Media Interestingness Prediction in ImageCLEFfusion 2023 with Dense Architecture-based Ensemble & Scaled Gradient Boosting Regressor Model

Notebook for the CS_Morgan Lab at CLEF 2023

Md. Ismail Siddiqi Emon [1] and Md Mahmudur Rahman [1]

[1] *Computer Science Department, Morgan State University, Baltimore, Maryland*

Abstract

The field of computer vision plays a key role in managing, processing, analyzing, and interpreting multimedia data in diverse applications. Visual interestingness in multimedia contents is crucial for many practical applications, such as search and recommendation. Determining the interestingness of a particular piece of media content and selecting the highest-value item in terms of content analysis, viewers' perspective, content classification, and scoring media are sophisticated tasks to perform due to the heavily subjective nature. This work presents the approaches of the CS_Morgan team by participating in the media interestingness prediction task under ImageCLEFfusion 2023 benchmark evaluation. We experimented with two ensemble methods which contain a dense architecture and a gradient boosting scaled architecture. For the dense architecture, several hyperparameters tunings are performed and the output scores of all the inducers after the dense layers are combined using min-max rule. The gradient boost estimator provides an additive model in staged forward propagation, which allows an optimized loss function. For every step in the ensemble gradient boosting scaled (EGBS) architecture, a regression tree is fitted to the negative gradient of the loss function. We achieved the best accuracy with a MAP@10 score of 0.1287 by using the ensemble EGBS.

Keywords

Ensemble, Fusion, Regression, Gradient boost, Deep fusion, Dense Architecture

1. Introduction

This work presents the CS_Morgan team's participation in the ImageCLEFfusion 2023 [1] under the ImageCLEF 2023 benchmark evaluation campaign [2]. We participated solely in the media interestingness (ImageCLEFfusion-int) task, which is mainly an image interestingness score prediction regression task applied to the media interestingness data associated with the Interestingness10k dataset [3]. Generally, the concept visual or media interestingness attempts to measure the ability of multimedia (e.g., image, audio, video, etc.) content to capture and keep the viewer's attention for longer periods of time [4,5]. A growing number of media contents makes it more difficult to calculate or

[1]CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: mdemo1@morgan.edu (A. 1); md.rahman@morgan.edu (A. 2).

quantify the interestingness, which is a high-level semantic concept and highly subjective [3]. The review of the literature in the domain of interestingness prediction, overview of the traditional fusion mechanisms, and investigation of several types of deep networks for creating the fusion systems are presented in [5].

To create better, stronger image interestingness prediction results, the goal of this task is to use ensemble learning techniques to enhance the performance of individual prediction systems, called inducers or weak learner algorithms. It has been demonstrated many times since the early days of machine learning (ML) research that ensembles of classifiers can be more accurate than individual models [6,7]. The ensemble methods use multiple learning algorithms as weak learners (inducers) to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. The algorithms generally add all the weak learners such that errors from each single learner or inducer are remunerated by other inducers, which eventually provides a robust performance by averaging out. The remainder of this paper describes the proposed methodology, result analysis, and conclusion with future work.

2. Methods

This media interestingness prediction task specifically concentrates on the fusion method Where the inducer's outputs are given. Participants' task is to combine these scores and predict the interestingness score of those visual contents. To do so, we proposed to use two different architectures, a Dense architecture, and a scaled Gradient Boosting Regressor and finally submitted three different runs based on those architectures.

