July 24th, 2019

ANSWERS TO REVIEWS

Dear editor,

In this document, we answer the points raised by the reviewers in this second revision stage. In blue, we indicate the changes made on the paper and we give explanations to the different comments and questions. We also indicate how the paper has been changed accordingly.

-Reviewer 1

- It is suggested to accept this paper for publication.

No observation to be done.

-Reviewer 2

- The authors emphasized a number of times that the paper is mainly focusing on feature space compression. And it is clarified by the authors that ICA is not embdeed into the network. It seems to me that the works is a simple application of ICA to the features extracted by the network presented in the author's previous work [19] and visualized, again, using the technique presented in [19]. As a result, the contribution of the work is too weak.

This work continues the research done about constructing a DL model that has good performance (QWK close to 0.8) but with a smaller model, that uses much less resources. The DL model we have constructed has been recently published in Neurocomputing (ref [19] indicated by the reviewer). In the present paper, we approach the compression of the last feature layer using ICA in order to generate a new space with independent components with two aims:

1) The ICA components will enable to disentangle the information of the features in order to be easily mapped into the original eye fundus image.
2) ICA components are independent and this property improves the performance of other classification techniques such as KNN, decision trees, etc [Sanchez,2004]. The next step in our research is to use the 3 ICA values of an image together with other patient's data coming from analytical tests (ie. Blood tests) as well as other patient description data.

Although goal 2 is still not possible due to the lack of data, we believe that the contribution of the paper goes in the line of the demonstration of the utility of ICA to keep the classification accuracy while at the same time permits a propagation of relevant features to the input space of the image. This explanation has been added at the end of the paper, as future work.

[Sanchez2004]: ICA as a preprocessing technique for classification, ICA conference, pp. 1165-1172, 2004 https://link.springer.com/chapter/10.1007/978-3-540-30110-3_147

Give feature compression is the main focus, the experimental results ane evaluation are weak as well. Only PCA is visually compared, no evaluation is conducted with other popular compression approaches like LDA, SVD etc. The authors keep saying that the components extracted by ICA is independent, which we all know. But, what's the advantages of using ICA for feature compression? Is the components more useful for clinical diagonosis, any evaluation done by ophthalmologists.

I am sure the ICA is not helpful for classification, as the simple expereiments in the paper suggest that the QWA of ICA components is not as good as that using the original features. The evaluation, again , is very simple, no comparition with popular networks like AlexNet, VGG, ResNet, is conducted, and no other feature compression techniques like PCA, LDA are involved as well.

The reviewer has pointed a very interesting issue of comparing other compression techniques. From the usual techniques, LDA uses a supervised approach in order to make the separation. In our case, this is already done by our classification layer and we do not want to change this structure of the system. Moreover, we want to use an unsupervised method in order to discover independently from classes the underlying causes, that is why we use PCA and ICA (note that SVD is equivalent to PCA). Therefore, we have added a comparison with the results obtained of compressing using 3 PCA components, having worse results (QWK=0.695). Therefore, finally these results confirm that ICA is better than other techniques, not only because we aim at minimizing the mutual information between components (to find the independent components for interpretability purposes) but also when comparing the classification performance.

The used network has a good performance for this particular classification and serves as a baseline for testing our feature compression technique. As we aer not focusing on improving the classification performance, we do not compare to other types of networks. Using the same baseline network than the one used in previously published articles facilitates the comparison with previous results. See the answer to next question too.

Why use a self-designed network for testing, when there are so many classical netwoks available in literature?

The DL model we have constructed and recently published in Neurocomputing is using about 400,000 parameters and 17 layers while the best state-of-the-art method uses 23 million parameters and 160 layers in the network. We considered important enough to study the impact of ICA with this type of small DL models, because they can then be executed with standard computers, without requiring the large computing power of the other models.
As the work presented is framed in a funded research project, the study is focused on the improvement of decision support and explanation of this particular network that we are constructing with the aim of being deployed in the Catalan health care system in the future. This will be possible after conducting a pilot test that is now under definition.

-Reviewer 3

The paper proposes a method for compressing images features into feature space. The resulting compresses features are used for classification. The validity of the proposed method is demonstrated using Diabetic Retinopathy image dataset.

