sensors-898288

Dual oxygen and temperature luminescence learning sensor with parallel inference

Reply to Reviewer 1

The authors would like to thank the reviewer for the positive feedback and the constructive questions and suggestions.

The reviewer commented that “the manuscript failed in explaining how the sensor is interesting for the user.” The authors are strongly convinced that the sensor is extremely interesting because it measures oxygen and temperature without a separate sensor (or an additional indicator) to measure the temperature. The measurement of the temperature is a perhaps trivial problem from the point of view of the physics or technology, but it remains a practical challenge. Any temperature sensor mounted on the housing or mechanichs of an optical sensor will always measure the temperature of the housing, which is never the temperature of the indicator. The determination of the temperature is a well-known source of error for sensors based luminescence quenching. More generally, the same approach can be used for parameters other than oxygen concentration and temperature, as also pointed out by the reviewer.

The question of the calibration is a legitimate one. Since the parameters of the neural network model are specific for the spot, when substituting it with a new unknown one, an adjustment to the network is most certainly needed. To train the neural network for the first time requires a long time (65 hours only for the data gathering part). But subsequently, what is called transfer learning could be used to quickly fine-tune the network if the conditions of the training and usage are not the same. A schematic representation of what transfer learning is is depicted in the Figure below (taken from page 148 of Michelucci, Umberto. *Advanced applied deep learning: convolutional neural networks and object detection*. Apress, 2019)

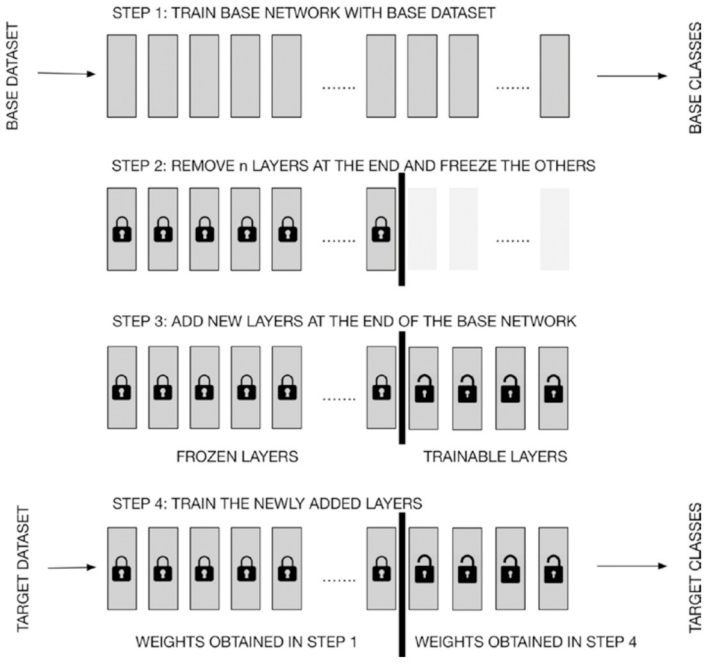


Figure 4-6. *A schematic representation of the transfer learning process*

From page 148 of Michelucci, Umberto. *Advanced applied deep learning: convolutional neural networks and object detection*. Apress, 2019

The authors have already done a few tests and have found out in very preliminary results that using only a few measurements (ca. 50 different values of T and [O2]) and just a few minutes of training would be enough for a new sensor to adjust the network trained on the old sensor. Those results are very preliminary and need to be verified since the work is in progress.

The method can most certainly be used for real-time sensing. Once the network has been trained, the response time depends mainly on the time required by the sensor to sweep the frequency range and collect the phase shifts. The algorithm requires almost no time. A sentence about the response time was added in section 3.2.

The reviewer asks about the performance at low oxygen concentration. The most relevant limiting factors for low concentrations are the sensitivity of the indicator and the experimental error on the value of the concentration during the acquisition of the data for the training. The spot used in this work (Pt-TFPP) is specified from the manufacturer to have a limit of detection of 0.03%. However, for concentrations below 0.5-1% air (saturation) or equivalently 0.1-0.2% O2 we would recommend another indicator with higher sensitivity in the lower range (for example Pd-TFPP). The method described here is however, independent of the sensor used. Similarly, the dynamical range is mainly due to the dynamical range of the indicator or spot itself.

