Retrieval of similar cases to improve the diagnosis of diabetic retinopathy

Álvaro Sanz-Ramos[1], Luis Ariza-López[1], Cristina Montón-Giménez[2], and Antonio A. Sánchez-Ruiz[1][0000−0003−0034−1254]

[1] Departamento de Ingeniería del Software e Inteligencia Artificial,
Instituto de Tecnología del Conocimiento,
Universidad Complutense de Madrid, Madrid, Spain
{alsanz16,luiariza,antsanch}ucm.es

[2] Hospital Universitario Severo Ochoa, Madrid, Spain
cristina.monton@gmail.com

Abstract. Diabetic retinopathy is the leading cause of new cases of blindness in adults and its early detection is fundamental. In this work we propose a technique to retrieve similar fundus images of already diagnosed eyes to support the physicians when they must diagnose a new patient. The similarity between images is computed using standard distances on image embeddings extracted from the last layers of a neural network. Our preliminary experiments seem to confirm that embeddings encode important medical information, the similarity between embeddings aligns with the specialist's concept of similarity, and the similar images retrieved are almost always relevant for the diagnosis.

Keywords: Case-based Reasoning, Deep Learning, Image embeddings, Diabetic retinopathy

1 Introduction

Diabetes is one of the most widespread and difficult to control diseases today. It is estimated that 783 million people will be living with diabetes by 2045, and the cost of treating the disease has tripled in the last 15 years [3]. Diabetic Retinopathy (DR) is a common complication of diabetes (affecting 40-45% of diabetic patients) caused by vascular damage from persistently elevated blood sugar. In fact, DR is already the leading cause of new cases of blindness in adults and costs more than $500 million in the United States alone.

Early detection is key in this disease, as the most obvious signs appear when the disease is already too advanced to be effectively treated. The most common detection technique is *ophthalmoscopy*, which consists of dilating the pupil to photograph the retinal *fundus* (which includes the central and peripheral retina, the optic disc and the macula) using a specialized camera. The diagnosis of the disease is then made by a physician, who carefully examines the photographs

[3] Diabetes Atlas https://diabetesatlas.org/.

to determine whether the patient has the disease and its degree of development (mild, moderate, severe or proliferative).

Diagnosis of the disease is not without difficulties. The equipment needed to take the photographs is expensive and the training of specialists capable of interpreting the images requires years of study. Some signs of the disease, such as microaneurysm or narrowing of the blood vessels, are difficult to detect even by trained professionals. In fact, studies have found that different physicians agree on the exact classification of DR in one eye in only 69% of cases [5]. For all these reasons, research has been underway for years on the development of automated methods to effectively detect DR to diagnose the disease earlier and thus treat the disease more effectively.

Based on previous research in other medical settings and the availability of labeled datasets, we believe that Case-based Reasoning (CBR) strategies could play an interesting role providing support during the diagnosis of DR. CBR is based on the idea that similar problems tend to require similar solutions, and it is usually easier to build those solutions by adapting solutions applied to previous similar problems than to build them from scratch from general domain knowledge. This problem-solving strategy is more intuitive and transparent than other approaches based on black-box models, being especially interesting in some domains such as, for example, medicine. For example, when a specialist must diagnose a new patient, we can provide similar cases already diagnosed that can be of support.

The retrieving of similar cases, however, is challenging when working with non-symbolic representations of information such as medical images. In this work, we propose the use of image *embeddings*, generated by a deep neural network architecture, and standard distances to retrieve similar cases. Embeddings are a compact, low-dimensional representation that tends to group images with similar characteristics. Moreover, training deep learning models can be a time-consuming and computationally expensive process. To speed up training, we preprocess the images and use an efficient network architecture pretrained on ImageNet. Transfer learning has proven to be a very effective technique even among domains as different as object recognition and medical image diagnosis. Finally, our preliminary experiments have yielded promising results: similarity between embeddings seems aligned with the specialist's concept of similarity when analyzing the images, and the images retrieved are almost always relevant.

In summary, the main contributions of our work are:

- Using a efficient network architecture pretrained on ImageNet to train a competitive DR classifier with low computational resources.
- Carefully selecting an embedding-based representation that allows a kNN classifier to obtain network-like performance.
- A preliminary evaluation with a specialist that seems to confirm that similarity between embeddings is aligned with similarity from a medical point of view.

The rest of the paper is organized as follows. Next section describes the related work. Section 3 describes in more detail the disease and its diagnosis.

Section 4 describes the dataset used and the preprocessing of the images. Section 5 describes the neural network architecture used to diagnose DR from images of the retina. Section 6 explains how to perform a semantic search to retrieve similar images using embeddings extracted from the last layers of the network. Section 7 describes the experiments we have conducted with the help of an ophthalmologist to validate our proposal. Finally, the paper ends with some conclusions and lines of future work.

2 Related work

Case-based reasoning (CBR) has been successfully applied in the medical field for clinical decision-making applied to therapy and diagnosis [19]. Respect to cancer treatment, researchers used clinical data to select classified breast cancer tissue and then reduced the cases by searching for similar DNA methylation patterns [2]. Other researchers performed adaptation of the adjustment of parameters before segmentation on an existing CBR system to segment renal parenchyma [15].

