or ensembling systems as a primary approach to enhance model performance. These systems involve the integration of multiple individual prediction systems, referred to as inducers, through fusion schemes.

Given these factors, the participants in this task are faced with several challenges that necessitate exploration. These challenges include *diversity*, which pertains to a collection of classifiers that generate varying predictions for the same instance; *voting mechanism*, which governs the utilization of individual outputs from the base models during prediction; *dependency*, which refers to the influence of a base model on the construction of the subsequent model in the fusion chain; *cardinality*, which denotes the number of individual base models composing the ensemble—a delicate balance must be struck, as incorporating too many models may diminish diversity within the fusion; and finally, the *learning mode* of the base models, which represents the characteristic that enables the classifiers to effectively adapt to new, previously unseen data while retaining previously acquired knowledge.

This paper presents an overview of the 2023 ImageCLEFfusion task including the data creation in Section 2, the evaluation methodology in Section 3, and the task and participation in Section 4. The results are described in Section 5, followed by conclusion in Sections 6.

2. Data description

The ImageCLEFfusion framework encompasses three distinct tasks, each utilizing different datasets and associated challenges:

ImageCLEFfusion–int: This task focuses on the Interestingness10k dataset [5]. Specifically, it utilizes image-based prediction data derived from the 2017 MediaEval Predicting Media Interestingness task [6]. The task provides prediction outputs from 29 systems that participated in the benchmarking task. To facilitate the training and evaluation of fusion systems, the available data is divided into 1,877 samples for training and 558 samples for testing.

ImageCLEFfusion–div: This task relies on the Retrieving Diverse Social Images dataset [7], specifically targeting the DIV150Multi challenge [8]. The task provides retrieval outputs from 56 systems, which are further divided into 60 queries for the training data and 63 queries for the testing data.

ImageCLEFfusion–cap: This task is derived from the ImageCLEF Medical Caption Task [9]. It involves the extraction of multi-label outputs from 84 inducers. The data used for this task consists of 6,101 images for the development set and 1,500 images for the testing set.

For the training sets, we provide a comprehensive package comprising ground truth data, inducer prediction outputs, detailed inducer performance metrics, and the requisite scripts for metric computation. Conversely, the testing sets solely include the inducer prediction outputs. The characteristics of the datasets used in these tasks are presented in Table 1. Participants have the freedom to generate their own validation sets by partitioning the training set according to their specific needs. However, to ensure a fair and reasonable selection of proposed fusion methods, participants are limited to a maximum of 10 runs for each of the three tasks.

Table 1
Data composition for the ImageCLEFfusion-int, ImageCLEFfusion-div tasks and ImageCLEFfusion-cap.

Task	Training set	Testing set	No. inducers
ImageCLEFfusion-int	1,877 images	558 images	29
ImageCLEFfusion-div	60 queries	63 queries	56
ImageCLEFfusion-cap	6,101 images	1,500 images	84

3. Evaluation Methodology

Participants were required to devise late fusion learning strategies based on the outputs of the inducers associated with the media samples for each of the subtasks. The evaluation of the participants' submissions was conducted using the Mean Average Precision at 10 (mAP@10) metric for the ImageCLEFfusion–int task, F1 at 20 (F1@20) and Cluster Recall at 20 (Cluster Recall@20) metrics for the ImageCLEFfusion-div task, and the F1 metric for the ImageCLEFfusion–cap task. The aforementioned metrics align with the evaluation measures employed for the individual datasets pertaining to each of the three tasks. Participants were encouraged to submit their solutions for all three tasks.

4. Participation

A total of 12 teams completed their registration for ImageCLEFfusion, demonstrating a strong interest in the competition. Among these teams, two successfully submitted their runs and completed the competition by submitting detailed working notes describing their methods. In terms of the interestingness task, both teams collectively submitted 13 runs, while one team submitted a total of 10 runs for the diversification task. No runs were recorded for the ImageCLEFfusion–cap task. For a comprehensive overview of the participating teams, please refer to Table 2.

Table 2
Groups that participated with runs in the ImageCLEFfusion tasks. We present the institutions represented by these teams, the number of runs for ImageCLEFfusion-int, ImageCLEFfusion-div, and ImageCLEFfusion-cap, as well as references to their papers.

Team Name	Institutions	Runs int	Runs div	Runs cap	Paper
CS_Morgan [10]	Computer Science Department, Morgan State University, Baltimore, Maryland, US	10	10	0	yes
SSN CSE-ML [11]	Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India	3	0	0	yes

5. Results

5.1. ImageCLEFfusion-int task

A total of 13 runs were submitted by two teams for the ImageCLEFfusion-int task. The highest achieved performance in terms of MAP@10 value was 0.1331, indicating a significant improvement of 40.65% compared to the baseline value of 0.0946. Despite the reduced number of participants compared to the previous year, the participating team achieved a performance that surpassed the majority of the participants in the previous year, but still under the state-of-the-art result of the last year achieved by [12]. The results for the participating teams for the ImageCLEFfusion-int task are presented in Tables 3.

SSN CSE-ML: The SSN CSE-ML team's most successful run attained a mAP@10 score of 0.1331, establishing itself as the highest-scoring submission among the participating teams in this year's competition for this particular subtask. Balasundaram et al. [11] utilized an ensemble learning model approach based on a Voting Classifier that leverages XGBoost, decision trees, and K-nearest neighbors algorithms, and uses grid search to find the best hyperparameters for each classifier, and the optimal voting scheme and weights for the Voting Classifier.