Comments:

Introduction: The novelty of the paper should be explicitly stated.

We explicitly introduced a sentence clarifying the novelty as follows:

"The novelty of the paper is based on the separation and visualization of the mathematically independent features that explain the DR classification."

Related works: Several recent notable works directly related to this article, which discuss state-of-the-art in applying deep learning for detecting diabetic retinopathy, and extraction of novel features from eye fundus images, are recommended to be included for citation:

- Visualizing deep learning models for the detection of referable diabetic retinopathy and glaucoma. JAMA Ophthalmology, 137(3), 288-292. doi:10.1001/jamaophthalmol.2018.6035
- Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. Eye (Basingstoke), 33(1), 97-109. doi:10.1038/s41433-018-0269-y
- Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images. IET Image Processing, 12(4), 563-571. doi:10.1049/iet-ipr.2017.0636
- Detection of saliency map as image feature outliers using random projections based method. ICENCO 2017 - 13th International Computer Engineering Conference: Boundless Smart Societies, 2018-January 85-90. doi:10.1109/ICENCO.2017.8289768
- Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. PLoS ONE, 12(6) doi:10.1371/journal.pone.0179790
- The image classification with different types of image features doi:10.1007/978-3-319-59063-9_44

We introduced 4 of your recommendation papers in the related work section. We have found that the other two were not so closely related with our article as they deal with feature extraction.

Figure 1: I suggest to use standard process (or data flow) diagrams to avoid ambiguity between processes and data.

As you suggested we changed the diagram according to data flow standards.

Section 4.2: suggested to move to Section 3 “Methods”, since it describes the architecture of the network.

As you suggested we moved this subsection to section 3

Figure 2: what is contribution? Weight?

We clarified substituting Contribution by Weight.

A formula for calculation of QWK value should be provided.

As suggested we added the mathematical definition of QWK.

Section 4: provide a confusion matrix of the classification results.

As suggested we added the confusion matrix to the results. We added the values of sensitivity, specificity and F1 score considering a simplification to a binary case (classes 0 and 1 versus classes 2, 3 and 4). We can observe better results when using ICA than with the original model.

Section 5: is the difference between K(orig) and K(ICA) statistically significant? Apply statistical procedures such as ANOVA to test.

We calculated the confidence intervals for both indexes. The results show that the difference is not significant at 95% confidence.

Conclusions: „the mathematically independent signs of the disease“ -> not clear, should be rewritten.

Changed to “mathematically independent features”.

- **Reviewer 4**

In general, the paper is very poorly written. The flow is fractured going from one topic to another without a transition. Section titles would have helped to organize the paper better.

The sections used are the ones established by the journal: introduction, related work, methods, results and discussion. We have revised the paper in order to try improve the transitions from one topic to another. We have also moved part of the introduction to the related work and methods sections, in order to make the paper easy to follow.

Specifics follow:

1. The authors reference their previous publication (19) to support their work. This is like self-referral and has little impact.

Attending the suggestions of the reviewers, we added new references about the state of the art of the research of DR classifiers. So the new version includes much more related work.

2. Different goals are stated for the paper. One is to compress the internal representation of a deep learning classifier. Use of ICA and PCA accomplish this goal is described. However; another goal to improve the "visualization for human interpretation" is not clearly met. Though several retinal images are presented as evidence of meeting this latter goal, there is no strong statistical evidence that this process can be generalized.

3. There is no evidence given that the regions identified by the process are useful to the ophthalmologist or eyecare specialist. The authors do not show that the regions marked contain pathology associated with DR and the DR level in question.

Issues 2 and 3 refer to the problem of validation of the interpretability of the visualization provided using this method. In this regard, at the moment we do not have a large set of DR images with marks of the lesions provided by the experts. We could only validate with a reduced set of 25 images. The lesions indicated by the doctors appeared also in the maps generated by the system, which confirmed that the method proposed is able to detect the areas of the image where the doctor has to focus.