The reviewer mentions the paper https://doi.org/10.1021/acssensors.9b02512. The authors thank the reviewer for the suggestion. A reference to the paper was added in an introduction as an example on how, by measuring multiple quantities (specifically the apparent lifetime and an intensity ratio) it is possible to perform dual sensing with one indicator. The approach was discussed by O.S. Wolfbeis, SPIE, Vol 1368 (1990), also added as a reference. The paper, although very interesting since proposing an ingenious material, clearly shows 1) you need to measure two physical quantities instead of just one as proposed by our approach; 2) the parametrization of the sensor response scales correspondently in complexity (a nonlinear fit of the two equations 3 and 4 is required, with a plethora of temperature dependent coefficients). The above mentioned references were added in the introduction.

Please find below the answer to the specific points:

Line 15-16: Within the manuscript, the method is applied to a single indicator and not to a multi-indicator system. Multi-indicator systems are more complex as inherent parameter interactions might occur that are not predicable leading to an error propagation throughout the evaluation procedure. As the method proposed here bases on a single indicator, I doubt whether the method can directly be transferred to a multi-indicator system without any adjustments.

It would be extremely interesting to apply the method to a multi-indicator system. Particularly for complex systems, with inherent parameter interactions, an approach based on neural networks may solve the difficulties of the parametrization. The authors do not have at disposal a multi-indicator system but would be very interested in trying the method on a multi-indicator system.

Line 25-27: unclear

The authors modified the text for clarity.

Line 43: Stern-Volmer is not an empirical model and T-dependencies neither.

The authors agree with the reviewer that Stern-Volmer is not an empirical model. In the paper, the term empirical was used in conjunction with other models proposed for oxygen luminescence quenching (see for example Carraway, E. *et al.* Analytical chemistry 1991, 63, 337–342 and Demas, J.N. *et al.* Analytical Chemistry 1995, 67, 1377–1380). Regarding the temperature dependence of the sensor response, it depends not only on the properties of the indicator (which may be described by an Arrenius-type dependence) but also on how it is immobilized and on the chemical matrix used. Therefore, in practice, the sensor-specific temperature dependence needs a parametrization which is determined empirically.

The authors realize that the usage of the term can be misleading and therefore removed it from the paper.

Linear 51: the manuscript describes a new approach on data evaluation not a new sensor.

For clarity, the authors have updated in the paper the term with “an approach for sensor development”. This work describes not only data evaluation but also the experimental setup requirements needed (including the data gathering algorithm) to get enough data automatically for the neural network training.

Line 66-68: Are all parameter-dependencies and especially all sensor-specific response characteristics learnable? What about the photo-degradation over time? Is it something the method might be able to compensate?

This work shows that the studied parameter-dependencies (T and [O2]) and especially all sensor-specific response characteristics are learnable.

The photo-degradation affects the phase shift, and with time, it is to be expected that the photo-degradation will reduce the accuracy of our sensor just as it happens with conventional luminescence oxygen sensors. In theory, it should be possible to use time as an additional dimension for the training of the neural network, so that the algorithm could learn the changes over time. We have not yet tried this, and therefore since we don’t want to speculate without results, we have not added this idea to the paper.

Line 72: Even though decay time-based approaches are more robust, might it be possible to use also an intensity-based approach?

Yes, this would be possible. The neural network model would have as input normalized intensities instead of phase shifts.

Line 76: The Temperature dependency depends on the chosen indicator and range. There are multiple systems known where the T-dependency can be described as a linear correlation with a negligible error propagation.

The text was modified removing non-linearly.

Line 85: What is the expected dynamic (in terms of decay time) of the indicators regarding the oxygen and the temperature dependency?

The dynamic depends entirely on the indicator chosen for oxygen sensing. A review of indicators for oxygen can be found in Wang, X.d.; Wolfbeis, O.S. Chemical Society Reviews 2014, 43, 3666–3761.

Line 87: Even though the setup was already described somewhere else, it would be great to include it in the SI. Please, add also the reference sensors that are needed to ensure that the target parameters are matched.

We agree with the reviewer. However, the editor requested explicitly to remove it so to include only unpublished material in the paper.

The oxygen concentration was adjusted with two calibrated Bronkhorst F-201CS Mass Flow Controllers. The temperature was measured with two industrial calibrated Pt-100 sensors connected to the temperature controller (PTC10, Stanford Research Systems, Sunnyvale, CA, USA). Additionally, prior to the experiment, the experimental setup was checked using certified Pst3 spots.

