Transformed Representations for Convolutional Neural Networks in Diabetic Retinopathy Screening

Gilbert Lim, Mong Li Lee, Wynne Hsu
School of Computing
National University of Singapore
{limyongs,leeml,whsu}@comp.nus.edu.sg

Tien Yin Wong
Singapore Eye Research Institute
Singapore National Eye Centre
tien_yin_wong@duke-nus.edu.sg

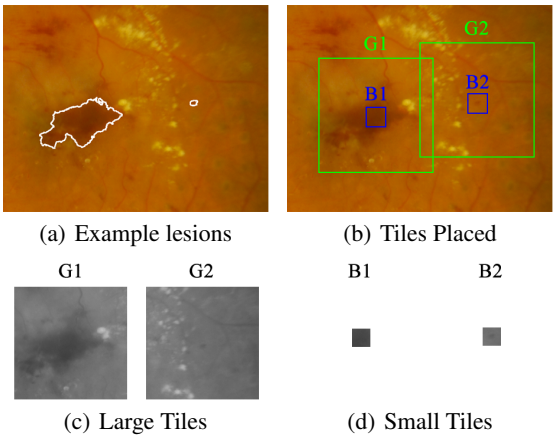
Abstract

Convolutional neural networks (CNNs) are flexible, biologically-inspired variants of multi-layer perceptrons that have proven themselves to be exceptionally suited to discriminative vision tasks. However, relatively little is known on whether they can be made both more efficient and more accurate, by introducing suitable transformations that exploit general knowledge of the target classes. We demonstrate this functionality through pre-segmentation of input images with a fast and robust but loose segmentation step, to obtain a set of candidate objects. These objects then undergo a spatial transformation into a reduced space, retaining but a compact high-level representation of their appearance. Additional attributes may be abstracted as raw features that are incorporated after the convolutional phase of the network. Finally, we compare its performance against existing approaches on the challenging problem of detecting lesions in retinal images.

Introduction

There is a pressing demand for automated systems that can efficiently and cheaply screen large populations for diabetic retinopathy, which may lead to blindness if left untreated. This is often done by manually examining retinal images. However, early signs of retinopathy are often less than obvious even to trained graders, and accurate diagnosis constitutes a complex vision task, where it is not readily apparant how to characterize regions of concern within the image. We therefore propose a solution involving deep convolutional neural networks, which have emerged as one of the best known architectures for tackling such issues.

CNNs have exhibited state-of-the-art classification performances on a wide array of real-world vision assignments, ranging from handwritten text to stereo 3D objects (Cireşan et al. 2011), and handling millions of natural images in thousands of categories (Krizhevsky, Sutskever, and Hinton 2012). They have also been proven in biomedical applications such as breast cancer mitosis detection (Cireşan et al. 2013) and neuronal membrane segmentation (Cireşan et al. 2012). The prevailing approach taken by the above implementations has been to exploit the huge quantity of data available at the lowest-possible pixel level, and train a CNN over as many examples as possible, explicitly converging the large number of free weights to minimize the error function with the labels of the training data.



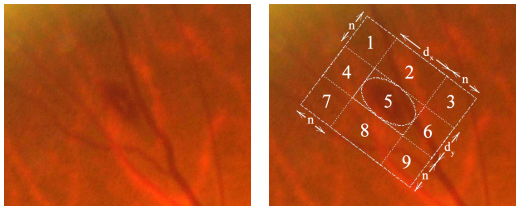
(a) Example lesions (b) Tiles Placed

(c) Large Tiles (d) Small Tiles

Figure 1: The optimal tile width problem

While certainly successful, however, processing all pixels within even a moderate-sized image remains a time-consuming process, even with GPUs. CNNs are generally applied to tile images of standardized sizes, such as in MNIST (LeCun et al. 1998), CIFAR-100 (Krizhevsky and Hinton 2009) and ImageNet (Deng et al. 2009). However, in all these cases, it is assumed that each tile contains an object that is at a suitable scale. This is not the case in retinopathy detection, where lesions of the same class may differ in size by orders of magnitude.

