A Late Fusion Framework with Multiple Optimization Methods for Media Interestingness

Maria Shoukat[1], Khubaib Ahmad[1], Naina Said[1], Nasir Ahmad[1], Mohammed Hasanuzzaman[2] and Kashif Ahmad[2]

[1]*Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan*
[2]*Department of Computer Science, Munster Technological University, Cork, Ireland*

Abstract

The recent advancement in Multimedia Analytical, Computer Vision (CV), and Artificial Intelligence (AI) algorithms resulted in several interesting tools allowing an automatic analysis and retrieval of multimedia content of users' interests. However, retrieving the content of interest generally involves analysis and extraction of semantic features, such as emotions and interestingness-level. The extraction of such meaningful information is a complex task and generally, the performance of individual algorithms is very low. One way to enhance the performance of the individual algorithms is to combine the predictive capabilities of multiple algorithms using fusion schemes. This allows the individual algorithms to complement each other, leading to improved performance. This paper proposes several fusion methods for the media interestingness score prediction task introduced in CLEF Fusion 2022. The proposed methods include both a naive fusion scheme, where all the inducers are treated equally and a merit-based fusion scheme where multiple weight optimization methods are employed to assign weights to the individual inducers. In total, we used six optimization methods including a Particle Swarm Optimization (PSO), a Genetic Algorithm (GA), Nelder-Mead, Trust Region Constrained (TRC), and Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm (LBFGSA), and Truncated Newton Algorithm (TNA). Overall better results are obtained with PSO and TNA achieving 0.109 mean average precision @10. The task is complex and generally, scores are low. We believe the presented analysis will provide a baseline for future research in the domain.

Keywords

Media Interestingness, Late Fusion, PSO, Genetic Algorithms, Nelder Mead, Trust Region Contrainted optimization

1. Introduction

In the modern world, thanks to social media and other multimedia content sharing platforms, we have access to a huge amount of multimedia content. However, accessing multimedia content of interest generally involves processing, analyzing, and filtering a huge amount of data. Filtering and retrieval of multimedia content of interest require specialized tools to analyze and extract semantic features/meanings from the content [1].

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
marvi1708@hotmail.com (M. Shoukat); imkhubaib1999@gmail.com (K. Ahmad); nainasaid@uetpeshawar.edu.pk (N. Said); n.ahmad@uetpeshawar.edu.pk (N. Ahmad); mohammed.hasanuzzaman@mtu.ie (M. Hasanuzzaman); kashif.ahmad@mtu.ie (K. Ahmad)

Thanks to the recent development in Multimedia Analytics, Computer Vision (CV) and Artificial Intelligence (AI) techniques, different semantic notions, such as sentiments [2], emotions [3], and interestingness-level [4], can be extracted from multimedia content. More recently, Deep Neural Networks (DNNs) have shown tremendous predictive capabilities in various multimedia content analysis tasks. Despite the proven performances, there are several tasks where a single Neural Network is not enough to accurately extract meaningful insights precisely. In order to increase the performance of individual models, researchers have been exploring the so called "fusion" techniques allowing multiple models to complement each others in such complex tasks [5]. The fusion techniques allow to combine multiple models to achieve higher accuracy compared to the individual models. Two popular methods to combine these models are the early fusion and late fusion techniques. In early fusion, the separate raw data is integrated into a unified representation before the learning process. On the other hand, in case of late fusion, fusion is performed at the decision level i.e, the output of different predictors is combined after the learning process to create a new and improved super predictor. The literature has reported the effectiveness of fusion technique in several applications, such as natural disasters analysis [6], event recognition [5], data analytics [7].

This paper is based on one of the tasks in ImageCLEF 2022 [8], which is a benchmark competition for image retrieval tasks. This year ImageCLEF proposes four different tasks. However, the work is based on Image Interestingness CLEFfusion 2022 task [9]. This is a regression task that aims at the prediction of image interestingness score. The task mainly focuses on the fusion of different inducers, whose scores are already provided, to jointly predict the interestingness of visual content. In this work, we propose both a naive fusion, where all inducers are assigned equal weights, and merit-based fusion techniques with optimized weights to combine the scores of the individual inducers for better prediction. For the merit-based fusion, we employ five different techniques to assign weights to the 29 inducers provided in the dataset by the organizers. These methods include evolutionary algorithms, namely Particle Swarm Optimization (PSO) and a Genetic algorithm (GA) based methods, Trust Region Constrained Optimization, Limited-memory Broyden–Fletcher–Goldfarb–Shann (LMBFGS) method, and Truncated Newton Algorithm (TNA) method.