Line 91: Have you experienced any background effects at low modulation frequencies (200Hz)?

We did not experience any background effects at low modulation frequencies.

Line 100: The section 2.3 requires elaboration. It might be nice to get a better understanding on how the data are analyzed. How are the (unknown) sample data analyzed after the training? Maybe you could use one sample as explanation and attach it to the SI. What about the visualization? Can you show how the final matrix pattern, identified by the algorithm, looks like?

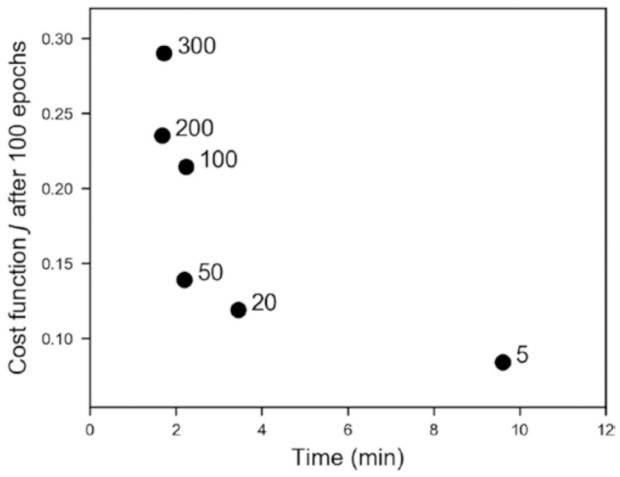
The raw data (phase shifts) without any processing are shown in Fig. 2 to 4. The data (the phase shifts either divided by 90° or by Theta_0) are the input of the network without any manipulation. The network learns during the training and is the equivalent of a calibration matrix. After the training the sensor is ready: when the sensor measures the phase shifts (again either divided by 90° or by Theta_0) the algorithm returns two values: O2 and T. In section 3 an extensive analysis of the predicted values of the neural networks was done, where we analysed the distributions of the resulting AE (similar to residuals, although in a non-common format). In general, the neural network has only two real numbers as outputs: [O2] and T. Figure 5 (the distributions of the AEs) is the most efficient way the authors found to visualize enough information on the many predictions.

Line 104-105: You submit your paper in a sensor journal who might be not familiar with the network architecture. Please, revise this sentence for a reader-oriented outreach and include more information on how the network architecture is supposed to work.

In the original version of the paper, the section was much longer with all the details. Unfortunately, the editor requested explicitly to remove this part completely since already published. The text has been slightly expanded for clarity adding details as suggested by the reviewer. An additional reference for the reader on the specifically mentioned architecture was added.

Line 113: Elaborate in more detail, why 32 is the optimal number. How did you optimize the batch size? According to your literature, a special tuning is needed here.

This is a very good question. The tuning was done in steps. One starts using the entire batch as input and reduce the batch size incrementally. In the beginning, after the same number of epochs, the metric goes down quite fast with decreasing batch-size, until at a certain point, the time needed for the training grows rapidly, while its value starts to remain constant. One chooses the smallest value for the batch size that allows the training to finish in a reasonable time.
To give the reviewer a better idea in the Figure below it can be seen the value of the cost function after 100 epochs vs. the training time for different batch size (the plot has been taken from Michelucci, U. *Applied Deep Learning—A Case-Based Approach to Understanding Deep Neural Networks*; Apress Media, LLC: New York, NY, USA, 2018; ISBN 978-1-4842-3789-2)



(From: Michelucci, U. *Applied Deep Learning—A Case-Based Approach to Understanding Deep Neural Networks*; Apress Media, LLC: New York, NY, USA, 2018; ISBN 978-1-4842-3789-2)

Note that the plot has been generated with a different dataset, but the behavior is quite general. One usually chooses the batch-size value at the “elbow” position in the plot. In our case, this value was 32. The fact that the number is a power of two is more due to tradition than to any practical reasons. In the text, how we have chosen 32 is explained in a much more concise way in lines 120-121. This process of optimising parameters is called hyper parameters and is discussed at length in the reference to the book mentioned above that is in the paper.

Line 123: What is the CPU and how many cores are needed? Please, provide concrete information focusing on the facts required here and avoid woolly formulations.

