May 2nd, 2019

ANSWERS TO REVIEWS

Dear editor,

We thank the reviewers for their valuable comments and appreciations to our paper, which have helped us to improve the paper. In this document, we answer the points raised by the reviewers. In blue we indicate the changes made on the paper and we give explanations to the different requirements. We have modified the paper accordingly to fulfil all the requests made by the reviewers.

Reviewer 1

This paper presents an interpretable deep learning-based classification framework with network compression for diagnostic of diabetic retinopathy. Interpretability of deep learning is indeed important for designing a deep model. To compress a deep network while preserving the classification accuracy is also significant.

The main objective of this paper is the presentation of a method for compressing the internal representation of the network into a space of mathematically independent components. The method of interpretability used in the paper was presented in a recently published article, which was not available at the moment of submission of this paper (https://doi.org/10.1016/j.neucom.2018.07.102). In order to clarify the key contributions, we modified the abstract (objectives and conclusions), clarifying that the proposed compression method is independent from the visualization method used, although we use a specific method for its visualization.

1. It is suggested to review some recent works about designing interpretable deep learning techniques as well as compression of deep models, and explicitly present the major novelties and contributions of this paper.

As previously explained, the main contribution of our article is not the design of an interpretable classifier. We have clarified the objective that was prone to error. Our main objective is feature space compression. Then, as suggested by the reviewer, we have added some references to the different available methods for dimensionality reduction, i.e. feature compression. Although our compression method is not linked to any particular interpretation method, we added also some references to novel interpretation methods like CAM, Grad-CAM and the one used in this work published recently in Neurocomputing (https://doi.org/10.1016/j.neucom.2018.07.102.). The related work section includes a brief explanation of all these techniques.

2. It is suggested to explicitly explain which useful features are exactly learned and provide significant contributions to image classification to diagnostic of diabetic retinopathy.

The purpose of compressing the features from 64 to 3 is not intended to find 3 axes with a clear correspondence to the medical signs that the ophthalmologists use to distinguish the different categories of Diabetic Retinopathy. It is, in fact, impossible to find 3 numbers that summarize the information of the different types of lesions, their size, their number, their location in the eye, the distribution (scattered or concentred), etc.

Our aim was to find a low number of vectors that could be used as input variables of a more complex classifier for the detection of Diabetic Retinopathy. Many classification methods require independent variables, which makes impossible to be used with the 64 features obtained by the DL model. On the contrary, in this paper we show that with 3 independent ICA vectors we can maintain the same level of quality (with a linear classifier). This compression of the initial 64 correlated features is the one that permits to have better explanatory tools to address the study of other classification techniques.

In fact, in our research group we have been working on the implementation of a fuzzy rule-based classifier that takes 8 indicators of the patient's physical condition and classifies in Diabetic Retinopathy levels (e.g. body mass index, creatinine level, age, etc.) [see papers 1,2]. Our goal is to add the 3 ICA vectors to these 8 indicators and study if we can improve the classification quality when integrating physical information and eye-fundus information together. However, we are not able to initiate this study now because we do not have the information of that 8 indicators together of the eye fundus image for the same patients. A hospital in our region has started to collect that data and maybe next year we can approach that step of the research.

We have included a brief explanation of this idea in the final section of the paper.

[1] Saleh, E., Blaszczynski, J., Moreno, A., Valls, A., Romero-Aroca, P., de la Riva-Fernandez, S., Slowinski,R., Learning ensemble classifiers for diabetic retinopathy assessment, Artificial Intelligence in Medicine, vol 85, pg 50-63, 2018

[2] Saleh, E., Valls, A., Moreno, A., Romero-Aroca, P., Integration of different fuzzy rule-induction methods to improve the classification of patients with diabetic retinopathy, Book: Recent advances in Artificial Intelligence Research and Development, Frontiers in Artificial Intelligence and Applications, vol. 300, pp. 6-15, IOS Press, 2017

3. It is suggested to review some recent deep learning-based works for diagnostic of diabetic retinopathy, and provide some comparisons.

The objective of this paper is the presentation of a procedure for compressing the network information. The diabetic retinopathy classifier used, referenced as "baseline model" was also presented and discussed in the previously referenced article (https://doi.org/10.1016/j.neucom.2018.07.102.) that at the moment of submission it was still not available. Such paper was published last week and now we added some references to it where such comparisons are already discussed.

Reviewer 2

In this manuscript, an ICA-based feature analysis and visualization method was proposed. The work is interesting; however, the writing of the manuscript was not clear and needs major improvement.

1. In introduction, the author claimed that the work can "compressing the internal representation of a deep learning classifier" using the ICA method. I am not sure what the author means for "compressing", is that mean the ICA can reduce the feature dimensions from 64 to 3? If so, it can be realized by simply adding a FC layer with 3 channels. Please give a clear explanation for the "compressing" effect of the proposed method.

Independent Component Analysis allows the generation of an arbitrary number of independent subcomponents with interesting properties: they can be modelled as non-gausian signals and they are statistically independent (maximizing KL divergence). Adding an additional layer does not guarantee the particular properties that are obtained from the independent component analysis. In the paper, it is shown that the nature of the neural network feature space is a highly correlated and redundant vector space. We modified the article to clarify this aspect. The importance of obtaining independent components has been mentioned in the discussion and future work explains how this feature permits the use these vectors in machine learning methods.