The rest of the paper is organized as follows : Section 2 presents the overview of the related literature. Section 3.2 gives details about different fusion techniques that have been used in this research work. Section 4 provides the experimental results. Finally, Section 5 concludes the work and outlines the future directions.

2. Related Work

Media interestingness prediction, which involves an automatic analysis of multimedia content for the identification of relevant content of users' interest, got great attention from the community over the last few years [10]. It plays a vital role in several applications, such as image retrieval and recommendation and media summarization, etc. In the literature, the topic has been analyzed from two different perspectives including psychological and computational aspects of media interestingness [10]. The first part mainly focuses on psychological studies involving theoretical analysis and reports on human emotions, choices, and interests. For instance,

Silvia et al. [11] linked interestingness level with emotions by providing a detailed analysis and overview of some unusual aesthetic emotions. The computational methods on the other hand involve multimedia analytics, CV, and ML techniques to analyze and extract semantic features from multimedia content for the prediction of interestingness level [12]. Extensive research exploring different aspects of the topic has been carried out in this direction. For instance, Liu et al. [13] analyzed the importance of feature extraction for the task by proposing a multi-view manifold learning framework. To this aim, the authors mapped multi-view data to a single common space by considering cross-view correlation to preserve the geometric structure and interestingness information. Wang et al. [14] discussed other two important aspects of media interestingness namely comparison information, and evaluation metric optimization. Despite sufficient improvement, the performance of most of the algorithms is not good as compared to other computer vision tasks.

As part of the efforts to improve the performances of media interestingness frameworks, a vast majority of the recent works rely on multiple models. To this aim, different fusion techniques have been incorporated to jointly employ multiple models for the task. For instance, Constantin et al. [12] proposed a deep fusion ensemble framework by exploring the potential of several deep networks including dense, attention, convolutional, and cross-space-fusion networks. Similarly, Almedia et al. [15] proposed a late fusion framework employing multiple ranking models trained on multimodal features for media interestingness score prediction.

In this work, we explore the potential of merit-based fusion by combining the predictions of several inducers using different weight optimization methods.

3. Methodology

Figure 1 provides the block diagram of the methodology adopted in this work. There are two main components of the methodology, namely (i) prediction by individual inducers, and (ii) fusion of the score obtained with the individual inducers for joint prediction. Our contribution mainly lies in the fusion part, where we employed several weight selection/optimization techniques to assign weights to the inducers. The scores of the individual inducers are already provided by the task organizers. In the next subsections, we provide a detailed description of all the methods.

3.1. Inducer Scores and Pre-processing

The inducers' scores are provided by the task organizers. In total, the task organizers provided prediction scores of 29 inducers. For the majority of the inducers, the prediction scores were floating numbers between 0 and 1. However, the interestingness scores for some of the inducers were out of range. To combine the scores of the inducers properly, all the values should be in the same range. To this aim, before combing the scores in a late fusion, we normalized the inducers' scores to bring them to the same range.

3.2. Fusion Techniques

In this work, we mainly focused on late fusion techniques where we tried several weight optimization methods to obtain a combination of weights that provides highest interestingness