2.1 Dense Architecture

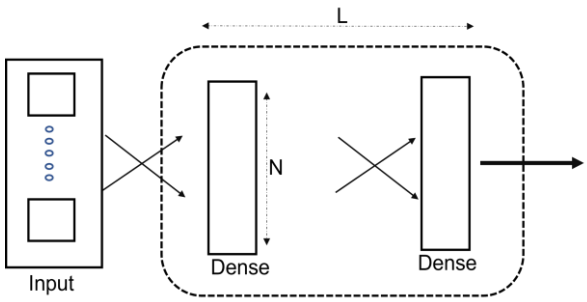

Figure 1: Dense Architecture

For the dense architecture, we put together a stack of dense layers (Fig. 1) and set up a dense ensemble with a predefined set of hyperparameters, such as a blend of different numbers of dense layers, neurons for each of these dense layers, batch norms, etc. A dense architecture contains all the neurons that are connected internally in a deep manner, which refers to the fact that every node or neuron in a dense architecture receives output from previous layers of neurons as input. For ensembling, we applied a combination of min-max (LFMinMax) of the output scores of all the inducers after the dense layers. To do this, we can implement as many dense layers as we need. For this task, we proposed two different variations, where the second architecture is basically an extended version of the first (base) architecture. The first fusion architecture includes several dense layers with a normal kernel initializer along with rectified linear unit (ReLU) activation functions. Then we compiled our architecture with the mean squared error (MSE) loss function and fitted with Adam as the optimizer [8]. Additionally, in our second architecture, we increased the number of dense layers. In addition to that, the number of neurons in each of those dense layer's increases, respectively. Later, we scaled our results to conform to the final submission format.

2.2 Ensemble Gradient Boosting Regressor (EGBS)

Boosting refers to the way an ensemble can 'boost' a weak learner into an arbitrarily accurate strong learner where the weak learner (inducer) is typically a decision tree with just one decision node. In boosting the estimators are trained in series with the training of a new member being influenced by overall performance so far [8]. The estimator performance also determines their contribution in the aggregation process. Gradient boosting seeks to optimize the training of new estimators in tandem with the aggregation process. The gradient-boosting regression trees resemble the idea of fusion in ML, more specifically late fusion, which derives from decision trees. This estimator builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function [9]. When it is used for predicting continuous target variable (as a Regressor) as a regressor, the cost function is Mean Square Error (MSE).

Mainly, this method creates several decision trees or more specifically random forests. To prevent overfitting, the main idea is to use the ensemble method used for decision trees and then average the regression results as shown in Fig. 2.

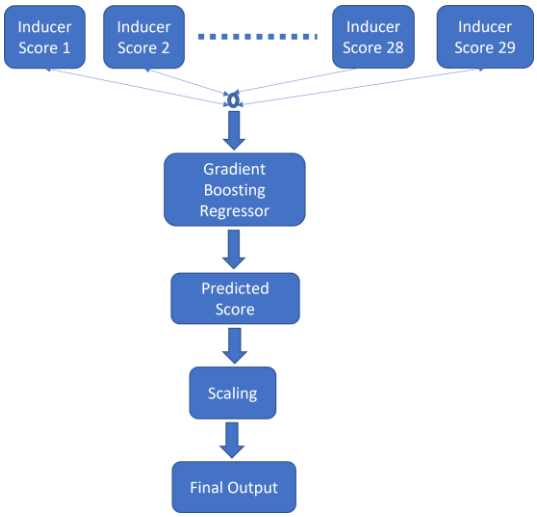

Figure 2: Ensemble of the Gradient Boosting Regressor

Here, the key idea of late fusion is working in a sequential way such that out of the M trees, the training of the first tree is performed by feature matrix ***X*** (e.g., inducer score) and labels ***y***. Using the predictions from the first tree, the residual error ***r0*** was calculated. Then the second tree completes training using feature matrix ***X*** and the residual error ***r0*** from the first tree as labels. From the prediction of the second tree, we calculate the residual error ***r1***, and so on. It is important to mention that a shrunk technique Shrinkage is applied, which shrinks the ensemble after the prediction of each tree by multiplying the learning rate ranges from 0 to 1. To meet a standard of model performance, there is a tradeoff between learning rate and number of regression trees, where a declining learning rate must be compensated with upward estimators or a higher number of regression trees. Eventually, all the trees are trained completely, and prediction is performed using the following equation, where *lr* = learning rate:

$$Y\,(pred) = y1 + (lr * r0) + (lr * r1) + \ldots\ldots\ldots\ldots + (lr * rN) \quad (1)$$

Then our prediction had to go through the final scaling pipeline to meet the actual prediction label using the following formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

3. Result Analysis

3.1 Dataset description

The data for this task is extracted and corresponds to the Interestingness10k dataset [3]. The organizer provided the output data (media interestingness prediction scores) from 29 inducers, which were gathered from the prediction outputs of the previous MediaEval Predicting Media Interestingness task under the benchmarking Initiative for Multimedia Evaluation [10]. The dataset splits into 1877 samples in the development set and 558 samples in testing set. The outputs from the 29 inducers for all the images in the development and test set are provided in .txt format, where each entry in this file contains the fields, video id, image id, classification (0 represents non-interesting and 1 represents interesting), and interestingness score by that respective inducer. Some inducers have a big range of results in the development set, which could result in unrealistic output due to the outliers. Therefore, it was necessary to convert them to the same range to meet our model consistency and get a higher accuracy.

