

Overview of ImageCLEFfusion 2022 Task – Ensembling Methods for Media Interestingness Prediction and Result Diversification

Liviu-Daniel Ștefan¹, Mihai Gabriel Constantin¹, Mihai Dogariu¹ and Bogdan Ionescu¹

¹AI Multimedia Lab, Politehnica University of Bucharest

Abstract

The 2022 ImageCLEFfusion task is the first edition of this task, targeting the creation of late fusion or ensembling methods in two different scenarios: (i) the prediction of media visual interestingness, and (ii) social media image search results diversification. The objective proposed to participants is to train and test their proposed fusion schemes on a set of pre-computed inducers, without creating or bringing inducers from the outside. The two scenarios correspond to a regression scenario in the case of media interestingness, where performance is measured via the mean average precision at 10 (MAP@10) metric, and to a retrieval scenario in the case of result diversification, where performance is measured via the F1-score and Cluster Recall at 20 (F1@20, CR@20). Overall 6 teams registered for ImageCLEFfusion, 5 of them submitting runs, while only one team submitted runs to both the interestingness and diversification tasks. A total of 39 runs were received, and an analysis of the proposed methods shows a great diversity among them, ranging from statistical weighted approaches, weighted approaches that use learning stages for creating the weights, machine learning approaches that join the inducer predictions like SVM or KNN, deep learning approaches, and even fusion schemes that join the results of other fusion schemes.

Keywords

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

1. Introduction

The current landscape of computer vision tasks seems to be dominated by end-to-end deep neural networks that take a media sample as input and output a prediction by processing the images or videos through the network layers. However, in several domains, the performance of single network systems reaches a plateau where performance increases are marginal, and performances can be considered comparatively low. This phenomenon may impede the adoption of such AI solutions, as companies and users may be dissatisfied with the results.

One of the main methods researchers use to enhance the performance of models is the use of late fusion (or ensembling) systems. These systems use a collection of individual prediction systems, called inducers, and join their results through fusion schemes. The usefulness of such approaches is proven in several scenarios, in traditional computer vision tasks, like action


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ liviu1_daniel.stefan@upb.ro (L. Ștefan); mihai.constantin84@upb.ro (M. G. Constantin); mihai.dogariu@upb.ro (M. Dogariu); bogdan.ionescu@upb.ro (B. Ionescu)

🆔 0000-0001-9174-3923 (L. Ștefan); 0000-0002-2312-6672 (M. G. Constantin); 0000-0002-8189-8566 (M. Dogariu)



© 2022 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)

recognition in videos [1], but more pervasive in tasks related to the human understanding of multimedia data. These types of tasks generally show lower performance for end-to-end systems when compared with traditional computer vision tasks, a phenomenon commonly attributed to their inherent subjectivity and multi-modality, as well as difficulties in creating reliable ground-truth annotations [2, 3]. Examples from the current literature that support this trend show fusion systems achieving top performance in several benchmarking competitions related to the processing of subjective multimedia data, including but not limited to media interestingness [4], memorability [5], and violent scene detection [6].

Given these factors, the ImageCLEFfusion task, currently in its initial edition at ImageCLEF 2022 [7], asks participants to create ensembling schemes that can accurately join the prediction outputs of a pre-computed set of inducers for two scenarios. The first scenario represents a regression task, applied to the media interestingness data associated with the Interestingness10k dataset [3], while the second scenario represents a retrieval task, applied to the results diversification data related to the Retrieving Diverse Social Images dataset [8].

This paper presents an overview of the 2022 ImageCLEFfusion task and is structured as follows. Section 2 presents the data used in this task, while Section 3 shows details with regards to the participation. Results and their analysis are presented in Section 4, and the paper ends with the main conclusions in Section 5.

2. Data Description

As we mentioned, the first edition of the ImageCLEFfusion task presents participants with two different scenarios: (i) ImageCLEFfusion-int, a regression task using media interestingness data, and (ii) ImageCLEFfusion-div, a retrieval task using search result diversification data. Our philosophy in these scenarios is that we want to compare the performance of ensembling engines in similar setups for all participating teams. Therefore, we provide the set of inducers that participants will use, and the addition of external inducers is not allowed. In a general sense, given a set of M media samples $S \in \{s_1, s_2, \dots, s_M\}$, and a set of N computer vision algorithms $A \in \{a_1, a_2, \dots, a_N\}$, we will provide all the prediction outputs, i.e., for each sample s_i , we will provide a set of predictions $y_i = \{s_{i,1}, s_{i,2}, \dots, s_{i,N}\}$. This process is presented in Figure 1. Participants are tasked with creating the ensembling function \mathcal{F} , that is able to join the predictions from individual inducers and create better predictions.