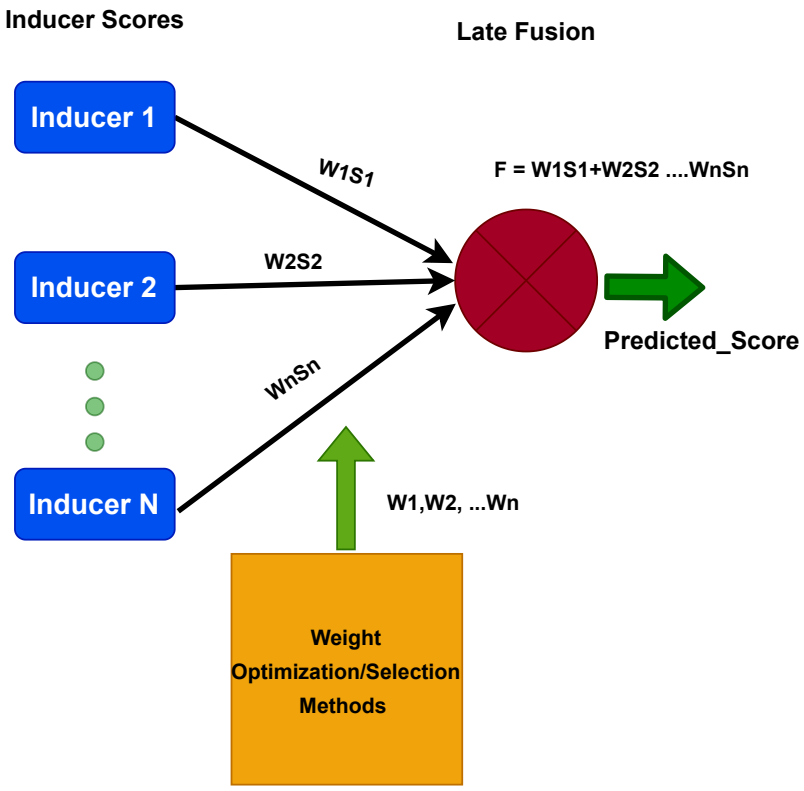

Figure 1: Block diagram of the proposed methodology.

score. To this aim, we picked several methods with proven performances in similar applications [16, 5]. As a baseline, we also considered a naive fusion method where equal weights are assigned to the inducers. Our late fusion method is represented by Equation 3.2.

$$S_c = W_1 S_1 + W_2 S_2 + W_3 S_3 + ... + W_n S_n \tag{1}$$

In the equation, S_c represents the combined interestingness score of different inducers, S_n is the score of the nth model and W_n is the corresponding weight used during the fusion. In our case, n=29. For the baseline, the weight W for all the inducers is the same i.e., $W_1 = W_2 = W_3 .... = W_n$. In case of merit based fusion, optimal weights are assigned to each individual inducer. In the next section, we provide the details of these optimization methods.

3.2.1. Genetic Algorithm

Our choice of Genetic Algorithm (GA) is based on our previous experience in similar applications [16, 17, 5]. GA, which is a meta-heuristic algorithm, is inspired by the natural evaluation process. In the natural evolution process, the fittest individuals of the current generation are firstly identified and then used for re-production in the next generation. A similar approach is adopted in GA-based optimization, where the algorithm searches for optimal value/set of

values minimizing a given function also called fitness function. To this aim, GA requires a training procedure to determine optimal values. The process starts with a randomly selected generation of the population (i.e., a set of values). The fitness of every individual in the current population is evaluated using the fitness function after which individuals are selected for the next generation with a modified genome. The process is repeated until either the maximum number of generations is reached or a respectable fitness value is achieved.

In this work, since we are dealing with a regression problem, the fitness function is based on Mean Squared Error (MSE) as shown in equation 2. Moreover, each possible combination (i.e., set of weights assigned to the 29 inducers) is a potential solution. The goal is to find the set of weights with minimum MSE.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(y_p - y_a)^2 \tag{2}$$

In the above equation, y_p represents the predicted interestingness score while y_a is the ground truth. For the implementation we used a python open source library, namely geneticalgorithm[1]. As we have a total of 29 inducers so we kept the dimensions to 29. Moreover, we used 'real' for the variable type and the variable boundary fixed between 0 and 1.

3.2.2. Particle Swarm Optimzation

The second weight optimization/selection method employed in this work is based on PSO [18]. The optimization method is inspired by the flocking of birds. Unlike GA, PSO does not use mutation and crossover operations rather aims at an improvement to a candidate solution according to a pre-defined criterion, iteratively. There are three main steps involved in the process. These include an (i) evaluation of each candidate solution on the basis of fitness criteria, (ii) updates in personal best and global best values and finally (iii) updating the position and velocity of each particle.

In our case, each combination of the weights (i.e, 29 values to be assigned to the inducers) is a candidate solution. Moreover, the fitness function is based on MSE as shown in 2. For the implementation of the method, we used open source library namely pyswarm[2]. As this algorithm support bounds like GA, we set the lower bound to 0 and the upper bound to be 1. Moreover, the maximum iterations hyperparameter is set to 10000 and the swarm size is kept at 300.

3.2.3. Nelder Mead Algorithm

Nelder Mead algorithm is a heuristic optimization technique and is appropriate for optimization problems where the gradient of the function is either unknown or cannot be reasonably computed. The algorithm can be used for both one dimensional and multi-dimensional optimization problems [19]. The algorithm starts with randomly generated simplex with number of vertices= $n+1$ points for an n-dimensional optimization problem. At every iteration, the algorithm moves the simplex one vertex at a time towards an optimal region in the search space with a goal to

[1] https://pypi.org/project/geneticalgorithm/
[2] https://pyswarms.readthedocs.io/en/latest/

minimize/maximize a certain objective function. At the end, the vertex of the simplex that yields that most optimal values is returned. For our experiments, $n = 29$ and the objective function definition is the same as in 2. For the implementation of the method, we used a Python open source library, namely, SciPy[3]. We set the value of absolute error in xopt between iterations that is acceptable for convergence to 1e-8 and the maximum iterations to 10000.