CBR has also been used in conjunction with deep learning techniques. Retrieval is an important aspect in which neural networks can play an important role. We find examples in the literature were Class-to-Class Siamese Networks [24] or Siamese Graph Neural Networks [8] have been used for this purpose. Other works study the capabilities and limitations of neural networks to learn adaptation knowledge in CBR systems [23,25].

Regarding the use of embeddings, it is known that the layer from with the embeddings are extracted may vary their quality [12]. In the medical field, there are several works that use embeddings. For example, to perform medical image report generation where a visual attention branch captures image embeddings [22]; to label medical images to introduce as much novelty as possible to an existing dataset [3]; or for abnormalities detection on chest radiographs [6].

Medical Image similarity has been subject of study due to the interest in accurately relating information in the different images for diagnosis and treatment. For example, color and texture histograms help to identify similar images from an endoscopy and reduce the entire volume of frames, so it will be easier for a diagnosing physician [9].

3 Diabetic retinopathy

Diabetic retinopathy (DR) is a common complication of diabetes, caused by high blood sugar levels damaging the blood vessels and nerve tissue of the retina. The earliest anatomical changes linked to the disease are the narrowing of the retinal arteries and the dysfunction of neurons of the inner retina. As the disease progresses, damage may reach the outer retina, provoking subtle visual dysfunction, and weaken the blood-retinal barrier, that protects the retina from many substances present in blood, as immune cells or toxins.

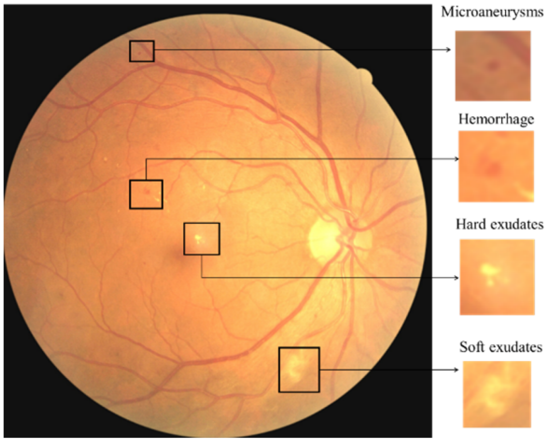


Fig. 1: Example of lesions associated to diabetic retinopathy [1].

In the later stages of the disease, the basement membrane of the retinal blood vessels thickens and the capillaries degenerate. This leads to a progressive ischemia of the tissues, which leads to degeneration of the neurons and glial cells of the retina. Narrower capillaries are prone to the appearance of microaneurysms, which may cause swelling or leak of them. DR is comorbid with macular edema, caused by the deposition of fluid and protein under the macula of the eye, causing it to thicken and swell.

According to the International Clinical Diabetic Retinopathy Disease Severity Scale [21], the severity of DR can be graded into five stages (0-4): no retinopathy (0), mild non-proliferative DR (NPDR) (1), moderate NPDR (2), severe NPDR (3), and proliferative DR (4). The grading depends on the number and size of different related lesions and complications.

Figure 1 shows some examples of lesions indicating the presence of DR. It is important to note that lesion detection is a challenge even for medical specialists after several years of training, especially in the early stages of the disease. Besides, the quality of the image is usually not this good and depends on several factors over with the physician has little control, including the age of the equipment and the patient cooperation.

4 Dataset and preprocessing

For training our model, we have chosen the EyePacs Dataset [4]. This dataset was offered for the Diabetic Retinopathy Detection Kaggle competition [4] and it can be freely used for research after accepting the conditions of the competition. The dataset has a total of 88,704 labeled images, divided into two sets: a training set consisting of 35,126 images and a test with 53,578 files.

[4] https://kaggle.com/competitions/diabetic-retinopathy-detection

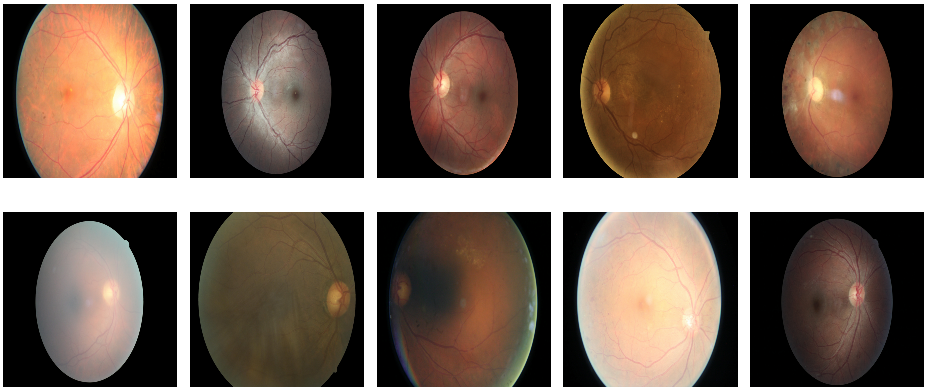

(a) Grade 0 (b) Grade 1 (c) Grade 2 (d) Grade 3 (e) Grade 4

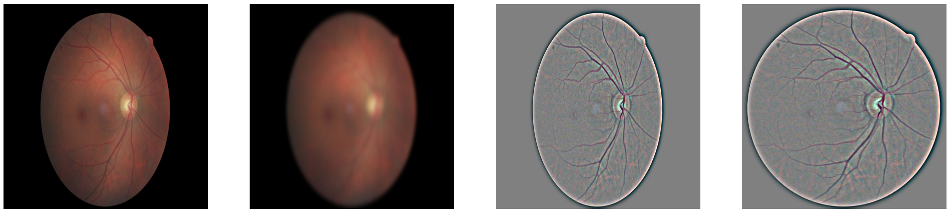
Fig. 2: Example of images of different classes. Notice the different color, brightness and scale of each image.