CS_Morgan: The best performing run from the CS_Morgan team achieved a mAP@10 of 0.1287. For this approach, Emon and Rahman [10] utilized an ensemble of decision trees trained sequentially, with each tree using the predictions from the previous tree to calculate residual errors. A shrinkage technique is applied to reduce the ensemble's impact after each tree's prediction. The ensemble's final predictions are obtained by averaging the regression results. Additionally, the predictions undergo scaling through min-max normalization.

Table 3
Participation in the ImageCLEF-int 2023 task: the best score from all runs for each team. We also included a baseline that consists of the average performance of all the provided inducers.

Team	*#Runs*	*mAP@10*
SSN CSE-ML	10	0.1331
CS_Morgan	3	0,1287
baseline	-	0.0946

5.2. ImageCLEFfusion-div task

A single team submitted a total of 10 runs for the ImageCLEFfusion-div task. The highest performance achieved by the team resulted in an F1@20 score of 0.5708, indicating a 7.4% improvement compared to the baseline value of 0.5313. Additionally, for the secondary metric CR@20, the corresponding system exhibited an improvement of 8.45%. The results for the participating teams in the ImageCLEFfusion-div task can be found in Tables 4.

SSN CSE-ML: The best performing run from the SSN CSE-ML team achieved an F1@20 score of 0.5708 and an CR@20 score of 0.449 using the same model construction as for the ImageCLEFfusion-int task, i.e., bulding an ensemble model based on three classifier models (XGBoost, decision tree, and K-nearest neighbors), and finally creating a Voting Classifier based on the best combination of voting scheme and weights obtained through a grid search.

Table 4
Participation in the ImageCLEF-div 2023 task: the best score from all runs for each team. We also included a baseline that consists of the average performance of all the provided inducers.

Team	*#Runs*	*F1@20*	*CR@20*
SSN CSE-ML	10	0.5708	0.449
baseline	-	0.5313	0.414

6. Conclusions

The second edition of the ImageCLEFfusion task garnered submissions from a total of two teams. The participants were presented with three tasks: the continuation of the previous year's regression challenge, which involved media interestingness data; the continuation of the retrieval challenge, which focused on image search result diversification data; and the addition of a new multi-label classification task centered around concepts detection in medical data. In total, the teams submitted 23 runs, with 13 runs for media interestingness and 10 runs for diversification. Unfortunately, no runs were recorded for the concept detection task. Despite the reduced number of participants compared to the previous year, with only two teams submitting runs for two out of the three presented tasks, the participating teams achieved commendable performance that surpassed the majority of the participants from the previous year. However, their performance still fell short of the state-of-the-art result achieved in the previous year.

7. Acknowledgments

This work is supported under the H2020 AI4Media "A European Excellence Centre for Media, Society and Democracy" project, contract #951911.

References

[1] B. Ionescu, H. Müller, R. Péteri, J. Rückert, A. Ben Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the ImageCLEF 2022: Multimedia Retrieval in Medical, Social Media and Nature Applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.

[2] B. Ionescu, H. Müller, A.-M. Drăgulinescu, W. wai Yim, A. B. Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A.-G. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L.-D. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, social

media, and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science (LNCS), Thessaloniki, Greece, 2023.

[3] L.-D. Ştefan, M. G. Constantin, M. Dogariu, B. Ionescu, Overview of imagecleffusion 2022 task-ensembling methods for media interestingness prediction and result diversification, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bologna, Italy, 2022.

[4] M. G. Constantin, L. D. Stefan, B. Ionescu, C.-H. Demarty, M. Sjoberg, M. Schedl, G. Gravier, Affect in multimedia: benchmarking violent scenes detection, IEEE Transactions on Affective Computing (2020).

[5] M. G. Constantin, L.-D. Ştefan, B. Ionescu, N. Q. Duong, C.-H. Demarty, M. Sjöberg, Visual interestingness prediction: a benchmark framework and literature review, International Journal of Computer Vision 129 (2021) 1526–1550.

[6] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, M. Gygli, N. Duong, Mediaeval 2017 predicting media interestingness task, in: MediaEval workshop, 2017.

[7] B. Ionescu, M. Rohm, B. Boteanu, A. L. Gînscă, M. Lupu, H. Müller, Benchmarking image retrieval diversification techniques for social media, IEEE Transactions on Multimedia 23 (2020) 677–691.

[8] B. Ionescu, A. L. Gînscă, B. Boteanu, M. Lupu, A. Popescu, H. Müller, Div150multi: a social image retrieval result diversification dataset with multi-topic queries, in: Proceedings of the 7th international conference on multimedia systems, 2016, pp. 1–6.

[9] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[10] I. S. Emon, M. Rahman, Media interestingness prediction in imagecleffusion 2023 with dense architecture-based ensemble & scaled gradient boosting regressor model, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[11] B. Prabavathy, G. G. Sai, N. Kishore, M. Olirva, A. M. Vaibhav, N. S. Murali, P. S. Harshith, Efficient fusion techniques for result diversification and image interestingness tasks, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[12] M. G. Constantin, L.-D. Ştefan, M. Dogariu, B. Ionescu, Ai multimedia lab at imagecleffusion 2022: Deepfusion methods for ensembling in diverse scenarios, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bologna, Italy, 2022.