Figure 1 shows two lesions with very different sizes and shapes. This creates a dilemma in the selection of the input tile width. Suppose we place two large tiles and two small tiles centered on them, as shown in Figure 1(b). Figure 1(c) shows the greylevel images corresponding to the green tiles. We note that the large lesion's characteristics are retained in G1. However, in G2, the small lesion is represented by too few pixels and registers as noise. On the other hand, if we select a smaller tile size, we might not be able to characterize the large lesion in its proper context, as its appearance at this scale is approximately uniform (B1 in Figure 1(d)).



(a) Original (b) Object Neighbourhood

(c) Transformed $\boldsymbol{R}$ (d) CNN Tile Input

Figure 2: Example of Representational Transformation

There has been relatively little prior work addressing the challenge of detecting multiple objects at various scales, with current efforts (Schulz and Behnke 2012; Szegedy, Toshev, and Erhan 2013) downsampling the entire image, which would cause smaller lesions such as microaneurysms to disappear entirely. In this paper, we propose instead to first identify promising lesion candidates, before transforming them to a constant-sized representation. This greatly reduces the amount of computation needed, as a large candidate would require no more analysis than a smaller one, and moreover allows us to extrapolate about unseen target objects from incomplete data. Experimental results on two real world datasets show that our approach matches or outperforms existing known state-of-the-art methods.

Overview

We first identify candidate regions that are likely to contain lesions (Multiscale C-MSER Segmentation). Each of the identified candidates are then transformed into tiles of fixed size (Representational Transformation). Finally, the obtained tiles are input to a CNN (Convolutional Neural Network Classification), and individual lesion results are then combined into an image-level diagnosis.

Multiscale C-MSER Segmentation

MSER have been demonstrated (Mikolajczyk and Schmid 2004) to be robust region detectors, and operate on the principle that visually-significant regions are distinguished by a boundary that is wholly either darker than or brighter than its surroundings. Although MSER are fine detectors as-is, they may produce regions at very similar scales, especially when gradients are subtle. A constrained variant, C-MSER (Lim, Lee, and Hsu 2012), has been developed on retinal images. In this paper, we utilize a multiscale extension that searches for C-MSER at different scales in a scale-space pyramid with each successive level created by downsampling the image by a factor of two. At each level of the scale pyramid, we set a minimum allowable region size that is smaller than any target object, so as to suppress spurious noise.

Representational Transformation

Multiscale C-MSER provides a list of candidate regions that are likely to contain some lesion objects. For each candidate region, we obtain a statistical approximation to the shape. Other than being more robust to situations where the lesions are adjacent to vessels, in which case taking an axis-aligned bounding box would significantly underestimate the size of the actual lesion, this also speeds up segmentation as we do not have to maintain pixel lists. The statistical approximation is achieved by keeping track of the raw moments $\mu_{x,y}$ of each region, with the angle θ and axis lengths d_x, d_y of the final approximating ellipsoid of each region at scale pyramid level L as defined by the formulae:

$$\theta = \frac{2[\mu_{1,1} - (2\mu_{1,0}\mu_{0,1})/|Q| + \mu_{1,0}\mu_{0,1}]}{[\mu_{2,0} - (2\mu_{1,0}\mu_{0,1})/|Q|) + (\mu_{1,0}^2)/|Q|] - [\mu_{0,2} - (2\mu_{1,0}\mu_{0,1})/|Q|) + (\mu_{0,1}^2)/|Q|]} \quad (1)$$

$$d_x = 2^L\sigma_x, d_y = 2^L\sigma_2 \quad (2)$$

Note that it is possible that the same region may be detected at multiple scales. To remove duplicates, for each C-MSER CQ_t, we pairwise compare the defining attributes $\{\theta, \mu_{1,0}/|Q|, \mu_{0,1}/|Q|, d_x, d_y\}$ of its approximating ellipse with those of all C-MSER at higher levels in the scale pyramid, and remove the higher-level C-MSER CQ_u if all its attributes are close enough:

$$|\theta_t - \theta_u| \leq 0.5 \quad (3)$$

$$|(\mu_{1,0}/|Q|)_t - (\mu_{1,0}/|Q|)_q| \leq \sqrt{d_x * d_y} \quad (4)$$

$$|(\mu_{0,1}/|Q|)_t - (\mu_{0,1}/|Q|)_q| \leq \sqrt{d_x * d_y} \quad (5)$$

$$0.8 \leq (d_x)_t/(d_x)_u \leq 1.25 \quad (6)$$

$$0.8 \leq (d_y)_t/(d_y)_u \leq 1.25 \quad (7)$$

We then transform each surviving multiscale C-MSER to a square representation $\boldsymbol{R}$ of width w (see Figure 2) where:

$$n = max(1.0, \frac{w}{5\sqrt{d_x * d_y}}) \quad (8)$$

$$d = max(\frac{w}{5}, 2\sqrt{d_x * d_y}) \quad (9)$$

This is achieved by dividing the object neighbourhood into nine rectangles, as shown in Figure 2(b), and performing an affine transform on each individual rectangle, such that it meets the required dimensions for $\boldsymbol{R}$. In situations where some neighbourhoods are extended beyond the image boundaries, we synthesize the unavailable pixels by randomly assigning them a value from adjacent available pixels. Finally, $\boldsymbol{R}$ is normalized on the statistics of the candidate region. Given the mean μ and standard deviation σ of the intensities of the pixels within the approximating ellipsoid, each pixel is mapped to a new intensity value I', from its initial intensity value I:

$$I' = 127 + 10(\frac{I - \mu}{\sigma}) \quad (10)$$

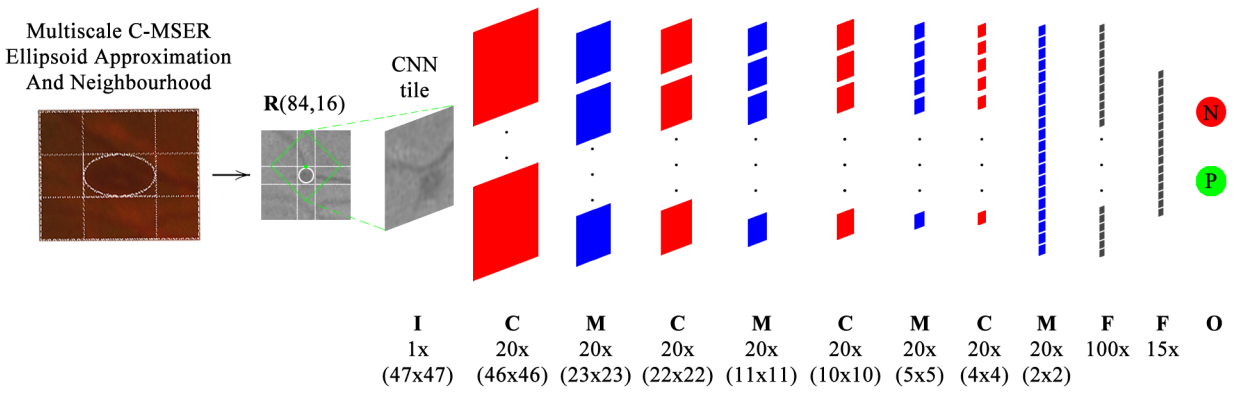


Figure 3: Convolutional neural network architecture
I = Input layer, C = Convolutional layer, M = Max-pooling layer, F = Fully-connected layer, O = Output layer

Some examples of lesion C-MSER, as well as their representative transformations, are shown in Figure 4. It can be observed that despite the variance in their original appearance (top row), the haemorrhages take on a remarkably congruent appearance after transformation (bottom row), while their neighbourhood context is also included.

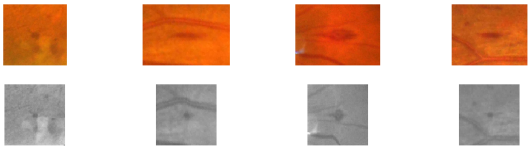


Figure 4: Examples of transformed lesions (haemorrhages)

Convolutional Neural Network Classification

The transformed representations form the inputs to convolutional neural networks (CNNs). CNNs are specialised feed-forward multi-layer perceptrons (MLPs) (Hubel and Wiesel 1968). They generally consist of alternating convolutional and subsampling layers, with at least one fully-connected hidden layer before the output layer. Each layer consists of a number of nodes that are linked, by connections, to nodes in the next layer. Each connection has an associated weight, which are independent for fully-connected layers, but shared through kernels in convolutional layers. This enforces locality and diminishes the number of free parameters for better generalization ability. We train our CNNs by online backpropagation of errors using gradient descent, and utilize max-pooling layers to select good invariant features (Scherer, Müller, and Behnke 2010).