The images come from a variety of cameras and may have different colors, resolutions and capture different part of the eyes. Also, different images may have different orientations, depending on the type of equipment used to take the photograph.

Both train and test images have been graded for DR by a professional ophthalmologist: the rating is a number between 0 and 4 representing no DR, mild, moderate, severe DR, or proliferative DR. Concerning class distribution, the dataset is heavily unbalanced, as no DR images account for 73.48% of the training set, while the Mild, Moderate, Severe and Proliferative DR have a representation of 15.06%, 6.95%, 2.49% and 2.02% respectively. Figure 2 shows examples of different fundus images of the dataset according to the degree of diabetic retinopathy diagnosed. The last nuance about the dataset is the strong correlation between the left and right eye grading (Pearson correlation coefficient $\rho = 0.85$), which made us explore *binocular* methods, that consider information coming from both eyes for grading.

The first step of the preprocessing is to crop the background as it does not give any information, however, this is not immediate, as the color varies in the images, although it is dark in all cases. After several approximations, we used a scheme inspired by the one carried by the winner of the Kaggle competition, Ben Graham [5], consisting in 3 steps: scaling down the image, so the radius has a fixed length by finding the diameter of the eye, then we overlap Gaussian noise over the image by convolving it with a kernel created by extracting values from a $Normal(\eta = 0, \sigma = 10)$ distribution and lastly, we homogenize the background and crop the image with radius equal to 0.9 times the radius of the eye and then set all the pixels outside the circle to gray.

[5] https://www.kaggle.com/competitions/diabetic-retinopathy-detection/discussion/15801



(a) Original image

(b) Blurred image

(c) Overlapped image

(d) Cropped image

Fig. 3: Process of preprocessing an image. The original image is rectangular, so it appears stretched when scaled to a squared image.

Figure 3 shows the entire preprocessing process for an image, except for changes in scale. The application of Gaussian noise successfully highlights the blood vessels and the main anatomical structures, as the optic nerve. It also removes some undesirable effects, as the changes in brightness. While this method sacrifices most color information, we have found that color is mostly dependent on the brightness conditions and does not contain diagnostic information by itself. Since we have limited data, it is convenient to use data augmentation techniques to create small variations of images. We apply the following transformations (from the Albumentations [6] library) to each image with probability 0.5: scaling the image by a random factor, rotating it by a random angle, flipping it vertically, horizontally or both. Also, normalization is applied after data augmentation and immediately before feeding the image to the model. The normalized image I' is calculated from the augmented image I as:

$$I' = \frac{I - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation, calculated separately for each channel over the whole training set.

5 Network architecture and training

To generate the embeddings we use an EfficientNetv2-B3 network [20], a medium-sized convolutional neural network which strikes a good compromise between accuracy and complexity. The output of the convolutional network is pooled using *global average pooling* and fed to a classifier, designed as a stack of four dense layers with decreasing width (Figure 4).

Using a multilayer classifier increases the risk of overfitting and usually does not improve performance, but serves as a gentle way to perform dimensionality reduction on the features extracted by the convolutional network.

[6] https://albumentations.ai

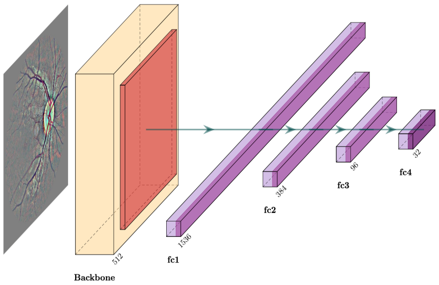


Fig. 4: Architecture for embeddings extraction.

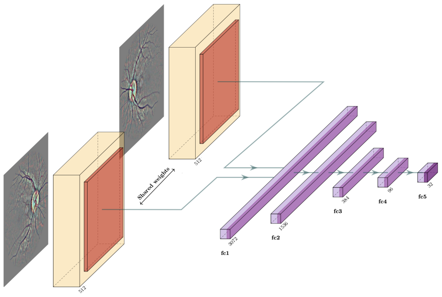


Fig. 5: Architecture for predictions. The classifier *blends* the information from both images

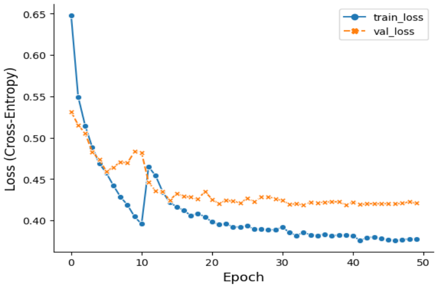


Fig. 6: Loss evolution during training.

