

# AIMultimediaLab at ImageCLEFfusion 2022: DeepFusion Methods for Ensembling in Diverse Scenarios

Mihai Gabriel Constantin<sup>1</sup>, Liviu-Daniel Ștefan<sup>1</sup>, Mihai Dogariu<sup>1</sup> and Bogdan Ionescu<sup>1</sup>

<sup>1</sup>AI Multimedia Lab, University “Politehnica” of Bucharest

## Abstract

The ImageCLEFfusion task addresses the challenging task of creating ensembling schemes and algorithms for fusing a large set of inducers in two particular scenarios: a regression scenario where the ground truth data consists of media interestingness annotations and a search result diversification scenario. We present our team’s approach for these two scenarios, consisting of a simple statistical baseline followed by the use of DeepFusion architectures and the way these architectures must be adapted for each scenario. The DeepFusion methods tested for media interestingness and result diversification are represented by deep neural networks consisting of dense, attention, convolutional, and Cross-Space-Fusion layers.

## Keywords

DeepFusion, deep ensembles, ImageCLEFfusion, media interestingness, result diversification

## 1. Introduction

While the development of deep neural networks has greatly improved system performance for many popular tasks, such as image recognition [1], there still are some domains where the performances of single end-to-end systems are comparatively low. This can happen in various domains, however, those related to the human perception of multimedia data represent one of the areas widely known for data that is harder to classify and predict with the help of computer vision methods. This is presented throughout the literature with examples like media interestingness [2], violence detection [3], and emotional content classification [4]. Many reasons are given for this trend, including the inherent subjectivity of the concepts, their multimodality, the lack of distinctive or unique features that define and influence them, and the complexity and novelty of these concepts.


Ensemble systems, or late fusion systems, are defined as systems where, given a set of end-to-end systems called inducers and their decisions on the training set, an ensembling engine is trained or programmed to join the decisions of the individual inducers in order to obtain a better set of predictions. These systems represent one of the most important approaches for increasing overall system performance and have been proven useful in many domains, including those related to the human perception of multimedia data [5].


---

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

✉ mihai.constantin84@upb.ro (M. G. Constantin); liviu1\_daniel.stefan@upb.ro (L. Ștefan); mihai.dogariu@upb.ro (M. Dogariu); bogdan.ionescu@upb.ro (B. Ionescu)

ORCID 0000-0002-2312-6672 (M. G. Constantin); 0000-0001-9174-3923 (L. Ștefan); 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)

Given these aspects, the ImageCLEFfusion 2022 task [6], which is part of the 2022 ImageCLEF evaluation campaign [7], proposes two different tasks related to the subjective perception of multimedia data, namely media interestingness prediction and search result diversification. Participants are given a set of pre-defined inducers and their prediction outputs and are tasked with developing and adapting ensembling methods that are able to more accurately predict the ground truth for these two subjective concepts. This paper describes the methods employed by the AI Multimedia Lab team for this task. We propose starting from a simple statistical weighted method [8] for combining the inducer predictions and then using the more novel DeepFusion approach [9], which has previously shown very good improvements over the performances of the inducers included in the system.

The rest of this work is structured as follows. Section 2 presents the datasets used for the experiments. The proposed methods are presented in Section 3, while their results on the ImageCLEFfusion tasks are presented in Section 4. Finally, the paper closes with Section 5, where we present the main conclusions of our experiments.

## 2. The ImageCLEFfusion 2022 task

ImageCLEFfusion consists of two different tasks related to the human perception of multimedia data, namely the prediction of media interestingness and search result diversification. The organizers provide a set of pre-defined and pre-computed inducers for these two tasks that correspond to end-to-end systems used for the Interestingness10k [2] dataset during the MediaEval 2017 Predicting Media Interestingness Task<sup>1</sup>, and to the systems used for MediaEval Retrieving Diverse Social Images task<sup>2</sup> [10]. Table 1 shows the main statistics of the two proposed tasks. As the table shows, a high number of inducers is available for both tasks, giving the opportunity of testing ensembling schemes that have an integrated learning component. The performance for interestingness prediction is validated with the MAP@10 metric, while the performance for result diversification is validated with the F1@20 primary metric and with Cluster Recall at 20 (CR@20) as the secondary metric.

	ImageCLEFfusion-int	ImageCLEFfusion-div
No inducers	29	56
Samples devset	1826 samples	104 queries
Samples testset	609 samples	35 queries
Primary metric	MAP@10	F1@20
Secondary metric	-	CR@20

**Table 1**

Main statistics of the ImageCLEFfusion 2022 tasks. ImageCLEFfusion-int represents the Media Interestingness task, while ImageCLEFfusion-div represents the Result Diversification task.