The training has been performed on a 2.2 GHz 6-Core Intel Core i7, with 32 GB of RAM. No GPU acceleration was used.
The text in the paper has been updated.

Line 130: Why did you use the mean of the absolute error instead of the residuals?

This is again a very good question. A residual analysis cannot be used while training a neural network. For the training of the neural network, one has always to decide from a single-number metric how good or bad each single prediction is. We have not chosen the Mean Squared Error since we did not want to give more weight to big errors, and we wanted a more balanced metric. Therefore, we have chosen the Absolute Error (AE).
A possible further analysis would consist in doing a residual analysis intended as "the difference between the ith observed response value and ith response value that is predicted" (as defined by James, Gareth, et al. in *An introduction to statistical learning* at page 62) analysing specifically different parameter ranges.

Line 141: Why did you choose accuracy as the only performance parameter? To my knowledge, the accuracy can be biased and shouldn't be the only figure of merit, when it comes to describe the performance of an algorithm (http://eecs.wsu.edu/~holder/courses/cse6363/spr04/pubs/Provost98.pdf)

This is a very good (although sometime overlooked) reference from 1998 that the authors knew already. First of all, please note that the neural network has been trained to solve a regression problem, minimizing the Mean Square Error, and not a classification problem. We talk about accuracy when we define the new metric (Error Limited Accuracy) since we wanted to classify each observation as right (below a certain given error from the expected value) or false (above a certain error from the expected value) therefore converting a regression problem into a classification one. This allowed us to estimate the maximum error with which the sensor could classify 100% of the observations. Please note also that we don't use the accuracy to compare models or choose the best one. All the neural network training has been structured as a regression problem (as stated above), and therefore, the cited paper applies only minimally to our work.

We would like to comment nonetheless quickly on the interesting paper. Note that as they state in the paper accuracy "assumes equal misclassification costs". This is the case in our work. We are only interested if an observation is below or above a certain error from the expected value (in absolute value). In medicine, for example, this is not the case. One wants to classify correctly sick and healthy people and the costs of missing one or another is quite different, since the cost of sending home someone sick is much higher (could result in death) than treating a healthy one (one talk about true positives and false negatives). In this case, the ROC curve, the confusion matrix, Sensitivity, Specificity, or F1 are surely better metrics. Accuracy is also a very misleading metric in case of a heavily unbalanced dataset, but this is also not our case. You can find an extensive discussion about metrics in this case at pages 239-245 here Michelucci, U. *Applied Deep Learning—A Case-Based Approach to Understanding Deep Neural Networks*; Apress Media, LLC: New York, NY, USA, 2018; ISBN 978-1-4842-3789-2.

Line 180: Have you thought about reproducibility / calibration-transfer? How is the sensor intended to be used by the user? Does the user have to make the whole calibration by themselves again?

The authors thank the reviewer for the question. To train the neural network for the first time a long time is required (65 hours only for the data gathering part). But subsequently, what is called transfer learning could be used to quickly fine-tune the network if the conditions of the training and usage are not the same. The authors have already done a few tests and have found out in very preliminary results that using only a few measurements (ca. 50 different values of T and [O2]) and just a few minutes of training would be enough for a new sensor to adjust the network trained on the old sensor. Those results are very preliminary and need to be verified since the work is in progress.

Line 198: Single indicator sensors are unusual for optical chemical sensors? How would the approach perform in multi-layer systems and deal with their inherent indicator-interaction that possibly occur?

The author would be thrilled to try the method on a multi-indicator system. We do not currently have any sensing element with multi elements but would be glad to measure one.

Line 199: A calibration curve of the final matrix would help to understand how the user can then analyze its unknown samples.

Actually, the idea of the neural network model is to overcome the difficulty in finding the calibration matrix. As can be seen from Fig 2 and 3 there are several combinations of O2 and T and \omega which result in the same phase shift. Even with a frequency sweep (Fig 3 and 4) the curves with different pairs of O2 and T values may result in almost indistinguishable curves.

The neural network weights play the role of the calibration matrix, although mathematically in a very different way, as it is effectively a composition of non-linear functions that produce the desired output after training.

Line 200: Please, provide more (concrete) information on the comparison. What is the typical error of commercial sensors? Which sensors did you use for comparison?

A comparison of several commercial fluorescence-based sensors for O2 is reported in the paper Wolfbeis O.S. Bioessays 2015 37: 921–928. The reference to the paper was added to the text.