For the ImageCLEFfusion-int task, we provide data extracted from the image prediction task from the 2017 MediaEval Predicting Media Interestingness task [9]. We use the prediction outputs from the 29 systems submitted during MediaEval. We then split the available data into 1,877 images contained in the training set, and 558 images in the testing set. For the ImageCLEFfusion-div task, we provide data extracted from the DIV150 challenge associated with the Retrieving Diverse Social Images dataset [8]. The prediction outputs of 56 systems inducer systems is provided, with 60 retrieval queries included in the training set and 63 queries included in the testing set. These details are presented in Table 1. Participants are free to create the validation sets as they choose, and can do this by splitting the training set according to their individual needs. In order to encourage a careful selection of the proposed fusion methods, participants are only allowed a maximum of 10 runs for each of the two tasks.

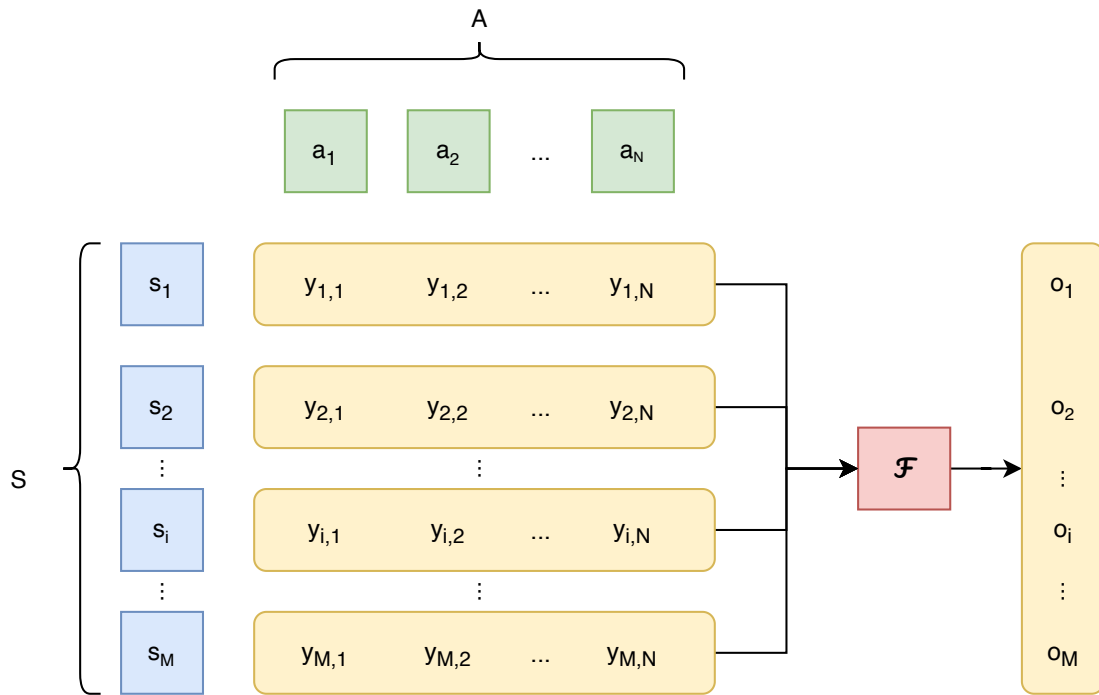


Figure 1: General presentation of ensembling systems. Samples are represented with blue color, inducer algorithms with green, prediction outputs with yellow and the ensembling function with red.

Table 1

Data composition for the ImageCLEFfusion-int and ImageCLEFfusion-div tasks.

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

Evaluation is carried out using mean average precision at 10 (MAP@10) for the interestingness task and F1-score as the primary metric and Cluster Recall as the secondary one, at 20 (F1@20 and CR@20) for the diversification task. These metrics correspond to the metrics used on the respective datasets for each of the two tasks. Both these metrics place greater importance to selecting and presenting the most appropriate media samples to users, 10 in the case of interestingness and 20 in the case of diversification. For ImageCLEFfusion-int we provide the trec_eval tool developed by NIST¹, that can compute several metrics, including MAP@10, while for ImageCLEFfusion-div we provide the div_eval tool, specially developed and designed for the DIV competitions.

For each of the two training sets we provide ground truth data, inducer prediction outputs, inducer performance with regards to the main metrics, and scripts necessary for performance calculation. On the other hand, for the testing sets we only provide the inducer prediction outputs.

¹https://trec.nist.gov/trec_eval/

3. Participation

Participation is satisfactory for the first edition of this task. Overall, 7 groups completed their registration to ImageCLEFfusion, 5 submitted runs, and 4 completed the competition by also submitting working notes describing their methods. For the interestingness task, 3 teams submitted a total of 14 runs, while 3 teams submitted a total of 25 runs for the diversification task. Only one of these teams chose to participate in both tasks. An overview of the submitting teams is presented in Table 2.