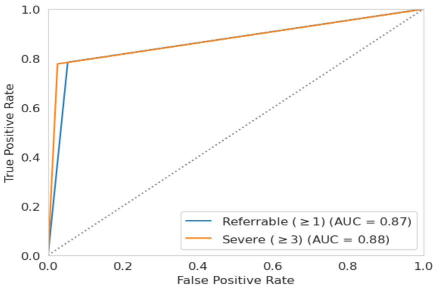


Fig. 7: The ROC curve shows promising results with an area under the curve of 0.87 for referrable cases (some sign of diabetic retinopathy) and 0.88 for severe cases (at least grade 3).

The EfficientNet implementation, from the open source library *Pytorch Image Models* [7], has been pretrained on ImageNet21K, a large dataset for general purpose image classification problems. This technique has been proved to improve performance and reduce training time in several types of datasets (including medical ones [10]) even if they are quite different to the pretraining one, as the network can reuse most of the features originally learned [16].

The model was trained on the preprocessed images from the EyePACS training dataset to grade diabetic retinopathy, using cross-entropy as the loss function. We found the best results using AdamW [11,14] as an optimization method, combined with L^2 weight decay and *label smoothing* [18] to prevent overfitting. We performed data augmentation to increase the effective size of the dataset, by randomly applying the following operations: a rescale by a random factor between 0.9 and 1.1, a rotation by a random amount of degrees and a vertical,

[7] https://github.com/rwightman/pytorch-image-models

horizontal or mirror flip, each with probability 0.5. After the augmentation, the images were resized to a 512×512 resolution, normalized and fed to the model.

Figure 6 shows the evolution of train and validation loss during training: the peak on iteration 10 is caused by the inclusion of late dropout [7,13] to the first two layers of the classifier in order to address overfitting. The model was trained for 50 iterations (around 10 hours) on a single Nvidia GeForce GTX 1070 GPU.

To evaluate the performance of the model, we designed a variation of the classifier with twice as wide layers as the original (Figure 5). The classifier is simultaneously fed the left and right eyes pooled features, so it can combine the information for both eyes, and will output a vector of *logits* containing the unnormalized probability of each grade of the disease for each eye. This classifier was trained independently using AdamW as the optimization method, freezing the weights of the convolutional network.

This classifier was fed five copies of each pair of images in the test dataset, each rotated independently by a random amount of degrees and normalized. We obtained the final prediction by computing $\sum_{i=0}^{4} i\,\mathrm{softmax}(y^{(k)})_i$ for each of the $y^{(1)}, \ldots, y^{(k)}$ vector of logits and averaging the results, exploiting the fact that the grades of the disease are ordered. Instead of rounding up the result to the closest integer, we used the set of thresholds that maximized Cohen's κ over the validation set (0.57, 1.37, 2.30, 3.12).

The final accuracy of the model is 83.31% and the obtained Cohen's kappa score is $\kappa = 0.8491$. The ROC curve (Figure 7) shows promising result, with solid detection of the disease from the early stages. We found that our model compares favourably to much larger architectures reported in the literature, requiring a fraction of the computational resources both for training and inference. To provide some context, these results would set us second in the *Diabetic Retinopathy Detection* Kaggle competition, with a difference of just 0.0005 from the first solution, which uses an ensemble of three neural networks, each one significantly larger than ours. While ensembles are a reliable way to improve performance, they are inconvenient since they increase training costs and, what is more important, multiply the time and resources needed for inference. Our approach shows the viability of using pretrained smaller models to obtain excellent results in a computationally efficient way.

6 Semantic search based on embeddings

To generate embeddings for each image, we use the neural network from the previous section (Figure 4) as a feature extractor by feeding each image of the dataset rotated by a random number of degrees and normalized. A differently sized representation is extracted by intercepting the activation of the different layers of the classifier. This process generates multiple embeddings for each image, of dimensions 1536, 384, 96 and 32.

Interestingly, we found that applying classical dimensionality reduction techniques (as UMAP, t-SNE or PCA) to the output of the convolutional neural network creates very low-quality embeddings. In contrast, the multilayer classi-

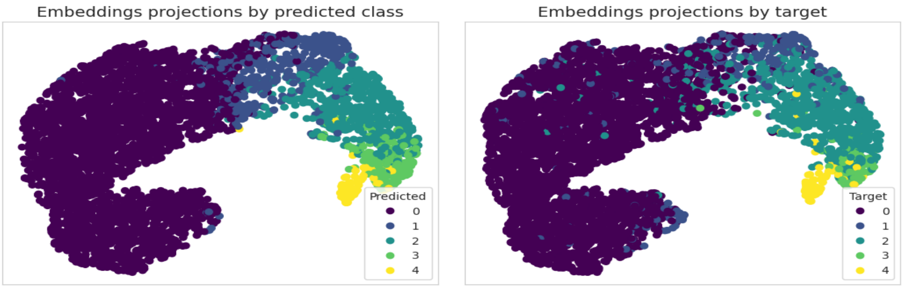
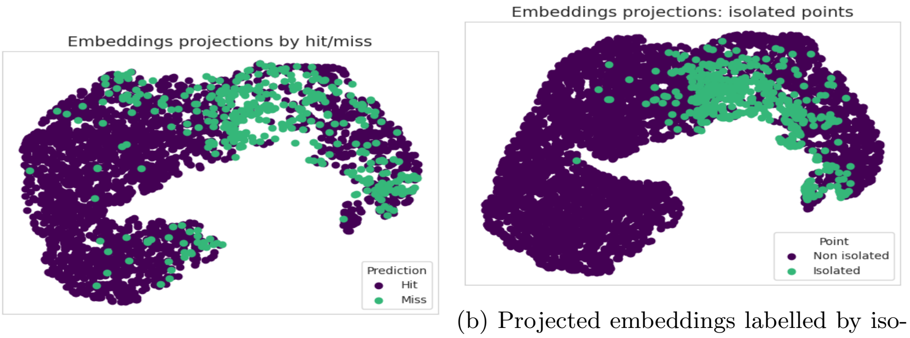