2. In introduction, the author claimed that the method "significantly reduce the data size to be used in the detection of the important image regions or pixels, being much more efficient in the generation of explanations". Why features with 3 dim is much more efficient in the generation of explanations? As far as I know, the feature visualization method such as CAM (Class Activation Mapping) can achieve the same goal no matter how many features were extracted.

As the reviewer points out, the statement is not described in the right way. What we wanted to express was that it is easier for an ophthalmologist to interpret three independent variables than 64 highly correlated ones. With the mapping of these three components into the original image, it is possible to identify the image areas that are related with the disease grade. Although the values of the 3 axes are not directly interpretable in medical terms, they help to identify the areas in of the eye fundus image where the doctors have to focus in order to find the RD disease signs. We have modified the text in the paper to explain this.

3. The related work section only reviewed 3 related works and analyzed only one work besides your previous work. I am sure there are more works for DR classification using DL. Furthermore, since visualization is one of your major contributions, the previous feature visualization works should be reviewed in this section.

As explained in the answer to Reviewer 1, we have clarified the objectives of the paper emphasizing that the main contribution is the introduction of a compression stage in the model and leaving a part the interpretation, which has been recently published in a previous article. As stated before, the proposal for compression is of general applicability and it is not affected by the visualization method used. For purposes of our work, we have used the pixel-wise derived score propagation method proposed in other published papers, but any other method can be also used. We have also included such clarification in the article.

4.Quadratic Weighted Kappa was used for evaluation of the proposed method. However, no detail explanation of the QWK score was given in the manuscript. And why do you use the QWK score while others using sensitivity and specificity? Using different evaluation metrics will make it difficult for comparing your work to others. More comparisons with state of the art are expected.

A brief explanation of the QWK index is given in the introduction with references to other papers where it is explained in detail. As explained now in the paper, QWK is a good index for summarizing in only one index the performance of multi-class ordinal regression problems. Sensitivity and Specificity are good for binary classification problems. In the case of multi-class classification different sensibility and specificity values have to be considered grouping together classes considering 1 versus all comparisons. QWK is of common index used in the medical community and in diabetic retinopathy bibliography for inter-rater agreement measure in ordinal regression problems with more than two classes. In (https://doi.org/10.1016/j.neucom.2018.07.102.) we included comparisons of our deep learning model with other models on other evaluation indexes, like sensitivity and specificity. It is out of the scope of this article the model design, therefore we have referred to this paper that now is available online.

5. In section 3.1, the author mentioned using PCA for appraise the redundancy. Is PCA used in the network? If so, why no PCA block in fig.1 (b)?

The paper was not written appropriately regarding the PCA use. PCA is not used in the network at any step. PCA was used as a prior exploratory method for understanding the nature of the feature vector space obtained from the deep learning network. We used PCA to compare between the number of components required to explain different significant percentages of variance and in this way evaluate the redundancy of the space. As indicated in the article, we could see that with only 10 components we can explain 99% of the variance. Such fact experimentally proves the highly redundant nature of the space. Although PCA gives direct information about the percentage of total variance explained by each component, it does not maximize the independence between them. For this reason, as we wanted to obtain mathematically independent components, ICA method was chosen to be included in the deep learning model. We added this clarification in the paper.

6. In fig.1 (b) the ICA was adding after feature extraction. Is the ICA embedded into the network? How the network was trained? Is it end to end?

The network is trained as a conventional network. After training, the internal representation, i.e. the feature space vector for each image, is calculated. Such values are then used for calculating the optimal number of independent components as well as its transformational matrix. This process is made only once after the training. Such optimal matrix calculated from the optimized network is the one used on inference time and for the visualization in the input space. We added this clarification in the text.

7. In section 4.2, the baseline network was self-designed. How was the baseline network designed? And what is the performance compared to other classic networks?

As previously stated, we used the same network that was presented in a previous accepted for publication article that was not still available at the time of submission (https://doi.org/10.1016/j.neucom.2018.07.102). We added a reference to the paper where comparisons with other models and performance indexes can be found.

8. In fig.2, why no contribution of all three components for class 1 and 3?

It is true that with these ICA vectors, Class 1 and 3 are not predicted, being only classes 0, 2 and 4 the ones that distinguishes the model. In this regard, it is important to note that Diabetic retinopathy is an ordinal regression problem, where classes 1 and 3 are defined as intermediate transitions between the other classes. Thus, we curiously have found that this automatic model maintains the quality of prediction in terms of the QWK index by just using the 3 most relevant classes, despite doctors are normally using the 5 categories. We clarified this point also in the text.

9. In section 4.4, the definition of “input space final scores” was not given. The fig.4 and 5 was confusing, please give more explanation.

We changed "input space final scores" to the more clarifying "input-space explanation maps" and added some explanation.

10. In discussion, the contributions of this work were not well discussed. Please refer to the “Highlights” listed at the beginning of the manuscript and give clear discussion for every point.

The highlights of the paper are the following ones:

1. Developing of a generalized way of compressing feature-space information of a deep learning classifier.
2. Determination of the independent vectors that are able to represent the classes of diabetic retinopathy from the information extracted from a deep learning model.
3. Visualization of each one of the independent components in the original image.

We have rewritten part of the discussion section to focus more on these 3 point, putting together the results obtained and commented in this section.