Table 2

Groups that participated with runs to the ImageCLEFfusion tasks. We present the institutions represented by these teams, the number of runs for interestingness and diversification, as well as references to their paper, where submitted.

Team Name	Institutions	Runs int	Runs div	Paper
AIMultimediaLab [10]	AI Multimedia Lab, University “Politehnica” of Bucharest	5	5	yes
klssncse [11]	Sri Sivasubramaniya Nadar College of Engineering, Chennai	0	10	yes
shreya_sriram [12]	Sri Sivasubramaniya Nadar College of Engineering, Chennai	0	10	yes
ssn_it	-	1	0	no
UECORK [13]	University of Engineering and Technology, Peshawar Munster Technological University, Cork	8	0	yes

4. Results

The results for the participating teams for interestingness and diversification are presented in Tables 3 and 4. In both cases we present the results of the participating team in comparison with baseline runs that consist of the average performance of all the provided inducers.

4.1. Results for the interestingness task

Three teams submitted 14 runs in total for the ImageCLEFfusion-int task, with the highest performance being a MAP@10 value of 0.2192, representing an improvement of 131% over the baseline of 0.0946. It is interesting to note that all teams scored runs above the baseline.

AIMultimediaLab The best performing run from the AIMultimediaLab team achieved a MAP@10 score of 0.2192 [10]. This team proposed two types of runs. While the first type is based on a simple weighted approach, where weights are determined through a grid-search approach, the second one is based on DeepFusion [14] DNN structures, that use Dense, Attention, Convolutional and Cross-Space-Fusion approaches. The best performing method from this team uses the DeepFusion-CSF model.

ssn_it The best performing run from the ssn_it team achieved a MAP@10 score of 0.1106. Unfortunately, no working notes paper was submitted for this run.

UECORK The best performing run from UECORK attained a MAP@10 score of 0.1097 [13]. The authors propose several approaches to weighted fusion. A baseline weighting method computes the average for all inducer outputs, while complex approaches use different methods for determining and optimizing the weights, like Genetic Algorithms, Nelder Mead algorithm [15], Truncated Newton optimization [16], and Particle Swarm Optimization [17]. The best performing method from this was recorded by two different approaches: PSO and TNC.

Table 3

Results for the ImageCLEFfusion-int task. We present the results only for the best performing proposed system for each team according to the MAP@10 metric. We also compare these results with the a baseline that consists of the average performance of all the provided inducers.

Team Name	Nmb. Runs	Best MAP@10
AIMultimediaLab	5	0.2192
ssn_it	1	0.1106
UECORK	8	0.1097
baseline	-	0.0946

4.2. Results for the diversification task

Three teams were involved in the diversification task too, submitting 25 runs in total. The highest performance shows a F1@20 score of 0.6216, representing an improvement of 17% over the baseline value of 0.5313. For the secondary metric, CR@20, the improvement of the corresponding system is almost equal, at 18%. Again we are happy to note that all teams presented systems that score above the baseline.

AIMultimediaLab The best performing run from the AIMultimediaLab team achieved a F1@20 score of 0.6216, and a CR@20 score of 0.4916 [10]. The same approaches as presented in Section 4.1 were used for the diversification task as well. For the diversification task, the best performing method from this team used the DeepFusion-Convolutional approach.

klsscncse The best performing run from the klsscncse team shows a F1@20 score of 0.5634 and a CR@20 score of 0.4414 [11]. The authors analyzed fusion models based on KNN Regressors, Classification and Regression Trees [18] and SVR [19]. All the submission from this team are represented by CART approaches, as these provided the best results in the preliminary studies.

shreya_sriram The best performing run from the shreya_sriram team shows a F1@20 score of 0.5604 and a CR@20 performance of 0.4373. The paper studies several methods, based on neural networks, namely MLP regressor, Ridge regressor with Grid Search

and KerasRegressor [20]. The results of all these fusion models are combined by a voting regressor, thus obtaining a meta-estimator that uses late fusion as inputs. All the submissions for this team are represented by different setups for the voting regressor.

Table 4

Results for the ImageCLEFfusion-div task. We present the results only for the best performing proposed system for each team according to the F1@20 metric, as well as the corresponding CR@20 metric. We also compare these results with the a baseline that consists of the average performance of all the provided inducers.