We have added to the paper the presentation of images marked by an expert together with the ICA areas found with our method. So, we show now the relation between the areas detected and the lesions in the image. Due to the reduced set used, we do not have enough evidence to establish a direct correlation between the lesions types and each of the 3 ICAs. It seems that each ICA1 and ICA2 found the lesions that appear in Mild and Moderate DR, while ICA0 finds another type of lesions found in Proliferative DR. Sometimes, it is also the case that some ICA marks are focusing on different indicators of the DR disease, but a medical explanation is now not available with this small set of images available.

Due to the work done in this validation, we have now added Dr. Pere Romero as coauthor, who is the head of the ophthalmologist department at Hospital Universtari Sant Joan de Reus.

4. The connection between the "disentangled feature components" and the enhanced interpretability of the images by the human is not well presented. The calculated components do no help to score the importance of a pixel. This claim is not well supported.

We have added the evaluation manually done by the head of the Ophthalmology department of the Hospital Sant Joan de Reus. Some images with the explanation and interpretation of the activated regions in each ICA are now explained.

5. The discussion of Kappa is one example of a topic being thrown into the discussion without any transition f rom the discussion of feature space compression.

We have moved the presentation of Kappa to a specific subsection. We have also made some changes to kappa values obtained with the different methods in the discussion section.

6. Quoting the paper: “We use a the pixel-wise score propagation method to visualize such IC in the input space.” With no definition of IC.

In the first appearing of the term we have included the definition: independent component (IC)

7. The figure showing the flow of their DL and classifier network is not clearly explained. In the second part of the figure when ICA is introduced, it can be interpreted as being part of the DL learning processes, even though the text says otherwise.

We clarified in the diagram that ICA calculation is made offline. We have also changed the symbols using more standard ones, to avoid misunderstandings.

8. There is also some confusion about the data and number of cases that overlap with the Gulshan (23) study that also used the same data source. Gulshan achieves a sensitivity and specificity of 97% and 93.5%, respectively on 125,000 graded cases. Is the set of 88,650 a subset of the 125,000 or a different set. If different what was the grading protocol and by whom? The Gulshan study sets an excellent baseline for comparison of the proposed technique.

Gulshan uses a superset of the data used by our study. Such superset is not publicly available.

It is important to note also that the superset used by Gulshan was revised and relabeled by a panel of ophthalmologists to have more reliable labels than the original ones. This may be the reason for the difference on the performance. However, we cannot use this improved data set.

9. Instead the authors state: “We choose the minimal n that allows achieving maximum performance. The optimal n for this problem is 3, achieving a QW Kval =0.790 not far from the achieved by the original model without dimensionality reduction (QW Kval =0.800).” Where the QWKval of 0.8 comes from is not references. Why sensitivity and specificity were not used is a major weakness.

According to the suggestion of reviewer #3 and your suggestion we added confusion matrices and values of sensitivity, specificity and F1 score for both models. We also calculated confidence intervals, to check that the difference is not significant at 95% confidence level.

10. Similarly, the results summary for the 88,650 cases is not given. They state that the cases were randomly selected. This does not guarantee an even distribution of disease levels. In fact the data summary, which is not given, would show a highly skewed distribution to early disease.

The data used is one of the bigger public datasets available for DR (EyePACS dataset: kaggle.com). The data distribution is the same of the dataset. Many publications are based on the same dataset, ie. on the same data distribution.

For informative purposes, we added in section 4.1 the information about data distribution.

11. The authors do not consider that the intermediate levels, which had no representation by the components, are not important. On the contrary level 3 (severe) is highly critical to differentiate from moderate or proliferative.

For classification purposes, obviously it is better to have a differentiation between the 5 classes, but for assistance in image interpretation by doctors, we are more interested in finding the regions with evidences of causes of DR lesions than the proper classification of the image into the 5 categories.

12. Finally, Retinal images do not show the pathology, blocked by the annotations.

We changed the presented score maps. Moreover, as said in question 3, we present now three different samples with different disease levels, i.e. mild, moderate and proliferative. In each one of the cases, it is presented the original image, an image tagged by an ophthalmologist marking the zones with the most important lesions and finally the three ICA score maps. After the generation of the score maps, the expert ophthalmologist (Dr. Romero) has analyzed the lesions zones to evaluate their correct identification. The results of such validation is presented in the paper for each one of the presented images.