A Deep Learning Interpretable Classifier for Diabetic Retinopathy Disease Grading

Jordi de la Torre[a,*], Aida Valls[a], Domenec Puig[a]

[a] *Departament d'Enginyeria Informàtica i Matemàtiques.*
Escola Tècnica Superior d'Enginyeria.
Universitat Rovira i Virgili
Avinguda Paisos Catalans, 26. E-43007
Tarragona, Spain

Abstract

Deep neural network models have been proven to be very successful in image classification tasks, also for medical diagnosis. The main weak point is its lack of interpretable explanations about the reported results, although they are able to give results with high statistical confidence. The vast amount of parameters of these models make difficult to infer a rationale interpretation from them. In this paper we present an interpretable classifier able to classify retina images into the different levels of diabetic retinopathy severity with good performance, as well as of explaining its results by assigning a score for every point in the hidden and input spaces, evaluating its contribution to the final classification in a linear way. The generated visual maps can be easily interpreted by an ophthalmologist in order to find the lesions present in the retina that are causing the disease.

Keywords: deep learning, classification, explanations, diabetic retinopathy, model interpretation

2010 MSC: 68T10

*Corresponding author

Email addresses: jordi.delatorre@gmail.com (Jordi de la Torre), aida.valls@urv.cat (Aida Valls), domenec.puig@urv.cat (Domenec Puig)