Team Name	Nmb. Runs	Best F1@20	CR@20
AIMultimediaLab	5	0.6216	0.4916
klsscse	10	0.5634	0.4414
shreya_sriram	10	0.5604	0.4373
baseline	-	0.5313	0.4140

5. Conclusions

This first edition of the ImageCLEFfusion task attracted a total of 5 teams that submitted runs, with 4 of them completing their submissions by creating a working notes paper. Two tasks were proposed to the participants, a regression-based task that uses media interestingness data and a retrieval task that uses search result diversification data. In total, 39 runs were submitted by the teams, 14 for media interestingness and 25 for diversification. Only one group chose to participate in both tasks.

We compared the submitted runs against a baseline composed of the average performance of all the inducers in the testing set. All the participant teams show performances above this baseline. For the interestingness task, the best result is a MAP@10 score of 0.2192, representing an improvement of 131% over the baseline. On the other hand, for the diversity task, the improvement is lower, 17% for the F1@20 metric and 18% for the CR@20 metric, corresponding to performances of 0.6216 and 0.4916, respectively. Thus, while the results for diversity are higher, from the percentual improvement standpoint, interestingness represents a higher success. We consider that this difference may be the result of the complexity of the inducer output data – interestingness data has a lower complexity and, therefore, perhaps easier to handle by the proposed methods. Also, it may be possible that the higher inducer performances for the diversification tasks leave little room for improvement compared with the inducers associated with the interestingness task.

Regarding the fusion methods proposed by the participants, we are happy to report a high degree of diversity among them. Proposed fusion schemes include simple statistical weighted approaches, approaches that use learning methods for creating the weights, algorithms that utilize both traditional (kNN, SVM) and deep learning (DeepFusion, KerasRegressor) models for combining the inducer prediction, and even a method that uses a voting regressor for combining the outputs of several other fusion schemes. This is indeed encouraging as we are looking forward to developments in future editions of the ImageCLEFfusion task.

Future editions of this task must, first of all, follow the same two datasets with the purpose of monitoring if and how the performances of the proposed systems increased. Also, we plan to add other tasks in the future based on a different machine learning task, whether it is as simple as classification or more complex multi-label or multi-class regressions.

Acknowledgments

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

References

- [1] S. Sudhakaran, S. Escalera, O. Lanz, Gate-shift networks for video action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1102–1111.
- [2] M. G. Constantin, L. D. Stefan, B. Ionescu, C.-H. Demarty, M. Sjöberg, M. Schedl, G. Gravier, Affect in multimedia: benchmarking violent scenes detection, *IEEE Transactions on Affective Computing* (2020).
- [3] 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.
- [4] S. Wang, S. Chen, J. Zhao, Q. Jin, Video interestingness prediction based on ranking model, in: Proceedings of the joint workshop of the 4th workshop on affective social multimedia computing and first multi-modal affective computing of large-scale multimedia data, 2018, pp. 55–61.
- [5] D. Azcona, E. Moreu, F. Hu, T. E. Ward, A. F. Smeaton, Predicting media memorability using ensemble models, *CEUR Workshop Proceedings*, 2020.
- [6] Q. Dai, R.-W. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, Y.-G. Jiang, Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning., in: *MediaEval*, volume 1436, 2015.
- [7] B. Ionescu, H. Müller, R. Peteri, 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.
- [8] 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.
- [9] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, M. Gygli, N. Duong, Mediaeval 2017 predicting media interestingness task, in: *MediaEval workshop*, 2017.

- [10] M. G. Constantin, L.-D. Ștefan, 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.
- [11] L. Kalinathan, P. Balsundaram, Y. Munees, S. S, S. Mr, S. Ramachandran, S. Sriram, R. Venkatakrishnan, A fusion approach for web search result diversification using machine learning algorithms, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.
- [12] S. Sriram, P. Balasundaram, L. Kalinathan, Ensembled approach for web search result diversification using neural networks, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.
- [13] M. Shoukat, K. Ahmad, N. Said, N. Ahmad, H. Uzzaman, K. Ahmad, A late fusion framework with multiple optimization methods for media interestingness, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.
- [14] M. G. Constantin, L.-D. Ștefan, B. Ionescu, Deepfusion: Deep ensembles for domain independent system fusion, in: International Conference on Multimedia Modeling, Springer, 2021, pp. 240–252.
- [15] S. Singer, J. Nelder, Nelder-mead algorithm, Scholarpedia 4 (2009) 2928.
- [16] J. Martens, Deep learning via hessian-free optimization., in: ICML, volume 27, 2010, pp. 735–742.
- [17] R. Poli, J. Kennedy, T. Blackwell, Particle swarm optimization, Swarm intelligence 1 (2007) 33–57.
- [18] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, Routledge, 2017.
- [19] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Advances in neural information processing systems 9 (1996).
- [20] F. Chollet, et al., Keras, 2015. URL: <https://github.com/fchollet/keras>.