We have further experimented with augmenting the CNN with information about the candidate regions that is lost during the representational transform, such as their initial size, by incorporating them as independent features in the second-last hidden layer. However, this were found not to significantly affect classification performance, suggesting that the transform already retains sufficient information in practice.

Experimental Results

Training of the CNNs was performed on an Intel Core i7-3930K 3.20GHz system with four AMD Radeon HD 7970 GPUs. The implementation is in C++ with OpenCL.

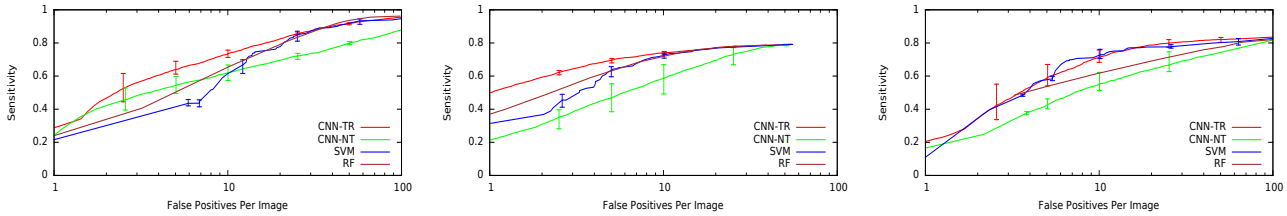
Datasets

For the purposes of evaluating our model, we require labeled sets of retinal images with multiple classes of lesions indicated. Two datasets were selected:

- **DIARETDB1** (Kauppi et al. 2007) is a public database for benchmarking diabetic retinopathy detection from digital images. It contains 89 retinal images of dimensions 1500×1152 independently annotated by four experts. We train on the even-numbered images, and evaluate on the odd-numbered ones, at a ground truth confidence of 0.75.
- **SiDRP** is composed of 2029 images of dimensions 3216×2136 from an ongoing screening program at country-level primary care clinics. Each image is given an overall classification, as well as lesion-level ground truth, by a trained grader. The images were divided into a training set of 1079 images, and a test set of 950 images.

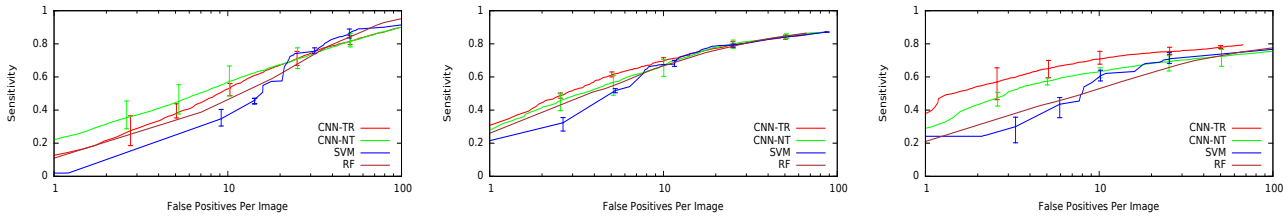
Lesion Level Classification

We evaluate the ability of CNNs to classify lesions based on the transformed representations (CNN-TR) of candidate regions produced by multiscale C-MSER, against CNNs trained directly on the untransformed pixels (CNN-NT), support vector machines (SVM) and random forests (RF). The inputs to CNN-TR are the 47×47 tiles centered on each pixel within the approximating ellipse of the $\boldsymbol{R}$ representational transformation of a candidate (see Figure 3), while the inputs to CNN-NT are on the untransformed candidate. During training, the input tiles are randomly rotated, as to reflect the rotational invariance of lesions. The probability score for each candidate is calculated as the mean over all examined pixels. For SVMs and random forests, we extract twenty features from the pre-transformed C-MSER candidate regions as training vectors. These features are: stability,



(a) Microaneurysms (b) Haemorrhages (c) Hard Exudates

Figure 5: Lesion-level classification results on SiDRP (error bars show min/max sensitivities on cross-validation folds)



(a) Microaneurysms (b) Haemorrhages (c) Hard Exudates

Figure 6: Lesion-level classification results on DIARETDB1 (error bars show min/max sensitivities on cross-validation folds)

size, circularity, compactness, perimeter, major axis, minor axis, aspect ratio, mean RGB values (3), mean Lab values (3), contrast (6). For both these methods, we perform a grid search on the free parameters, and retain the combinations with optimal sensitivity-false positives tradeoff.