Fig. 8: Projections of the embeddings of a sample (N = 5,000) from the training dataset, labelled by predicted DR grade and target (real grade)



(a) Projected embeddings labelled by hit/miss on the original image (grade predicted correctly)

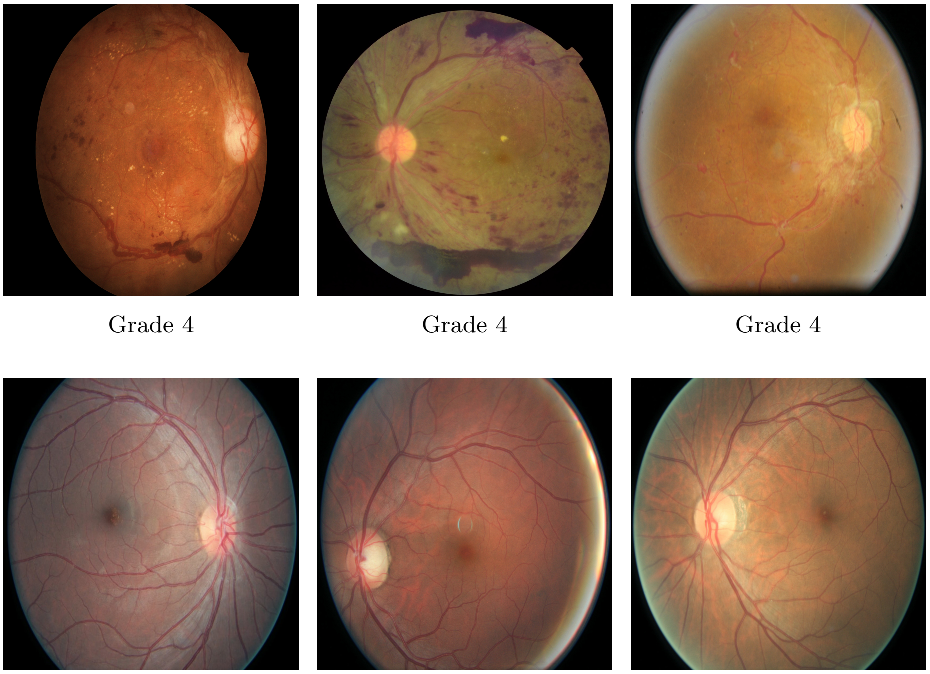
(b) Projected embeddings labelled by isolation (distance to the closest point is over two standard deviations of the average distance)

Fig. 9: Projections of the embeddings of a sample (N = 5,000) from the training dataset, labelled by hit/miss and isolation

fier works as a non-linear context-aware technique for dimensionality reduction that produces much better results.

The generated embeddings have a different nature than the ones created by an *autoencoder* architecture, as the latter encodes the input data (the graphical structure of the image itself) while the former encodes the diagnostic information of the image, ideally disregarding extraneous information, as varying brightness or contrast conditions.

Since the output layer of the classifier is purely linear, the last layers of the classifier should represent images of different classes in (approximately) linearly separable regions. Therefore, we should expect that some diagnostic factors (at least the grading of the image itself) are encoded in the spatial structure of the embeddings. Figure 8 puts this hypothesis to test: it shows a visualization of the 32-dimensional embeddings for a sample of 3.000 points, projected to the plane using UMAP [17] and labelled both by predicted and target class. The



Grade 4 Grade 4 Grade 4

Grade 0 Grade 0 Grade 1

Fig. 10: Two instances (one per row) of retrieval of the two most similar images to a given one (leftmost image), measuring similarity as the distance between 96-dimensional embeddings using cosine distance.

image leads us to two relevant assertions: (1) the network successfully creates a representation of the input image that spatially encodes diagnostic information of the disease and (2) the spatial structure is robust enough to persist after severe dimensionality reduction.

Figure 9a explores the distribution of embeddings of wrongly predicted images. These images concentrate over some small regions of the space of the embeddings, which can be used to classify "high risk" areas, where the performance of the model may be suboptimal. Interestingly, these zones mostly coincide with the areas of distribution of isolated points (Figure 9b), characterized as those points which distance to the closest point is over two standard deviations the average one.

In order to further test our assertions, we implemented a k-NN classifier using the 96-dimensional embeddings and different metrics. The class is obtained as the weighted average of the neighbors, weighted by the inverse of the distance.