3.2.4. Trust Region Constrained Optimization Algorithm

Trust Region Constrained method belongs to the family of optimization methods that are based on trust regions. Trust regions-based methods solve optimization problems by defining a region around their current best solutions, where they can approximate the fitness function up to a certain extent. The methods then take a step forward within the region. In contrast to line-based solutions, the step size is determined beforehand of the improvement in the direction. At this stage, the model is considered to be a good representation of the original objective function if a significant decrease in the objective function is observed.

In this work, for the implementation of the method we used scipy library[4]. The Trust Region Constrained algorithm requires initial weight values so for all the 29 inducers same value of 0.0345 is used. For the bounds, we used a lower bound of 0 and upper bound of 1. Moreover, we set the maximum iterations to 10000.

3.2.5. Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm

The Broyden, Fletcher, Goldfarb, and Shanno, or BFGS Algorithm, is a local search optimization algorithm. It falls under the category of a Quasi-Newton optimization methods which deal with the optimization of second order derivative of the objective function. These type of algorithms are suitable for optimization problems where the second order derivatives can not be reasonably quantified. Unlike the first order methods which make use of the first order derivative to find the optimal values of the objective function, these algorithm rely on second order derivatives. For a multivariate function, the second derivatives of all the input variables are maintained in a matrix called Hessian. In order to find the optimal values of the objective function, the BFGS algorithm calculates the inverse of this matrix. This is done by approximating the inverse using gradient thereby eliminating the need for inverse calculation at each step of the algorithm. The size of the Hessian and its inverse is proportional to the number of input parameters to the objective function. For a function with many input parameters, the BFGS then becomes impractical due to very high memory demands. Therefore, a variant of BFGS called Limited BFGS is utilized in this work. This method does not require storing the entire approximation of the inverse matrix.

The definition of the objective function for our experiments is the same as given in 2. For implementation of the method, we used a Python open source library, namely, SciPy[5]. Similar to Trust Region Constrained Optimization algorithm, the Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm also requires initial weight values that are set at

[3]https://scipy.org/
[4]https://scipy.org/
[5]https://scipy.org/

0.0345 for all the 29 inducers. Moreover, the value of absolute error is set to 1e-8 and the maximum iterations to 10000. For the bounds, we set the lower and upper bounds to 0 and 1, respectively.

3.2.6. Truncated Newton Algorithm

The method is also called Hessian-free optimization algorithm and is more suitable for applications involving a large numbers of independent variables [20]. The method uses an iterative process to solve the Newton's equation, which involves finding the roots of a differentiable function, for updating the parameters of the cost function. The term "truncated" refers to the fact that the inner solver is run for a limited number of iterations, which means that the algorithm needs to produce good approximation in limited iterations. The definition of the objective function for our experiments is the same as given in 2. For implementation of the method, we used a Python open source library, namely, SciPy. The Truncated Newton algorithm also requires initial weight values, which are set at 0.0345 for all the 29 inducers. Moreover, we set the value of absolute error in xopt between iterations that are acceptable for convergence to 1e-8 and the maximum iterations to 10000. The lower and upper bounds are set at 0 and 1, respectively.

4. Experiments and Results

4.1. Dataset

The individual inducers' scores, which are provided by the task organizers, are extracted from the Interestingness10k dataset [10]. In total, prediction scores for 2435 images are provided from 29 inducers, each representing the visual interestingness-level for the images. The dataset is provided in two seperate sets namely (i) development set, and (ii) test set. The development set is composed of the scores from all of the inducers for 1877 images while the test set covers 558 images only. In the development set, each data sample for each inducer provides four values including a video ID, image ID, classification score (0 or 1), and predicted interestingness score by the inducer.

4.2. Experimental Results

Table 4.2 provides the official results of the proposed methods in terms of mean average precision at the cutoff 10 (MAP@10). One of the main objectives of the experiments is to evaluate the potential of the different state-of-the-art optimization methods in this application.