Given the different compositions of the inducer output files associated with these two tasks, different input processing schemes will be employed, which will allow our proposed methods to function and learn in an input-agnostic fashion.

<sup>1</sup><http://www.multimediaeval.org/mediaeval2017/mediainterestingness/>

<sup>2</sup><http://www.multimediaeval.org/mediaeval2017/diverseimages/>

### 3. Proposed Methods

Two methods are proposed for handling these two challenges. The first one consists of a simple approach based on statistical weighted averages, where weights are assigned according to the pre-computed performance of the inducers on the development set. Thus, this statistical approach needs no training phase but only a grid-search type of approach. The second method is represented by the DeepFusion [9] approach, where dense, attention, convolutional, and Cross-Space-Fusion layers are employed, thus creating a set of architectures that are trained on the development set, allowing the networks to learn the way inducers correlate and interact.

#### 3.1. Statistical Approach

The statistical approach is based on creating a simple scheme, where weights are assigned to each inducer based on their pre-computed performance on the development set. After finding the optimal weights on the development set, these weights are applied to the testing set, thus obtaining the final predicted scores. In theory, given a set of inducers  $I = [i_1, i_2, \dots, i_n]$ , and their performances on the development set, arranged in descending order  $P = [p_1, p_2, \dots, p_n]$ , the rank according to their performance  $R = [r_1, r_2, \dots, r_n]$ , a set of weights can be applied to each inducer's prediction  $W = [w_1, w_2, \dots, w_n]$ , where weight  $w_1$  is the largest, as it belongs to the best performing inducer on the development set,  $w_2$  the second largest, as so on.

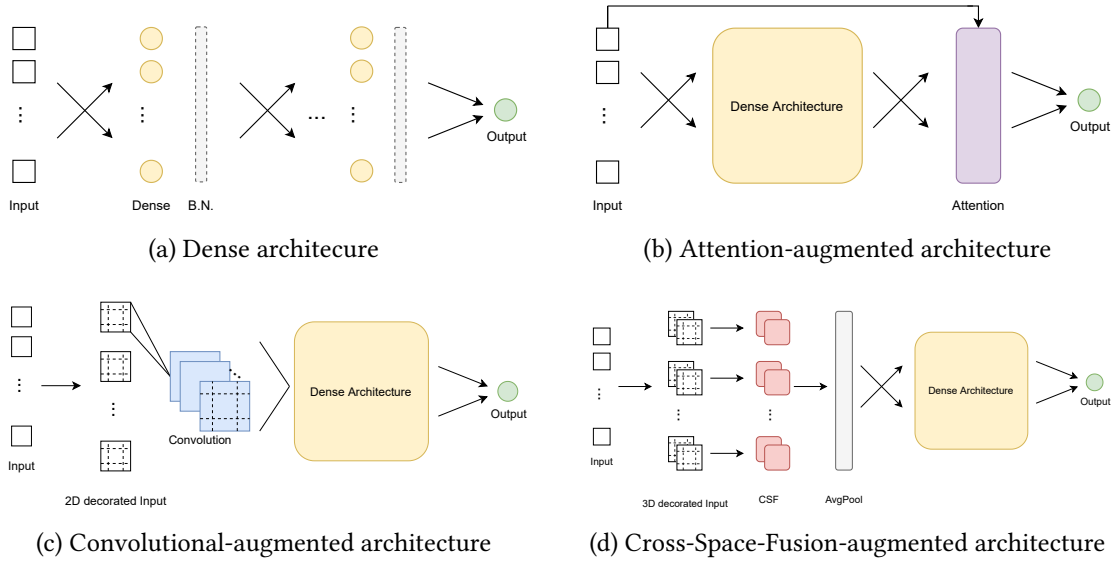
In order to calculate the weights according to the performance of the inducers we used equation 1, where we perform a grid search for parameter  $\epsilon$  taking the following values 0.01, 0.05, 0.1, 0.15, 0.20, 0.25. Following this step each inducer's prediction output is multiplied by the corresponding weight, obtaining a new set of inducers defined by equation 2. Of course, due to the high value of the *epsilon* parameter in some cases the worst performing inducers may end up being neutralized and populated with zeroes. For such cases, those inducers are neglected from the final ensembling scheme. Finally, in order to obtain the predictions on the testing set, the weighted prediction values of the inducers that have not been zeroed by the  $\epsilon$  parameter are averaged for each sample.