The results, shown in Table 1, are impressively solid, achieving even a higher accuracy than the neural network for some choices of k and a reasonable value for Cohen's κ. The results are robust between metrics and choices of k but

	k = 3		k = 11		k = 15	
	Accuracy	κ	Accuracy	κ	Accuracy	κ
Manhattan	0.8322	0.7810	0.8437	0.7945	0.8439	0.6078
Euclidean	0.8324	0.7825	0.8439	0.7951	0.8440	0.6084
Cosine	0.8336	0.7840	0.8431	0.7958	0.8425	0.6059

Table 1: Results obtained by a k-NN classifier ($k = 11$) on 96-dimensional embeddings.

using other dimensions for the embeddings severely degraded performance, which reinforces the importance of using adequately sized embeddings for each task.

We can use the fact that the spatial structure of the embeddings contains diagnostic information to model *semantic search*, the retrieval of images with similar diagnostic signs to a given one, as a *nearest neighbor search* problem.

Figure 10 shows the result of applying this technique to find the two images in the train set most similar to a base image from the test set, by using cosine distance to measure the distance between embeddings. We found that in 80.49% of the cases, the image recovered as the closest one using cosine distance has the same grading than the original one and is within one level in 93% of the cases.

As one might expect, there is an inverse correlation between the distance of the most similar image to the base and both having the same label ($\rho = -0.21$). In fact, an increase of the cosine distance between both images of 0.01, reduces the probability of both having the same grade by 50.39% on average.

Surprisingly, given the robustness of k-NN to the choice of metrics, we found that different distance functions retrieved generally different images and the measured similarity of a given image can significantly vary between metrics. For example, we found that the closest image according to the cosine distance only was between the top 5 closest images according to the Euclidean distance in 38.53% of cases.

7 Experiments and results

Two preliminary experiments have been conducted with the collaboration of an ophthalmologist with experience in the treatment of retinal diseases.

The purpose of the first experiment is to check whether the similarity computed from the image embeddings is consistent with the specialist's intuition of similarity when analyzing the same images. To this end, we show the specialist 1 unlabeled base image from the test set and 5 labeled images from the training set. The 5 images from the training set were selected according to their similarity to the test image so that the most similar image and one in the first 4 quintiles appear. The specialist is then asked to sort the images from the training set in order of similarity to the base image. The specialist can view all the images as many times as she wants with no time limit. This process was repeated 10 times with 10 different images from the test set.

Base	Specialist order	Embeddings order
19351	{18311 46571 129001 277821} {282271}	18311 277821 129001 46571 282271
208831	{246561} {162231 21921r 32104r 351321}	246561 351321 21921r 32104r 162231
30476r	{11219r} {13843r 204641 25222r 323581}	11219r 25222r 13843r 323581 204641
314661	{24711r 29906r 334691} {309191} {10047r}	29906r 24711r 334691 309191 10047r
36651r	{27153r} {13714r 133121 3395r}	27153r 133121 13714r 67881 3395r
37151r	{33251r} {3291 6569r 29141r}	33251r 29141r 6569r 3291 314281
38582r	{74871} {263811 279731} {282271} {172211}	74871 279731 263811 172211 282271

Table 2: Summary of the results of experiment 1. The first column contains the ids of the base images. The second column shows the specialist's order (no order is assumed within each set). The last column shows the order according to the cosine similarity and the embeddings.

At the beginning of the experiment, the specialist asked what was exactly meant by "similarity between images". There are many different criteria that can be taken into consideration: same eye (right or left), similar age of the patient, same type of lesions, same degree of DR development, images made by the same type of imaging device, ... As our neural network calculates the embeddings in the context of a system to diagnose the degree of DR, we instructed the specialist to only take into consideration lesions related to the diagnosis of that disease. Once the similarity criterion was set, the specialist made us realize that she could not sort the images corresponding to healthy eyes, as none of them had lesions. After analyzing the situation, we asked the specialist to group the images according to their similarity to the test image, but without having to order the images in the same group.

Of the 10 test images analyzed, 3 were discarded because they did not have adequate quality for diagnosis. The results of the other 7 images are shown in Table 2. In most cases, the specialist grouped the images into 2 or 3 sets using the type and number of lesions present in each image as the main criterion (her diagnosis of the degree of RD did not always match the image label). These sets should be considered as equivalence classes with no internal order but linearly ordered with respect to its similarity with the base image.

To measure the agreement between the expert's similarity ranking

$$\{d_1^{(1)},\dots,d_{n_1}^{(1)}\},\dots,\{d_1^{(m)}\dots,d_{n_m}^{(m)}\}$$

with the order retrieved from the embeddings $\mathcal{O} = d_{j_1}^{i_1} < d_{j_2}^{i_2} < \cdots < d_{j_l}^{i_l}$, we define the cost of a transposition as $c(d_{j_1}^{(i_1)}, d_{j_2}^{(i_2)}) = \|i_1 - i_2\|$ and calculate the sequence of transpositions of minimum cost transforming the order $d_1^{(1)} < d_2^{(1)} < \cdots < d_{n_1}^{(1)} < d_1^{(2)} < \cdots < d_{n_m}^{(m)}$ into $\mathcal{O}$.