3.2 Results of the Submission

Table 1: Results of the submitted runs in terms of MAP@10 score.

Run ID	MAP@10
Dense Architecture Fusion 1	0.0595
Dense Architecture Fusion 2	0.0595
Ensembled Gradient Booster Scaled (EGBS)	**0.1287**

We submitted three different runs based on two architectures described in Section 2. Table 1 shows the three different runs. The first run (***Dense Architecture Fusion 1***) is based on the dense architecture fusion, which includes several dense layers with a normal kernel initializer along with ReLU activation function. For the second run (***Dense Architecture Fusion 2***), we increased the number of dense layers to the range of 10 to 25, which was previously in the range of 5 to 10. Further, the number of neurons in each of those dense layer's increases, respectively, in the range of 25 to 1000. Then the train data was fitted, and output prediction was carried out. Finally, our data was scaled, which was incorporated into the system to show the ultimate results. The third run (***Ensembled Gradient Booster Scaled (EGBS)***) is based on the Gradient Boosting Regressor method described in Section 2.2. From Table 1, we can observe that the EGBS yields the best MAP@10 score of 0.1287.

4. Conclusion and Future Work

In our study, we implemented several ensemble methodologies for the given media interestingness regression task under ImageCLEFFusion 2023 benchmark evaluation [1]. Here we carry out both dense architecture and a gradient-boosting regressor for our target task. While training and experimenting, the hyperparameter tuning helped us achieve our objective. It was identified that using weight optimization and tuning the hyperparameter of the gradient-boosting regressor provides the best score for our

regression task. In our future work, we plan to investigate a few diversified deep learning-based architectures, including self-attention mechanisms where one of the potential methods would be using attention with optimized gradient regression residuals.

Acknowledgements

This work is supported by the National Science Foundation (NSF) grant (ID. 2131307) “CISE-MSI: DP: IIS: III: Deep Learning-Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support."

References

[1] Liviu-Daniel Ștefan, Mihai Gabriel Constantin, Mihai Dogariu, and Bogdan Ionescu. ‘Overview of ImageCLEFfusion 2023 Task - Testing Ensembling Methods in Diverse Scenarios’. In Experimental IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings (CEUR-WS.org), Thessaloniki, Greece, September 18-21, 2023.

[2] Bogdan Ionescu, Henning Müller, Ana-Maria Drăgulinescu, Wen-wai Yim, Asma Ben Abacha, Neal Snider, Griffin Adams, Meliha Yetisgen, Johannes Rückert, Alba García Seco de Herrera, Christoph M. Friedrich, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Steven A. Hicks, Michael Alexander Riegler, Vajira Thambawita, Andrea Storås, Pål Halvorsen, Nikolaos Papachrysos, Johanna Schöler, Debesh Jha, Alexandra-Georgiana Andrei, Ahmedkhan Radzhabov, Ioan Coman, Vassili Kovalev, Alexandru Stan, George Ioannidis, Hugo Manguinhas, Liviu-Daniel Ștefan, Mihai Gabriel Constantin, Mihai Dogariu, Jérôme Deshayes, Adrian Popescu, Overview of the 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, September 18-21, 2023.

[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] Y. Shen, C. H. Demarty and N. Q. K. Duong, Deep learning for multimodal-based video interestingness prediction, 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 2017, pp. 1003-1008, doi: 10.1109/ICME.2017.8019300.

[5] M. G. Constantin, L.-D. Ştefan, B. Ionescu, Exploring Deep Fusion Ensembling for Automatic Visual Interestingness Prediction. In: Ionescu, B., Bainbridge, W.A., Murray, N. (eds) Human Perception of Visual Information. (2022) Springer, Cham. https://doi.org/10.1007/978-3-030-81465-6_2

[6] D. Wolpert. Stacked generalization (1992), Neural networks, 5(2):241–259.

[7] L. Breiman, Bagging predictors (1996), Machine learning, 24(2):123–140.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv:1412.6980 (2014), ICLR (Poster) 2015

[9] Implementing Gradient Boosting in Python, URL: https://blog.paperspace.com/implementing-gradient-boosting-regression-python/

[10] MediaEval Benchmarking Initiative for Multimedia Evaluation, URL: http://www.multimediaeval.org/mediaeval2019/