$$w_n = 1 - r_n \cdot \epsilon \quad (1)$$

$$\hat{I} = [i_1 \cdot w_1, i_2 \cdot w_2, \dots, i_n \cdot w_n] \quad (2)$$

#### 3.2. DeepFusion

The DeepFusion approach [9] has previously shown very good results in improving the performance of the input inducers. We propose to adapt and test these architectures for the two given tasks. The general setup of the architecture is presented in Figure 1, describing the four types of layers that compose it: dense, attention, convolutional, and Cross-Space-Fusion layers. Unlike the statistical approach, no inducers are zeroed out, and all of them are taken into account. It thus falls into the network's learning tasks to reduce as much as possible the inducers that actually harm the final results. The training process uses 50 epoch for each of the four proposed architectures.



**Figure 1:** The DeepFusion architecture. We present the four main types of architectures associated with this method, namely Dense, Attention, Convolutional and Cross-Space-Fusion.

The Dense approach shown in Figure 1a consists of a series of fully connected or dense neural network layers, with the traditional role of creating a simple MLP-like architecture that can accurately classify the given testset samples, based on the learning process that is carried out on the devset. We created several variants of this approach, featuring a varying number of dense layers (namely 5, 7, 10, 12, and 15), a varying number of neurons per layer (namely 25, 50, 100, 500, 1000), and activating or deactivating the batch normalization layers for this architecture.

The Attention architecture shown in Figure 1b starts its training from the best performing Dense architecture and augments it with soft attention layers that have the role of controlling the inducers according to the learning that this architecture performs on the development set. The soft attention vector takes the values  $attn_i \in [0, 1]$ , thus having the option to vary the weights corresponding to inducers but also to completely negate them.

For the following two architectures, the Convolutional and Cross-Space-Fusion, input decoration schemes as presented in [11, 9] are used. These decoration schemes allow the input to the networks to gain spatial correlations by decorating the prediction of each inducer with predictions from correlated inducers and the correlation scores. The correlation scores between inducers are calculated using the provided primary metrics for the two tasks.

The Convolutional architecture, shown in Figure 1c uses the two-dimensional decorated input and applies convolutional filters to this input in order to compute and learn the way correlated inducers may interact. The output of the convolutional filters is then pooled through average pooling operation, thus allowing it to be sent to the Dense architecture previously discovered. A similar operation is carried out for the Cross-Space-Fusion approach, which uses a three-dimensional decoration scheme. This time the decorated input is fed to a series of CSF filters, each one being different for each inducer. This allows the network to learn more than just one type of correlation processing scheme, learning one for each inducer.

	Media Interestingness		Result Diversification		
	Run ID - int	MAP@10	Run ID - div	F1@20	CR@20
devset-inducer-low	-	0.022	-	0.4262	0.3103
devset-inducer-avg	-	0.0946	-	0.5313	0.4140
devset-inducer-high	-	0.1465	-	0.6092	0.4823
Statistical	183903	0.1406	183898	0.5971	0.4622
DeepFusion-Dense	183908	0.2191	183902	0.6159	0.4862
DeepFusion-Attention	183915	0.2116	183901	0.6182	0.4879
DeepFusion-Convolutional	183917	0.2103	<b>183900</b>	<b>0.6216</b>	<b>0.4916</b>
DeepFusion-CSF	<b>183921</b>	<b>0.2192</b>	183899	0.6188	0.4889

**Table 2**

Main statistics of the results obtained on the Interestingness and Diversification tasks. The proposed approaches are compared with the performance statistics for the development set.

### 3.3. Input Normalization

We quickly learned, in the training phase, that certain operations are needed for each individual task in order to optimize the performance of the networks and help their learning process. While for the interestingness task, this process was quite easy and only consisted of normalizing the predictions of the inducers in a  $[0, 1]$  interval, the problem was more difficult to adapt to the predictions of the diversification task.

Thus, a few observations have to be made about the inducer data from the diversification task. While the data from the interestingness task has a normal spread of values, many inducers from the diversification task start with a value of 1 for the most relevant image for a query and decrease in an exponential manner to the end of the relevance list. We theorize that this causes two possible problems: (i) The values at the top of the list are very high, thus making those images hard to re-rank at lower positions in case this is needed; (ii) The images at the middle, and of course at the end of the relevance lists have scores that are almost identical, making them easy to re-rank during the learning process, even though this may not be needed. We thus decided to recreate the scores associated with the relevance lists using the following custom normalization method. Given the rank  $R = [r_0, r_1, \dots, r_{49}] = [0, 1, \dots, 49]$  predicted by a single inducer for a single query, we create a new set of scores and overwrite the initial scores according to the equation:

$$s_i = 1 - r_i \cdot \omega \quad (3)$$

In this case, we tested several values for the  $\omega$  parameter, namely 0.005, 0.01, 0.05, 0.1. Given that in some of these cases the  $\omega$  parameter may cause some scores to be zeroed out or go below zero, we neutralize those particular values. We performed a series of preliminary tests in order to obtain the best value for this parameter, and in the end we obtained  $\omega = 0.05$ .