We found this distance to be 0 in all but one case. Notably, the image retrieved as the most similar one using the embeddings is classified by the expert

Base	Diagnosis	Retr. images	Retr. tags	Similar?	Useful?	Modify diagnosis?
1935r	0	1831l 16115r	0 0	yes	yes	no
20883l	2	24656l 23605l	3 2	yes	yes	no
30476r	3	11219r 4155l	3 3	yes	yes	no
31466l	0	29906r 12448r	0 0	yes	yes	no
36651r	2	27153r 17522l	3 2	no	no	no
37151r	1	33251r 23541l	2 1	yes	yes	no
38582r	0	7487l 25632r	0 0	yes	yes	no

Table 3: Summary of the results of experiment 2. The first 2 columns contain the id of the base image to be diagnosed and its initial diagnosis. Columns 3 and 4 show the ids and labels of the 2 most similar images retrieved. The last 3 columns show, respectively, whether the specialist considered the retrieved images to be similar to the original one, whether seeing those images was useful and if, after viewing the images, she wanted to modify her initial diagnosis.

in the group of the most similar images in all cases, which sustains the claim that our strategy retrieves diagnostically similar images.

The purpose of the second experiment is to make a preliminary study of the usefulness and quality of the retrieved images to help the specialist in her diagnosis. To do this, we again showed the specialist the same 7 base images from the previous experiment and for each of them we followed the following protocol: (1) ask for an initial diagnosis of the image, (2) show the two most similar labeled images from the training set, (3) ask the specialist if she thought they were similar to the base image, (4) ask the specialist if it was useful to be able to see those images, and (5) ask if, in view of those images, she wanted to modify her initial diagnosis.

The results of the experiment are collected in the Table 3. The initial diagnosis of the specialist always corresponds to the label of one of the images retrieved using the embedding-based similarity and never differs by more than one degree from the other. Furthermore, in none of the cases are images of diseased eyes retrieved from healthy eyes or vice versa. All retrieved images were considered similar to the originals, except in the case of image 36651 in which the specialist indicated that the retrieved images had different types of lesions. In all other cases, the specialist considered that being able to view those similar images could be helpful for diagnostic support. In no case did the specialist modify her initial diagnosis.

Asked about the possibility of modifying the diagnosis if the system retrieved images with very different labels, the specialist said that she would continue to rely on her judgment unless the system could provide a really convincing explanation. Finally, asked about her overall impressions, she told us that she had been surprised by the system's ability to find similar images and that this feature could be useful for residents.

8 Conclusions

In this work we propose a technique to retrieve fundus images of already diagnosed eyes similar to undiagnosed ones, with the aim of providing relevant cases to a specialist who must diagnose whether a patient suffers from diabetic retinopathy (DR). The search for similar images is based on the use of embeddings extracted from the last layers of a neural network trained to diagnose DR. To speed up the network training, we preprocess the images, use an efficient network architecture (with a good trade-off between performance and size) and start from a pre-trained model on ImageNet. In addition, to improve the quality of the predictions we use images from both eyes of the patient.

The embeddings produced by the network encode information relevant for the diagnosis and group images corresponding to the same degree of the disease in contiguous areas of the latent space. Applying standard distances on the embeddings, we can perform semantic searches and retrieve images with similar medical characteristics. The choice of the distance function significantly influence the images retrieved but we have not found convincing evidence on which function better reflects diagnostic similarity.

Preliminary experiments performed with the collaboration of an ophthalmologist have produced promising results that encourage us to continue this line of research. The similarity between embeddings (measured by the cosine distance) seems aligned with the specialist's concept of similarity when analyzing the images, and the similar images retrieved are almost always relevant.

During the experiments we have learned that the concept of inter-image distance for a specialist is not so simple to define and can be based on a multitude of factors. Moreover, it is much easier for the specialist to define a partial ordering relationship between images than to order them by similarity. In general, the retrieval of diagnosed images is useful to reinforce the specialist's opinion, but it is not sufficient to change the specialist's mind.

However, there is the risk of leading the expert to confusion, by presenting images that the specialist does not consider related. Our conjecture, supported by our experiments, is that the mismatch between embedding distance and semantic similarity is more likely to happen for embeddings in certain “high risk” areas of the latent space. These regions share defining characteristics: they are formed by relatively isolated points, they contain points from different classes and are much more likely to contain misclassified points. The identification of these areas can be used to detect the cases where retrieval based in semantic embeddings may offer suboptimal results.

Acknowledgements

This work has been partially supported by the Spanish Committee of Science and Innovation (PID2020-114596RB-C21, PID2021-123368OB-I00) and the UCM (Group 921330). We would also like to thank everyone involved in the creation and publication of the EyePacs Dataset.