As expected, overall lower results are obtained with the baseline method where all the inducers are treated equally by assigning them equal weights. Though there is no significant, difference in scores obtained between the least performing and the baseline, the merit-based fusion scheme seems more promising compared to simply averaging the individual scores. As far as the performances of the merit-based fusion methods are concerned, overall better results are obtained with PSO and TNC methods. One of the key advantages of TNC is its optimization capabilities in dealing with functions involving independent variables. This could be one of

Table 1
Experimental results in terms of mean average precision at the cutoff 10 MAP@10.

Fusion Method	MAP@10
Equal Weights	0.081
Trust-Constr weighted Fusion	0.095
PSO weighted Fusion	**0.109**
GA weighted Fusion	0.093
LBFGSB weighted Fusion	0.095
Nelder Mead weighted Fusion	0.090
TNC weighted Fusion	**0.109**

the main reasons for its better performance in the application as all the inducers are treated independently in the task. The highest score obtained in this work is 0.109 MAP@10, which indicates the complexity of the task.

5. Conclusions and Future Work

In this paper, we presented the experimental results of multiple fusion techniques for the media interestingness task presented in CLEF Fusion 2022. We used both a naive fusion scheme and merit-based fusion methods. Overll the results are much lower on the task compared to other computer vision tasks, which shows the complexity of the task. During the experiments, we observed better results for a merit-based fusion scheme where different weight optimization techniques are employed to assign weights to the individual inducers based on their performances. This verifies our assumption that individual performance should be considered in combining the prediction scores of the individual inducers.

In the future, we want to further explore different aspects of the application to further enhance the results. One potential direction could be an intelligent selection among the inducers instead of considering all of them.

References

[1] C.-H. Demarty, M. Sjöberg, M. G. Constantin, N. Q. Duong, B. Ionescu, T.-T. Do, H. Wang, Predicting interestingness of visual content, in: Visual content indexing and retrieval with psycho-visual models, Springer, 2017, pp. 233–265.

[2] S. Z. Hassan, K. Ahmad, S. Hicks, P. Halvorsen, A. Al-Fuqaha, N. Conci, M. Riegler, Visual sentiment analysis from disaster images in social media, Sensors 22 (2022) 3628.

[3] P. Bhattacharya, R. K. Gupta, Y. Yang, Exploring the contextual factors affecting multimodal emotion recognition in videos, IEEE Transactions on Affective Computing (2021).

[4] R. S. Kiziltepe, L. Sweeney, M. G. Constantin, F. Doctor, A. G. S. de Herrera, C.-H. Demarty, G. Healy, B. Ionescu, A. F. Smeaton, An annotated video dataset for computing video memorability, Data in Brief 39 (2021) 107671.

[5] K. Ahmad, M. L. Mekhalfi, N. Conci, F. Melgani, F. D. Natale, Ensemble of deep models for

event recognition, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14 (2018) 1–20.
[6] N. Said, K. Pogorelov, K. Ahmad, M. Riegler, N. Ahmad, O. Ostroukhova, P. Halvorsen, N. Conci, Deep learning approaches for flood classification and flood aftermath detection., in: MediaEval, 2018.
[7] Y. Wang, Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17 (2021) 1–25.
[8] 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.
[9] 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.
[10] 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.
[11] P. J. Silvia, Looking past pleasure: anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions., Psychology of Aesthetics, Creativity, and the Arts 3 (2009) 48.
[12] 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.
[13] Y. Liu, Z. Gu, Y.-m. Cheung, K. A. Hua, Multi-view manifold learning for media interestingness prediction, in: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, 2017, pp. 308–314.
[14] 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.
[15] J. Almeida, L. P. Valem, D. C. Pedronette, A rank aggregation framework for video interestingness prediction, in: International conference on image analysis and processing, Springer, 2017, pp. 3–14.
[16] K. Ahmad, M. A. Ayub, K. Ahmad, J. Khan, N. Ahmad, A. Al-Fuqaha, Merit-based fusion of nlp techniques for instant feedback on water quality from twitter text, arXiv preprint arXiv:2202.04462 (2022).
[17] K. Ahmad, K. Khan, A. Al-Fuqaha, Intelligent fusion of deep features for improved waste classification, IEEE access 8 (2020) 96495–96504.
[18] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95-

international conference on neural networks, volume 4, IEEE, 1995, pp. 1942–1948.
[19] S. Singer, J. Nelder, Nelder-mead algorithm, Scholarpedia 4 (2009) 2928.
[20] J. Martens, et al., Deep learning via hessian-free optimization., in: ICML, volume 27, 2010, pp. 735–742.