## 4. Results and discussion

Results on the testing set are presented in Table 2, where they are compared with the lowest, best and average performance of the inducers on the development set, as they are given by

the task organizers. These three inducer performances should be used as a general baseline for comparison, as they do not reflect the actual performances on the testset. The actual performances of the inducers on the testset are not released yet, as this data will most likely be used in future versions of this competition.

First of all it is interesting to note that, while the statistical approach generally obtains good results, surpassing both the lowest and the average inducer performances, it does not surpass the best highest inducer performance. However, better performances are achieved by the DeepFusion approaches. Thus, even the basic DeepFusion-Dense architecture outperforms the top inducer, by a very large margin in the case of Interestingness (49.55% over the top inducer) and by a smaller margin in the case of Diversification (1.09% over the top inducer). The best results are achieved by two different approaches. For Interestingness, the best performing approach uses the DeepFusion-CSF method, and attains a MAP@10 of 0.2192, representing a 49.62% increase over the top inducer. On the other hand, the best performing approach for Diversification is represented by the DeepFusion-Convolutional approach, with a final F1@20 score of 0.6216, representing a 2.03% increase over the top inducer, and a CR@20 score of 0.4916, representing a 1.92% increase over the top inducer.

The clear observation from these results is the significant gap in performance gains between the two tasks. We theorize that one of the reasons for such a gap may be represented by the data itself. Thus, for Interestingness, the results were very low to begin with, thus making performance gains much easier to achieve. Furthermore, there is a much larger comparative gap between the best and lowest performing inducers in the two tasks. This may have given the networks clearer candidates for exclusion during the learning process for Interestingness when compared with Diversification. Also, it is worth noting that the inducer predictions had to be adapted by using a custom input normalization scheme for the Diversification task. While this scheme did help the networks and increased the results, significantly improved variants of this scheme may be developed in the future.

## 5. Conclusions

In this study we proposed and compared a set of ensembling methods, based on simple statistical weighted approaches and on the DeepFusion deep neural network-based fusion architecture. These approaches are tested and validated on the ImageCLEFfusion 2022 task, featuring a task related to Media Interestingness and one related to Result Diversification. Results on the testing set show a major increase in performance provided by the DeepFusion architecture on the Interestingness task, while a lower increase is provided for the Diversification task. We thus demonstrated that the DeepFusion architectures can be successfully adapted to tasks that are significantly different in their scope and type of approach with regards to data representation.

## Acknowledgments

This work was funded under projects DeepVisionRomania “Identifying People in Video Streams using Silhouette Biometric”, grant 28SOL/2021, UEFISCDI, Solutions Axis, and AI4Media “A

European Excellence Centre for Media, Society and Democracy”, grant 951911, H2020 ICT-48-2020.

## References

- [1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision (IJCV)* 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.
- [2] 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.
- [3] 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).
- [4] E. Dellandréa, L. Chen, Y. Baveye, M. V. Sjöberg, C. Chamaret, et al., The mediaeval 2016 emotional impact of movies task, in: *CEUR Workshop Proceedings*, 2016.
- [5] M. G. Constantin, L.-D. Ştefan, B. Ionescu, Exploring deep fusion ensembling for automatic visual interestingness prediction, in: *Human Perception of Visual Information*, Springer, 2022, pp. 33–58.
- [6] L.-D. Ştefan, M. G. Constantin, M. Dogariu, B. Ionescu, Overview of the 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.
- [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] J. Kittler, M. Hatef, R. P. Duin, J. Matas, On combining classifiers, *IEEE transactions on pattern analysis and machine intelligence* 20 (1998) 226–239.
- [9] M. G. Constantin, L.-D. Ştefan, B. Ionescu, Deepfusion: Deep ensembles for domain independent system fusion, in: *International Conference on Multimedia Modeling*, Springer, 2021, pp. 240–252.
- [10] 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.
- [11] L.-D. Ştefan, M. G. Constantin, B. Ionescu, System fusion with deep ensembles, in: *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 256–260.