References

1. Alyoubi, W.L., Abulkhair, M.F., Shalash, W.M.: Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors 21(11), 3704 (2021), https://doi.org/10.3390/s21113704
2. Bartlett, C.L., Liu, G., Bichindaritz, I.: Classifying breast cancer tissue through dna methylation and clinical covariate based retrieval. In: Watson, I., Weber, R. (eds.) Case-Based Reasoning Research and Development. pp. 82–96. Springer International Publishing, Cham (2020)
3. Chinn, E., Arora, R., Arnaout, R., Arnaout, R.: Enrich: Exploiting image similarity to maximize efficient machine learning in medical imaging. medRxiv (2021)
4. Cuadros, J., Bresnick, G.: Eyepacs: An adaptable telemedicine system for diabetic retinopathy screening. Journal of Diabetes Science and Technology 3(3), 509–516 (2009), https://doi.org/10.1177/193229680900300315
5. Gangaputra, S.S., Lovato, J.F., Hubbard, L., Davis, M.D., Esser, B.A., Ambrosius, W.T., Chew, E.Y., Greven, C.M., Perdue, L.H., Wong, W.T., Condren, A.B., Wilkinson, C.P., Agrón, E., Adler, S., Danis, R.P.: Comparison of standardized clinical classification with fundus photograph grading for the assessment of diabetic retinopathy and diabetic macular edema severity. Retina 33, 1393–1399 (2013)
6. Gozzi, N., Giacomello, E., Sollini, M., Kirienko, M., Ammirabile, A., Lanzi, P., Loiacono, D., Chiti, A.: Image embeddings extracted from cnns outperform other transfer learning approaches in classification of chest radiographs. Diagnostics 12(9) (2022), https://www.mdpi.com/2075-4418/12/9/2084
7. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR abs/1207.0580 (2012), http://arxiv.org/abs/1207.0580
8. Hoffmann, M., Malburg, L., Klein, P., Bergmann, R.: Using siamese graph neural networks for similarity-based retrieval in process-oriented case-based reasoning. In: Watson, I., Weber, R. (eds.) Case-Based Reasoning Research and Development. pp. 229–244. Springer International Publishing, Cham (2020)
9. Ionescu, M., Glodeanu, A., Marinescu, I., Ionescu, A., Vere, C.: Similarity analysis for medical images using color and texture histograms. Curr Health Sci J. 48(2), 196–202 (2022)
10. Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., Ganslandt, T.: Transfer learning for medical image classification: a literature review. BMC Medical Imaging 22(1), 69 (2022), https://doi.org/10.1186/s12880-022-00793-7
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
12. Leake, D., Wilkerson, Z., Crandall, D.: Extracting case indices from convolutional neural networks: A comparative study. In: Keane, M.T., Wiratunga, N. (eds.) Case-Based Reasoning Research and Development. Lecture Notes in Computer Science, vol. 13405, pp. 81–95. Springer (2022), https://doi.org/10.1007/978-3-031-14923-8_6
13. Liu, Z., Xu, Z., Jin, J., Shen, Z., Darrell, T.: Dropout reduces underfitting. CoRR abs/2303.01500 (2023), https://doi.org/10.48550/arXiv.2303.01500
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019)

15. Marie, F., Henriet, J., Lapayre, J.C.: A new adaptation phase for thresholds in a cbr system associated to a region growing algorithm to segment tumoral kidneys. In: Watson, I., Weber, R. (eds.) Case-Based Reasoning Research and Development. Springer International Publishing, Cham (2020)
16. Matsoukas, C., Haslum, J.F., Sorkhei, M., Söderberg, M., Smith, K.: What makes transfer learning work for medical images: Feature reuse & other factors. CoRR abs/2203.01825 (2022), https://doi.org/10.48550/arXiv.2203.01825
17. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of open source software 3(29), 861 (9 2018)
18. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, NeurIPS. pp. 4696–4705 (2019), https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html
19. Noll, R., Schaaf, J., Storf, H.: The use of computer-assisted case-based reasoning to support clinical decision-making – a scoping review. In: Keane, Mark T.and Wiratunga, N. (ed.) Case-Based Reasoning Research and Development. pp. 395–409. Springer International Publishing, Cham (2022)
20. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research, vol. 139, pp. 10096–10106. PMLR (2021), http://proceedings.mlr.press/v139/tan21a.html
21. Wilkinson, C.P., Ferris, F.L., Klein, R., Lee, P.P., Agardh, C.D., Davis, M.D., Dills, D.G., Kampik, A., Pararajasegaram, R., Verdaguer, J.T.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 110 9, 1677–82 (2003)
22. Yang, Y., Yu, J., Zhang, J., Han, W., Jiang, H., Huang, Q.: Joint embedding of deep visual and semantic features for medical image report generation. IEEE Trans. Multim. 25, 167–178 (2023), https://doi.org/10.1109/TMM.2021.3122542
23. Ye, X., Leake, D., Crandall, D.: Case adaptation with neural networks: Capabilities and limitations. In: Keane, M.T., Wiratunga, N. (eds.) Case-Based Reasoning Research and Development. pp. 143–158. Springer International Publishing, Cham (2022)
24. Ye, X., Leake, D., Huibregtse, W., Dalkilic, M.: Applying class-to-class siamese networks to explain classifications with supportive and contrastive cases. In: Watson, I., Weber, R. (eds.) Case-Based Reasoning Research and Development. pp. 245–260. Springer International Publishing, Cham (2020)
25. Ye, X., Leake, D., Jalali, V., Crandall, D.J.: Learning adaptations for case-based classification: A neural network approach. In: Sánchez-Ruiz, A.A., Floyd, M.W. (eds.) Case-Based Reasoning Research and Development. Springer International Publishing, Cham (2021)