RESPONSE LETTER TO REVIEWERS:

First of all, we would like to thank the reviewers for their critical and valuable comments and suggestions. We have revised the manuscript according to their comments and questions. The modifications and revisions are listed as follows according to each of their comments.

Reviewer #1: The study concentrates into the proposal of weighted kappa as loss function for ordinal regression in deep learning.

It derivates the equations required for applying first order optimization algorithms. It solves a diabetic retinopathy ordinal classification problem. A performance comparison of the model trained by both log loss and kappa loss is shown. It checks stability of the kappa loss function using different input and batch sizes.  
  
My main concern is to see the extention of the study for another case, database and to confirm the results. This will be important for convencining the readers.

The references are recent.

The language of the study is understandable.

Following the suggestion of the reviewer, we have performed the same analysis with two other datasets, comparing the performance of the optimization with log-loss and with qwk-loss functions. We have taken two new datasets. Both datasets came from different Kaggle challenges where the evaluation function used for measuring the performance is the quadratic weighted kappa. The two new datasets are solved using neural networks of increasing complexity comparing the optimization using the logarithmic loss and the quadratic weighted kappa loss.

Therefore, section4 has been re-organized. It is divided into 4 subsections, one for each case study and the last for general comments about the findings on the 3 tests done.

Results confirm that the use of the qwk-loss function during the training of the network produced better classification rates. The QWK values obtained in the test set are larger when training with QWK than when training with log-loss. In this way, the neural networks learn the order relation between the different output classes, being able to achieve make a more accurate classification of the new samples.

We thank the reviewer for this suggestion. We believe that the paper has significantly improved by adding these other validation studied, which also demonstrate that the method is valid not only for image data but also for other large numerical and categorical databases.

Reviewer #5: Thanks for this interesting work.

Experiments and results did not convince me to accept the paper although I initially liked and willing to accept the paper after reading introduction and methods.

As explained above to the first reviewer, section4 has been re-organized to include 2 other case studies that support the results and conclusions found.

In the revised manuscript, article we present three different problems that use different types of data and with different goals (all of them require a classification into an ordered set of categories). Optimization is done with log-loss and qwk-loss, using neural networks of increasing complexity in order to test different situations. In all cases better results are obtained from optimizing qwk-loss than from optimizing log-loss.

-Introduction and methodology sections are easy to follow. However, Results section is very difficult to follow. Log-loss and QWK are confusing at the first read, also figure 2 is confusing in my first read. There should be better explanation style.

Can you explain why "log" and "qwk" loss value start at different values, it may be because of formulations. But figure 2 is really a good proof to show the superiority of qwk? Also, the vertical line shows "QWK" in your figure, should be error I guess?

Figure 2 of the original paper was presented using a scaling of the y-axis in order to see in greater detail the saturation zone. This made invisible in the graph the (0,0) coordinate, where both functions converge. We presented this figure not as a proof of better classification but as informative of the process of optimization of the network. Based on your feedback, in the new version of the article we substituted this figure by others that support our claims. They are the histogram of the confusion matrix of the test set and the box-plots of the confidence intervals of the quadratic weighted kappa index of the three use cases over their corresponding test sets.

-Deep learning literature at least big picture is missing.

Following your suggestion we introduced a new paragraph in the introduction with a big picture of the deep learning methodology and new bibliography references.

-Also, there should be better way to explain what you are trying to do in Figure 3. Confusing figure. It is apparent that qwk is usually got better achievements, but difficult to see details in that figure.

We modified the figure in order to make data more visible. We also changed the tags to make them more informative and clear.

-Experiments and results did not convince me to accept the paper although I initially liked and willing to accept the paper after reading introduction and methods. More data may be assessed. Or there should be some other ways to convince the reviewers and the readers. In this version, I would recommend to work on the experiments and generate some good ideas to convince the referee and readers. Classification rates, timing, energy loss should be evaluated at the same time.

As said before, following your suggestion we introduced two new case studies with different types of data and with quite different purposes, with the aim of showing the significance of the method proposed in different fields. The problems are solved using models of incremental complexity where the evaluation function is the quadratic weighted kappa. We have studied the increase of performance in the three problems. In all three studies qwk-loss optimized models performed better (with high level of statistically significance) than log-loss optimized ones. We kept our analysis centered on the ordinal classification index QWK since it is the usual one for this kind of ordered rating of elements into a finite set of categories (see Kaggle competition evaluation). Energy loss is the same than QWK in this case. Time evaluation has not been included because it greatly depends on the size of the data and on the computer capacity, not on the loss function optimized. Although the computational cost of calculating qwk is higher than computing the log-loss this cost is insignificant compared with the amount of computation of the forward and backward calculations of the overall network.

For visualizing the differences between the two losses we introduced a figure with the prediction histograms (true class vs predicted classes) of the confusion matrix of the more complex use case (retinopathy classification) over the test set in order to show how in case of failure in the prediction the new loss functions predict “closer to the true” classes than the standard one. We think that this figure helps to clarify the difference between the results of both functions.

Details:  
-Could you rephrase the paragraph, it is long and difficult to follow:  
"Weighted Kappa has been normally used as an index to measure the inter-rating agreement between two raters in a multi-class classification problem where the categories have a defined a priori ordering, in such a way that the classes to categorize are a high level abstraction of some sort of intrinsic information that we want to extract from data."

We changed this paragraph to:

“Weighted Kappa is a usual index to measure the agreement between to ordinal classification assignments made by two different raters. It is especially suitable when there are more than two categories that have an intrinsic order, defined a priori according to some high level criterion. The use of Weighted Kappa index as loss function should enable to reveal this high level ordering knowledge that lays beyond the set of categories.”  
  
-To eliminate confusion, can you please replace dot with commo in the population numbers? Or any better idea?  
"The training set contains a total of 35.126 images" -> "The training set contains a total of 35,126 images".

We replaced dots by comma in the thousands numbers.  
  
-Can you enlarge those examples below, such as ImageNet and many bigger data sets?  
CIFAR-10 (Graham,2014), CIFAR-100 (Clevert et al., 2015), STL10 (Dundar  
et al., 2015), SVHN (Liao and Carneiro, 2015).  
We introduced a new reference to ImageNet.

-Typos should be checked by an native English.

English has been carefully checked.