The lesion-level classification results are shown in Figures 5 and 6. For SiDRP, the classification performance of CNNs is approximately the same for microaneurysms, whether or not representational transformation is applied. This is to be expected as microaneurysms are defined by their small size, and they already fit comfortably within the selected tile size. However, classification performance is significantly better on transformed input for haemorrhages and hard exudates. For DIARETDB1, the results are generally similar for all classifiers. This may be due to the fact that the ground truth of this dataset is noisy with large variations among the 4 human graders giving rise to situations where unmarked lesions may turn out to be true lesions and marked lesions may, in fact, be false lesions.

Image Level Classification

Finally, we report the image-level classification performance on the SiDRP dataset. The ground truth image-level classification is provided by the graders following established guidelines. Each image is labeled as one of the following set, in increasing order of severity: {*NoDR, Minimal, Mild, Moderate, Severe, Proliferative*}. An image is deemed abnormal if it is of severity scale *Moderate* and above, and normal if the severity scale is *Mild* and below. Images that are determined by the human graders to be ungradable are ignored. In screening, the objective is to retain as few normal images as possible, while identifying all abnormal retinal images.

Using the models trained for lesion-level classification, we search over probability thresholds for each of the three lesion classes, as well as the number of microaneurysms, to assemble the image-level classifiers. Figure 7 shows the sensitivity-specificity tradeoff as we vary the the probability thresholds. We observe that for CNN-TR, we are able to achieve 100% sensitivity with a specificity of 30%; for a sensitivity of 90%, the specificity is 68%. This performance is comparable to that of human graders, and superior to that which can be obtained with SVMs and random forests.

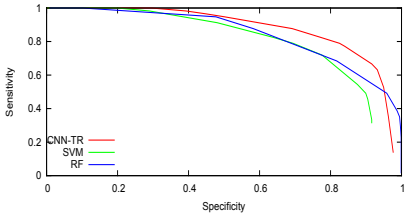


Figure 7: Image-level classification results on SiDRP

Conclusions

We have presented a general representational model that enables effective convolutional neural network classification of target objects at arbitrary scale within images. The effectiveness of this approach was demonstrated in a real-world application of detecting lesions in retinal images, for those classes that do indeed vary significantly in scale.

Acknowledgements

This research was supported by SERI grant R-252-000-396-490.

References

Cireşan, D.; Meier, U.; Masci, J.; Gambardella, L. M.; and Schmidhuber, J. 2011. Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, 1237–1242. AAAI Press.

Cireşan, D.; Giusti, A.; Schmidhuber, J.; et al. 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in Neural Information Processing Systems 25*, 2852–2860.

Cireşan, D.; Giusti, A.; Gambardella, L. M.; and Schmidhuber, J. 2013. Mitosis detection in breast cancer histology images with deep neural networks. In *MICCAI*, 411–418.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

Hubel, D. H., and Wiesel, T. N. 1968. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology* 195(1):215–243.

Kauppi, T.; Kalesnykiene, V.; Kamarainen, J.-K.; Lensu, L.; Sorri, I.; Raninen, A.; Voutilainen, R.; Uusitalo, H.; Kälviäinen, H.; and Pietilä, J. 2007. The diaretdb1 diabetic retinopathy database and evaluation protocol. In *BMVC*, 1–10.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, 1106–1114.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Lim, G.; Lee, M. L.; and Hsu, W. 2012. Constrained-mser detection of retinal pathology. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2059–2062. IEEE.

Mikolajczyk, K., and Schmid, C. 2004. Comparison of affine-invariant local detectors and descriptors. In *Proc. European Signal Processing Conf*.

Scherer, D.; Müller, A.; and Behnke, S. 2010. Evaluation of pooling operations in convolutional architectures for object recognition. In *Artificial Neural Networks–ICANN 2010*. Springer. 92–101.

Schulz, H., and Behnke, S. 2012. Learning object-class segmentation with convolutional neural networks. In *11th European Symposium on Artificial Neural Networks (ESANN)*, volume 3.

Szegedy, C.; Toshev, A.; and Erhan, D. 2013. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems 26*, 2553–2561.